Ranking Games

Margit Osterloh and Bruno S. Frey

Eval Rev published online 4 August 2014
DOI: 10.1177/0193841X14524957

Article

Evaluation Review
1-28

DOI: 10.1177/0193841X14524957
erx.sagepub.com

Ranking Games

Margit Osterloh[1,2] and Bruno S. Frey[1,2]

Abstract

Background: Research rankings based on bibliometrics today dominate governance in academia and determine careers in universities. *Method:* Analytical approach to capture the incentives by users of rankings and by suppliers of rankings, both on an individual and an aggregate level. *Result:* Rankings may produce unintended negative side effects. In particular, rankings substitute the "taste for science" by a "taste for publication." We show that the usefulness of rankings rests on several important assumptions challenged by recent research. *Conclusion:* We suggest as alternatives careful socialization and selection of scholars, supplemented by periodic self-evaluations and awards. The aim is to encourage controversial discourses in order to contribute meaningful to the advancement of science.

Keywords

academic governance, rankings, motivation, selection, socialization

Today rankings based on publications and citations dominate research governance. They serve as the basis for assessing the performance and impact of scholars, departments, and universities. Assessment of research

[1] Zeppelin University, Friedrichshafen, Germany
[2] CREMA—Center for Research in Economics, Management and the Arts, Zurich, Switzerland

Corresponding Author:
Bruno S. Frey, Department of Economics, University of Zurich, Hottingerstrasse 10, Zurich CH-8032, Switzerland.
Email: bruno.frey@econ.uzh.ch

is no longer left to the researchers in the respective fields alone but also to people who do not comprehend the content of research (Maier-Leibnitz 1989; Whitley 2011).

In this article, we focus on research rankings building on the production (number of publications in "top journals") or the reception (citations) of scholarly articles. In particular, we focus on journal rankings according to their "impact factor." Publications in "high-impact" journals today are crucial for scholarly careers as well as the reputation of departments and universities (e.g., Hudson and Laband 2013).

Four reasons are usually advanced why such rankings are deemed to be necessary. First, it is argued that because of the high specialization of research and the lack of expertise in areas that are different from the own research field, it is efficient to rely on research rankings. One should trust the "collective wisdom" of the community more than one's own judgment. Collective wisdom is taken to be summarized in research rankings based on publications and citations (Laband 2013). Second, research rankings fuel competition among scholars, lead to more and better research, and promote what is called an "entrepreneurial university" (Clark 1998; Stensaker and Benner 2013). Third, research rankings give the public a transparent picture of scholarly activity. They make scientific merits visible to people who have no special knowledge of the field like politicians, public officials, deans, university administrators, and journalists (see, e.g., Worrell 2009). Quantified criteria are useful, as they provide impersonal knowledge beyond the bounds of locality and power of insiders (Porter 1995, xi). They attack "club government," understood as closed professional self-regulation, old boys' networks, and unfounded claims to fame (Power 2005). Fourth, academic rankings make universities and departments more accountable for their use of public money. They help to allocate resources more efficiently (e.g., Research Assessment Exercise [RAE] 2008).

Based on these reasons, research rankings today are widely applied to take decisions on hiring, tenure, income of scholars, and allocation of resources. They shape the cognition and the activities even of those who criticize them harshly (Sauder and Espeland 2009). In many countries, recent reforms have linked scholars' salaries to the number of their publications in highly ranked journals. Scientists having published much in such journals are hired in order to raise publication and citation records. Such stars are highly paid though they often have little involvement with the respective university (Stephan 2008).

There are different kinds of journal rankings. Either there is a subjective peer approach, predicated on the view of experts about the quality of a

journal. Or there is an objective approach based on citation-based metrics.[1] The oldest and most popular one is the Journal Impact Factor (Garfield 1972). It is defined as the average number of citations of articles published in the journals in Thomson Reuter's Journal Citation Index over 2 years. It gives all citations equal weights. Other, more refined measures weight citations by the influence of the citing journal like the Eigenfactor score, or try to measure production combined with impact like the Hirsch index (Hirsch 2005). Such measures are also used to rank individual researchers.

All kinds of research rankings based on bibliometric measures rely on the following assumptions. First, the peer judgments on which research rankings are based—be it the quality of a journal or the quality of a single article—are of high quality. Second, the aggregation process of single peer judgments to metrics is correct and identifies good authors and journals. Third, the ranking of a journal in which an article is published is a suitable indicator of the quality of the article in question. Fourth, rankings do not produce unintended side effects with respect to individuals or institutions. Fifth, research rankings lead to a fruitful competition between researchers, fuel creativity, and promote better research.

Research rankings have come under scrutiny because these assumptions have been challenged. A lively discussion about the quality of research rankings is taking place (e.g., Butler 2007; Adler and Harzing 2009; Albers 2009). This discussion mostly focuses on the second aspect, in particular, the methods used to aggregate peer reviews to rankings and the tools available to improve them (e.g., Lane 2010; Hudson 2013; Laband 2013). The other aspects are considered only in a few cases. It is rarely asked whether rankings may produce unintended negative side effects, even if the indicators for research quality were perfect (e.g., Espeland and Sauder 2007; Osterloh 2010). Only recently discussions arise whether high-ranked journals really publish the best research. An interview with Randy Schekman, the 2013 winner of the Nobel Prize in Physiology or Medicine recently heated up the debate. He ferociously attacked high-impact journals (The Guardian 2013, see also Kleinert and Horton 2014). Whether there are alternatives to rankings to signal its quality remains open.

This article discusses two issues. In the first to fifth section, it is asked whether and to which extent the five assumptions upon which research rankings rest are realistic. Second, in the sixth and seventh section, we discuss whether there are alternatives to research rankings as a main instrument of research governance.

Are Peer Reviews Reliable?

All kinds of research rankings are piggybacked on peer reviews that are considered the founding stone of academic research evaluation. According to sociologists and economists (e.g., Nelson 1959; Arrow 1962; Merton 1973; Dasgupta and David 1994) in academic research, the evaluation by the market has to be substituted by the evaluation by peers who constitute the "republic of science" (Polanyi 1962). Success in academia is reflected by success in the market often only after a long delay or sometimes not at all (Bush 1945; Nelson 2004). In contrast, it is assumed that the quality of a piece of research is rapidly identified by peer reviews. As a consequence, the assessment of the value of an article as well as the quality of a journal in which articles are published depends on the judgment of peers.

However, the quality of peer reviews has come under scrutiny (e.g., Frey 2003; Bedeian 2004; Starbuck 2005, 2006; Tsang and Frey 2007; Gillies 2005, 2008; Abramo, Angelo, and Caprasecca 2009; Bornmann and Daniel 2009; Helbing and Balietti 2011). These studies' findings mainly consider peer reviews of scholarly articles. To our knowledge, there exist no empirical findings about peer-based assessments of journals. We suggest that the problems are comparable. Empirical findings about the quality of peer reviews of scholarly articles have disclosed the following problems (Campanario 1998a, 1998b; Osterloh 2010).

Low Interrater Reliability

Peer reviews differ considerably from each other. The judgments of two peers only correlate to a factor between 0.09 and 0.5 (Starbuck 2005). The correlation between reviewers' recommendations in clinical neuroscience "was little greater than would be expected by chance alone" (Rothwell and Martyn 2000, 1964). For rejected articles, the correlation is higher (Cicchetti 1991). Peer reviewers conform more on identifying bad than excellent scholarly contributions (Moed 2007). An empirical study on the acceptance of articles shows that luck of the referee draw plays a big role (Bornmann and Daniel 2009).

Low Prognostic Quality

Provided the quality of a contribution can be indicated by later citations, the correlation is also quite low; it is between .25 and .30 (Starbuck 2006,

83–84). Many articles published in "top" journals are rarely cited, which means that the reviewers do not judge the future impact satisfactorily (Hudson and Laband 2013). The percentage of "dry holes" (i.e., articles in refereed journals that have never been cited) in economic research during 1974–1996 has remained constant, although the resources to improve the screening of articles have risen (Laband and Tollison 2003).

Low Consistency Over Time

Highly ranked journals rejected many articles that thereafter received distinguished prizes, among them the Nobel Prize (Gans and Shepherd 1994; Campanario 1996; Lawrence 2003). On the other hand, Nobel Prize winning authors published articles in top journals that included severe mistakes discovered years later (Campanario 1998b).

Confirmation Bias

Referees score articles according to whether the results conform or conflict with their beliefs (Campanario 1998b). Methodological shortcomings are identified by referees in 71% of articles that contradict mainstream thinking but only in 25% of articles supporting orthodox views (Mahoney 1977). Articles threatening the previous work of reviewers tend to be rejected (Lawrence 2003). Selfish and rational reviewers have little interest to give good advice on how to improve an article. The result of a simulation study suggests that peer review is no better than a coin toss except if the share of rational, unreliable, and uninformed reviewers lies well below 30% (Thurner and Hanel 2011). As a consequence, unorthodox research has a small chance of being published, in particular if editors accept articles only if they are favorably reviewed by three or more reviewers (Campanario 1998b).

Editors Sometimes Make Serious Errors

The famous economist John Maynard Keynes rejected many innovative articles while he was editor of the *Economic Journal* (Campanario 1998b). The "Social Text" affair revealed that the physicist Alan D. Sokal published a parody in a (nonrefereed) special issue of the journal *Social Text*. The editors published it as a serious scholarly article not realizing that the article was a hoax (Sokal 1996).

Based on this evidence, the quality and credibility of peer reviews has become the subject of much scrutiny. An overview article on peer reviews states bluntly "Journal peer review is thus an unreliable quality control mechanism" (Campanario 1998b, 299). This might explain why quantifiable "objective" data have become popular, namely bibliometric measures based on articles and citations published in refereed journals. It is argued that by aggregation of independent judgments, individual reviewers' biases can be mitigated because they allow for error compensation, enable a broader perspective (e.g., Weingart 2005), and represent the collective wisdom (Laband 2013). This might be justified as long as the aggregation process of independent peer judges to rankings is correct.

Is the Aggregation Process of Peer Judges to Metrics Correct?

Several authors (e.g., Butler 2007; Donovan 2007; Adler and Harzing 2009) scrutinize the technical and methodological problems of the aggregation process and the usefulness of the metrics identified.

Technical errors occur when the scholars citing and cited are matched, leading to a loss of citations to a specific publication. For example, Thomson Reuters' Web of Knowledge is accused of having erroneous information (Monastersky 2005; Taylor, Perakakis, and Trachana 2008). It is unlikely that there is an equal distribution of the errors. Kotiaho, Tomkin, and Simmons (1999) found that names from unfamiliar languages lead to a geographical bias against non-English speaking countries. Small changes in measurement techniques and classifications can have considerable consequences for the position in rankings (Ursprung and Zimmer 2006; Frey and Rost 2010).

The methodological problems of constructing meaningful and consistent indices to measure scientific output have also been widely discussed (Lawrence 2003; Frey 2003, 2009; Adler and Harzing 2009).

First, there are selection problems. Usually, only journal articles are selected for incorporation in citation-based metrics, although books or proceedings may contribute considerably to scholarly work. Other difficulties include the low representation of small research fields, non-English articles, regional journals, and journals from other disciplines even if they are highly ranked in their respective disciplines. In addition, the role of older literature is not taken into account. Second, citations can have a supportive or negative meaning or merely reflect herding. According to the "Matthew effect," the probability of being cited is a function of previous

citations (Merton 1968). Simkin and Roychowdhury (2005) estimate that 70–90% of the articles cited had not been read properly by the person doing the citing. They base their estimation on the analysis of misprints in citations. Third, citations often are simply mistaken. Evans, Nadjari, and Burchell (1990) found that in surgery journals, 38% of all citations randomly examined proved to contain errors. Consequently, incorrect citations are endemic (Woelert 2013). Fourth, it is heavily discussed which metrics should be used to identify good scholarly contributions, authors, and journals. Researchers are advised on how to behave to maximize different metrics (e.g., Lalo and Mosseri 2009). Fifth, there are difficulties comparing numbers of publications, citations, and impact factors not only between disciplines but also between subdisciplines (Bornmann et al. 2008).

Technical and methodological problems can be mitigated, although it will take time and be expensive. For that reason, a temporary moratorium has been proposed "until more valid and reliable ways to assess scholarly contributions can be developed" (Adler and Harzing 2009, 72). However, there remain considerable problems with research rankings which would exist even if the aggregation process of peer reviews and citations to indicators worked perfectly and the metrics are meaningful to identify good research.

Is the Ranking of a Journal a Suitable Measure of the Quality of the Paper Published in This Journal?

Using the impact factor as a proxy for the quality of a particular journal or an article published in this journal is very common. The Institute for Scientific Information describes the impact Factor as "a systematic and objective means to critically evaluate the world's leading journals" (Baum 2011, 450). A publication in a scholarly journal with a high impact factor is taken to be "good." A publication in a journal with a low impact factor is considered to be unimportant. This interpretation has become internationally accepted (e.g., Abramo, Angelo, and Caprasecca 2009; Archambault and Larivière 2009; Jarwal, Brion, and King 2009). Today, in some countries, the number of publications in "high-impact journals" determines the distribution of public resources as well as the career of scholars (Hudson and Laband 2013). Some universities, for example, in Australia, China, and Korea, hand out cash bonuses for publications in top journals in order to raise their position in international rankings (Fuyuno and Cyranoski 2006; Franzoni, Scellato, and Stephan 2010). The Chinese Academy of

Sciences pays researchers who succeed in publishing an article in one of the top journals the equivalent of US$30,000 (The Guardian 2013).

However, using the impact factor of a journal (or the importance of a journal according to subjective peer evaluation) as a measure of the quality of a single article leads to substantial misclassification. This would be the case even if the technical and methodological problems of aggregation of citations did not exist or if the evaluation of the quality of a journal by peers would be reliable. Judging by the citations accumulated, many top articles are published in nontop journals, and many articles in top journals generate very few citations in management research (Starbuck 2005; Singh, Haddad, and Chow 2007), economics (Laband and Tollison 2003; Oswald 2007; Baum 2011), and science (Seglen 1997; Campbell 2008; Kriegeskorte 2012). A study of the *International Mathematical Union* states that the use of impact factors could be "breathtakingly naive" because it leads to large error probabilities (Adler, Ewing, and Taylor 2008, 14). Oswald's (2007) study includes data on citation intensity spanning over a quarter of a century. It thus takes into account that the citations may be strongly lagged. It turns out that many articles were never cited over the 25 years. In the most important journals, more than a third of the articles published are cited less than 20 times by other scholars over this longtime horizon. Baum (2011) compared the citations of articles in top management journals. He concludes: The "extreme variability in article citedness permits the vast majority of articles—and journals themselves—to free-ride on a small number of highly cited articles" (Baum 2011, 449).

Basic knowledge in statistics teaches us that when a distribution of data is highly skewed, it is unwarranted to deduce anything about a specific article. Nevertheless, this quality measure is widely used. Only recently it has been criticized in a broader context (The Guardian 2013). The chief editor of *Science*, Bruce Alberts (2013, 787) clearly stated in an editorial published in May 2013: "Such metrics ... block innovation." The "San Francisco Declaration on Research Assessment" (DORA 2012, 2) demanded in December 2012: "Do not use journal-based metrics, such as journal impact factors, as a surrogate measure of the quality of individual research articles, to assess an individual scientist's contribution, or in hiring, promotion or funding decisions."

Are Rankings Robust Concerning Reactivity?

When indicators become politically important, people change their behavior. So-called reactive measures (Campbell 1957) have an effect according

to the saying: "When a measure becomes a target, it ceases to be a good measure" (Strathern 1966, 4). This is in particular true if the measurement is not accepted voluntarily (Espeland and Sauder 2007). The effect takes place at the level of individuals and institutions.

Reactions by Individual Scholars

Reactivity on the level of individual scholars may take the form of goal displacement or counterstrategies to "beat the system." Both forms are aggravated by motivational reactions.

Goal displacement means that people maximize indicators that are easy to measure and disregard features that are hard to measure (Perrin 1998), a problem known in economics as the multiple-tasking effect (Holmstrom and Milgrom 1991; Ethiraj and Levinthal 2009). There is much evidence of this effect in laboratory experiments (Schweitzer, Ordonez, and Douma 2004; Fehr and Schmidt 2004; Ordonez et al. 2009)[2] and in the field. For example, public service teachers are responding with "teaching to the test" when they are assessed according to quotas of students who pass a certain exam (Nichols, Glass, and Berliner 2006; Heilig and Darling-Hammond 2008). In academia, "slicing strategy" is an example, whereby scholars divide their research into as many articles as possible in order to enlarge their publication list. Empirical field evidence from an Australian study has shown this to be so (Butler 2003). The mid-1990s saw a linking of the number of peer-reviewed publications to the funding of universities and individual scholars. The number of publications increased dramatically, but relative citation rates decreased. Using more refined measures like impact factors or the Hirsch index would help only in the short run. People would adapt soon to such measures as soon as they become relevant.

Counterstrategies are more difficult to observe. They consist in altering behavior itself to "game the system." Examples discussed in public service as reactions to evaluations from outside are chronically ill patients do no longer qualify for health care plans (Chen et al. 2011; Eijkenaar et al. 2013), teachers tell bad students not to participate in the tests (Figlio and Getzler 2002), or lower quality students are excluded from the measurement sample and put into special classes (Gioia and Corley 2002). Scholars distort their results to please, or at least not to oppose, possible reviewers. Bedeian (2003) found that no less than 25% of authors revised their articles following the suggestions of the referee, although they were aware that the change was not correct. To raise the chance of being accepted, authors refer

to possible reviewers, as the latter are prone to evaluate articles that cite their work more favorably (Lawrence 2003).[3] Frey (2003) calls this behavior "academic prostitution." Authors may be discouraged from undertaking and submitting novel and unorthodox research (Horrobin 1996; Prichard and Willmott 1997; Armstrong 1997; Gillies 2008; Alberts 2013).

The effects of goal displacement and counterstrategies are aggravated by motivational consequences of rankings: the decrease of intrinsically motivated curiosity. This kind of motivation is generally acknowledged to be of decisive importance in research (Spangenberg et al. 1990; Stephan 1996; Amabile 1998; Simonton 2004). In academia, a special motivation called taste for science exists (Merton 1973; Dasgupta and David 1994; Stephan 1996; Roach and Sauermann 2010). It is characterized by a relatively low importance of monetary incentives and a high importance of peer recognition and autonomy. People are attracted to research for which, at the margin, the autonomy to satisfy their curiosity and to gain peer recognition is more important than money (Bhagwat et al. 2004). These scholars are prepared to trade-off autonomy against money, as empirically documented by Stern (2004) and Roach and Sauermann (2010): Scientists pay to be scientists. The preference for the autonomy to choose one's own goals is important for innovative research in two respects. First, it leads to a useful self-selection effect of creative researchers.[4] Second, autonomy is the most important precondition for intrinsic motivation, which in turn is required for creative research (Amabile et al. 1996; Amabile 1998).

Rankings can negatively affect the motivation of scholars, in particular when high-ranking positions are linked to monetary rewards. In psychology and in psychological economics, considerable empirical evidence suggests that there is a crowding-out effect of intrinsic motivation by externally imposed goals. This is the case when goals are linked to incentives that do not give a supportive feedback and are perceived to be controlling (Frey 1992, 1997; Hennessey and Amabile 1998; Deci, Koestner, and Ryan 1999; Gneezy and Rustichini 2000; Gagné and Deci 2005; Falk and Kosfeld 2006; Ordonez et al. 2009).[5] Though so far there is no direct empirical evidence of a crowding-out effect by research rankings, numerous findings in other fields suggest that output-related incentives tend to crowd out intrinsically motivated curiosity. First, in contrast to qualitative peer reviews, rankings do not give supportive feedback because they do not tell scholars how to improve their research. Second, the content of research tends to lose importance. It is substituted for by extrinsic rewards (Lindenberg 2001; Ariely et al. 2009). That is, the taste for publication or

the ''taste for rankings'' replace the taste for science. Consequently, dysfunctional reactions like goal displacement and counterstrategies are enforced because they are not constrained by intrinsic preferences. The incentive to ''game the system'' in an instrumental way may get the upper hand (Frey, Homberg, and Osterloh 2013).

Reactivity at the Level of Institutions

Reactivity at the institutional level is closely related to Goodhart's (1975) Law in monetary policy and to the Lucas (1976) Critique in econometric modeling. It takes several forms. To the extent that rankings are used as a measure to allocate resources and positions, they create a lock-in effect. Even those scholars and academic institutions that are aware of the deficiencies of rankings do well not to oppose them. If they did, they would be accused not only of being afraid of competition but also of not contributing to the prestige and resources of their department or university. Therefore, it is a better strategy to play the game. For example, editors encourage authors to cite their respective journals in order to raise their impact factor (Garfield 1997; Smith 1997; Monastersky 2005). Universities engage scholars who have published in top journals in order to increase their ranking position. An example is the King Saud University in Riyadh that offers cash to highly cited researchers for adding the university's name to their research articles. In this way, the university has climbed from rank 2,910 in 2006 to rank 186 in 2010 in international research rankings (Bhattacharjee 2011). Such incentives risk crowding out the commitment to a scholarly discourse in favor of collecting publications and citations.

In addition, a negative *walling-off effect* sets in. Scholars are inclined to apply rankings to evaluate candidates in order to gain more resources for their research group or department. In addition, it is easier to count the publications and citations of colleagues than to evaluate the content of their scholarly contributions. Scholars thereby delegate their own judgment to the counting exercise behind rankings (Browman and Stergiou 2008; Laband 2013). This practice is defended by arguing that specialization in science has increased so much that even within disciplines it is impossible to evaluate the research in neighboring fields (van Fleet, McWilliams, and Siegel 2000; Swanson 2004). However, this practice in turn reinforces specialization and furthers a walling-off effect between disciplines and subdisciplines (Hudson 2013). By using output indicators instead of communicating on the content, the knowledge in the various fields becomes increasingly disconnected. This hampers the ability to create radical

innovations. Such innovations often cross the borders of disciplines (Amabile et al. 1996; Dogan 1999), presupposing a considerable overlap of knowledge between the disciplines (Schmickl and Kieser 2008).

Do Research Rankings Lead to a Fruitful Competition?

Rankings are by definition one-dimensional and thus tend to reduce diversity. In contrast to decentralized peer reviews, they press the heterogeneity and ambiguity of scholarly work into a simple order (Fase 2007). Such an order is easy to understand by the public similar to football leagues or hit parades. However, in contrast to such endeavors, scholarly work is characterized by controversial disputes that are essential for scientific progress. Rankings tend to suppress such disputes because they generate dominant views—not by disputes about the contents but by counting numbers. This contradicts the idea of research as institutionalized skepticism (Merton 1973). In contrast to rankings, peer reviews produce a great heterogeneity of scientific content and views. Heterogeneity fuels scholarly debates. This is the reason why some authors believe that the lack of reliability of peer reviews indicates solid science (e.g., Eckberg 1991; Campanario 1998a). Scholars with unorthodox views are less discouraged by negative peer reviews than by unfortunate rankings. If rejected by the reviewers of one journal, the referees of another equivalent journal might accept the article, in particular because luck plays a considerable role in having an article accepted (Bornmann and Daniel 2009). A scholar who unsuccessfully applies to one university often is successful when applying to another university with a similar reputation. Such diversity of scholarly endeavors is of special importance during radical innovations or paradigm shifts in the sense of Kuhn (1962).

The danger of reducing heterogeneity in research by centralized and hierarchical evaluation systems can be studied in the British Research Assessment Exercises:[6] The share of heterodox, not strictly neoclassical, economics sank drastically since the assessment of departments became based mainly on the ranking of journal publications. Heterodox journals have become less attractive for researchers due to the fact that they are less cited than mainstream journals (Lee 2007; see also Gioia and Corley 2002; Holcombe 2004). Moreover, the establishment of new research areas may have been inhibited. Research with uncertain outcomes has been discouraged in contrast to projects with quick payoffs. (Hargreaves Heap 2002). The problems of homogeneity are the greater, the more dominant research

rankings are, the more research assessments are centralized, and the more they are politically influential.

Suggestions to Overcome the Negative Consequences of Research Rankings as Instruments of Research Governance

There exist several proposals to mitigate the negative effects of research rankings. The first focuses on the lack of heterogeneity of rankings. The proposal is to use a number of rankings because their results differ markedly (e.g., Adler and Harzing 2009). This holds in particular as far as the ranking of individuals is concerned (Frey and Rost 2010). It may even be argued that the number of rankings should be augmented, so that each individual one loses its importance (Osterloh and Frey 2009). However, this proposal induces high costs not only on evaluators but also on the researchers being evaluated. Much energy and time would be consumed in reporting, negotiating, reframing, and presenting performance indicators. All of this distracts from the performance desired.

A second proposal aims at restricting the use of rankings to experts only. Nonexperts should not use rankings as ready-to-go indicators (van Raan 2005; Bornmann et al. 2008). Only experts who observe standards of good practice for the analysis, interpretation, and presentation of rankings should be allowed to employ rankings. However, it remains open how this suggestion could be enforced.

The third proposal suggests a combination of qualitative peer reviews and rankings or so-called informed peer reviews. It aims to balance advantages and disadvantages of peer reviews and rankings (Butler 2007; Moed 2007). ''Informed peer review'' might avoid the problem of too much homogeneity of evaluations. It may use indicators in an exploratory way (Frey, Homberg, and Osterloh 2013). However, it may also aggravate the problem if reviewers use citations or the impact factor of a journal as a proxy for the quality of an article. In the British Research Assessment Exercise, it became obvious that citations were the most important predictor of the evaluation outcomes by peers (Sgroi and Oswald 2013).

All three suggestions fail to provide a transparent and easy-to-understand picture of research quality. Moreover, they fail to avoid negative strategic and motivational reactions of scholars. Finally, they do not communicate that scholarly work has to be evaluated in a way including diversity and discourse, which are essential elements of scholarly research.

A Radical Suggestion for Research Governance

To overcome the problems discussed with respect to signaling quality of scholarly work, we refer to insights from managerial control theory (e.g., Ouchi 1979; Eisenhardt 1985; Osterloh 2010; Frey, Homburg, and Osterloh 2013). According to this approach, three different kinds of controls may be distinguished: output control, process control, and input control.

Output control is useful if well-defined unambiguous indicators are available to the evaluator. Such controls are attractive to nonexperts. However, as discussed, research rankings are far from delivering such unambiguous indicators.

Process control is useful when outputs are not easy to measure and to attribute, but when the controller has an appropriate knowledge of the transformation process of inputs into outputs. Process control is applicable only for peers who are familiar with the state of the art in the respective research field.[7] Peer reviews are the scholarly form of process control. However, though peer reviews are central to research, they have come under scrutiny and have recently been heavily criticized even in the popular press (see, e.g., *The Economist* 2013a, 2013b).

If neither output control nor process control works sufficiently well, input control has to be applied.[8] This kind of control is usually used when easy-to-measure outputs are not available, processes are not precisely observable, or peers are not able to evaluate the processes sufficiently. Input control is the main attribute of professions like life-tenured judges (e.g., Benz and Frey 2007; Posner 2010), medical doctors, teachers, and priests (Freidson 2001) characterized by a high information asymmetry between the professionals and the lays. To become a member of a profession, it is necessary to pass long-term education, selection, and self-selection, which ensure that one has deeply internalized professional norms as intrinsic values. Institutional rituals confirm these norms in order to signal that they have become part of the professional identity. The aim of the socialization and selection process is to induce professionals to follow these norms even if there are no controls and sanctions (Parsons 1968). Professionals are expected to resist the pressure and the persuasions of the market. At the same time, they are protected against competition by limiting market forces. This can be achieved, for example, by strict entrance qualifications or by providing professionals with basic resources without having to compete for them (Freidson 2001). Such a socialization and selection process is the precondition for professionals being granted a high degree of autonomy. It is only limited by professional norms comparable to judges' work.

Input control in the case of research governance means that aspiring scholars should be carefully socialized and selected by peers acting as role models. Scholars have to learn and to demonstrate that they master the state of the art, have preferences conforming to the taste for science (Merton 1973), and are able to direct themselves. This socialization and selection process may be supported by an open post-publication evaluation (Kriegeskorte 2012).[9] Those scholars with an ''entrance ticket'' to the republic of science can be granted much autonomy to foster their creativity and pursue their intrinsically motivated curiosity. Basic funds must be given in order to guarantee some degree of independence (Horrobin 1996; Gillies 2008).

The ''Principles Governing Research at Harvard'' state: ''The primary means for controlling the quality of scholarly activities of this Faculty is through the rigorous academic standards applied in selecting its members'' (p. 1).[10] This conforms to the idea of input control. Such control has empirically been shown to work also in Institutes for Advanced Studies and in research and development organizations of industrial companies (Abernethy and Brownell 1997). These observations are consistent with empirical research in psychological economics. They show that on average intrinsically motivated people do not shirk when they are given autonomy (Frey 1992; Gneezy and Rustichini 2000). Rather, they increase their commitment when they experience that they are trusted (Osterloh and Frey 2000; Falk and Kosfeld 2006; Frost, Osterloh, and Weibel 2010).

Input control has advantages and disadvantages. The *advantages first* consist in reducing the inefficient ranking games. They induce young scholars to learn the professional standards of their discipline assisted by their peers. Second, input control applies during limited time periods and occasions only. It takes into account that peer control is problematic for assessing academic quality. Third, input control decentralizes peer evaluation, for example, when submitting articles or applying for jobs. The heterogeneity of scholarly views central to the scientific communication process is maintained.

The *disadvantages* consist first in the danger that some scholars who have passed the selection might misuse their autonomy, reduce their work effort, and waste funds. This disadvantage is reduced when the selection process is conducted rigorously and entails the capability and intrinsic motivation of the scholars to direct themselves. Second, input control may lead to groupthink and cronyism (Janis 1972). This danger is reduced if the diversity of scholarly approaches within the relevant peer group is enhanced. Third, input control is a kind of informed peer review. It risks relying too much on publications in top journals or citations as a proxy for

an article's quality. This danger can be avoided if these indicators are used in an exploratory way only.[11] Fourth, the public as well as university administrators do not get an easy-to-understand indicator of scholarly activities. People outside the scholarly community have to acknowledge that it is difficult to evaluate academic research. Only scholars can fulfill this task in a serious way. As a result, university presidents, vice chancellors, and deans should consist of accomplished scholars and not of managers. Scholars have a good understanding of the research process. Goodall (2009) shows that for a panel of 55 British research universities, a university's research performance[12] is improved after an accomplished scholar has been hired as president. Fifth, peers have to judge for their own—or have to be silent about the quality of a piece of research, which they cannot evaluate themselves. However, to do so is more an advantage than a disadvantage. It avoids what can be seen as a "fatal conceit" (Hayek 1991).

To compensate for the disadvantages of input control, two measures are worth considering. The first measure consists in periodic self-evaluations including external evaluators. The major goal is to induce self-reflection and feedback among the members of a research unit. The second measure compensates to a certain extent for the limited visibility of input control to the public. Awards like prizes and titles, as well as different kinds of professorships and fellowships (from assistant to distinguished), signal the recognition of peers to nonexperts (Frey and Osterloh 2010; Frey and Gallus 2013). This procedure leads to an overall evaluation avoiding the manipulation of particular metrics (Frey and Neckermann 2008). Empirical evidence suggests that the two measures do not crowd out intrinsic motivation (Chan et al. 2013). They match motivational preconditions of the taste for science, above all consisting in peer recognition and appreciation of autonomy (Stern 2004; Roach and Sauermann 2010). They motivate even those who do not win such an award.[13]

Conclusions

Research rankings based on citations and publications in top journals dominate modern research governance. They shape careers, resource allocation, and reputation by introducing a quasi-market into a field characterized by notorious market failure. The result is a distorted competition. Such research rankings do not suitably measure performance. They neither give the public nor scholars from other fields a picture of scholarly quality. They do not really make universities more accountable for the resources used but foster

"gaming the system." Nor do they support creative or interdisciplinary work. Instead, they suppress diversity and deviations from the mainstream.

We find, first, that peer reviews on which rankings are based are not reliable. Second, the aggregation process of peer judgment cannot compensate for the lack of reliability of peer reviews. Third, the impact of a journal should not be taken as a proxy for the quality of an article published in this journal, even if the aggregation process was correct. Fourth, rankings trigger unintended negative side effects. In particular, they crowd out intrinsically motivated curiosity and commitment to the values of the republic of science. Fifth, they destroy diversity and controversial disputes, which are the founding stones of progress in research.

The peer review system has many faults. Nevertheless, research needs members of the republic of science who really are peers and engage in the scholarly discourse. This goal can be accomplished by input control in the form of socialization and selection following transparent rules. If input control fails, it cannot be compensated for by output control, for example, research rankings. Therefore, accountability to the public does not mean to provide numbers comparable to hit parades. Rather, it means being responsible for correct and fair processes during the selection process. The approach suggested is in line with Kay's (2010) message that most of our goals are best achieved when we approach them indirectly.

No return to the old system of academic oligarchy or the old boys' network will occur, provided the heterogeneity of scholarly approaches is maintained and the rules and procedures of input control are enacted with diligence and transparency.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. For an overview and comparison of the various measures of journal quality, see Hudson (2013); Helbing and Balietti (2011).
2. Locke and Latham (2009) provide counterevidence to Ordonez et al. (2009). In their view, goal setting has no negative effects. They disregard, however, that

within an organization, goal setting may well be suitable for simple but not for complex tasks. The latter case is discussed in Ethiraj and Levinthal (2009).

3. Such problems of sabotage in tournaments have been extensively discussed in personnel economics, see Lazear and Shaw (2007).
4. The importance of low monetary incentives for a selection of intrinsically motivated employees is discussed in Lazear and Shaw (2007).
5. The crowding-out effect sometimes is contested, for example, by Eisenberger and Cameron (1996). However, the empirical evidence for complex tasks and actors intrinsically motivated in the first place is strong (Deci, Koestner, and Ryan 1999; Weibel, Rost, and Osterloh 2010). For a survey of the empirical evidence, see Frey and Jegen (2001).
6. The new version is called Research Assessment Framework (RAF).
7. Process control can be exerted also in a mechanistic, bureaucratic way. However, the rules to be followed must be determined and supervised by the professional experts (Freidson 2001).
8. Ouchi (1979) calls this kind of control "clan control."
9. While we appreciate the idea of an open post-publication peer evaluation that Kriegeskorte (2012) suggests, we disagree with his suggestion for ratings.
10. See http://www.fas.harvard.edu/research/greybook/principles.html (accessed January 8, 2014).
11. The British RAF requires members of the peer review panels not to use information on impact factors. Nevertheless, citation data are provided, see Sgroi and Oswald (2013).
12. Measured according to the score the university has achieved in the British Research Assessment Exercise.
13. The money attached to awards is less important than the reputation of the award-giving institution; see Frey and Neckermann (2008).

References

Abernethy, M. A., and P. Brownell. 1997. "Management Control Systems in Research and Development Organizations: The Role of Accounting, Behavior and Personnel Controls." *Accounting, Organizations and Society* 22:233–48.

Abramo, G. D., C. A. Angelo, and A. Caprasecca. 2009. "Allocative Efficiency in Public Research Funding: Can Bibliometrics Help?" *Research Policy* 38: 206–15.

Adler, R., J. Ewing, and P. Taylor. 2008. "Citation Statistics." 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). *Statistical Science* 21:1–14.

Adler, N. J., and A.-W. Harzing. 2009. "When Knowledge Wins: Transcending the Sense and Nonsense of Academic Rankings." *Academy of Management Learning* 8:72–95.

Albers, S. 2009. "Misleading Rankings of Research in Business." *German Economic Review* 3:352–63.

Alberts, B. 2013. "Editorial: Impact Factor Distortions." *Science* 340:787.

Amabile, T. 1998. "How to Kill Creativity." *Harvard Business Review* 76:76–87.

Amabile, T., R. Conti, H. Coon, J. Lazenby, and M. Herron. 1996. "Assessing the Work Environment for Creativity." *Academy of Management Journal* 39: 1154–84.

Archambault, É., and V. Larivière. 2009. "History of the Journal Impact Factor: Contingencies and Consequences." *Scientometrics* 79:639–53.

Ariely, D., U. Gneezy, G. Loewenstein, and N. Mazar. 2009. "Large Stakes and Big Mistakes." *Review of Economic Studies* 76:451–69.

Armstrong, J. S. 1997. "Peer Review for Journals: Evidence on Quality Control, Fairness, and Innovation." *Science and Engineering Ethics* 3:63–84.

Arrow, K. 1962. "Economic Welfare and the Allocation of Resources for Invention." In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, edited by R. Nelson, 609–26. Princeton, NJ: Princeton University Press.

Baum, J. A. C. 2011. "Free-riding on Power Laws: Questioning the Validity of the Impact Factor as a Measure of Research Quality in Organization Studies." *Organization* 18:449–66.

Bedeian, A. G. 2003. "The Manuscript Review Process: The Proper Roles of Authors, Referees, and Editors." *Journal of Management Inquiry* 12:331–38.

Bedeian, A. G. 2004. "Peer Review and the Social Construction of Knowledge in the Management Discipline." *Academy of Management Learning and Education* 3: 198–216.

Benz, M., and B. S. Frey. 2007. "Corporate Governance: What Can We Learn from Public Governance?" *Academy of Management Review* 32:92–104.

Bhagwat, J. G., S. Ondategui-Parra, K. H. Zou, A. Gogate, L. A. Intriere, P. Kelly, S. E. Seltze, and P. R. Ros. 2004. "Motivation and Compensation in Academic Radiology." *Journal of the American College of Radiology* 1:493–96.

Bhattacharjee, Y. 2011. "Saudi Universities Offer Cash in Exchange for Academic Prestige." *Science* 334:1344–45.

Bornmann, L., and H. D. Daniel. 2009. "The Luck of the Referee Draw: The Effect of Exchanging Reviews." *Learned Publishing* 22:117–25.

Bornmann, L., R. Mutz, C. Neuhaus, and H. D. Daniel. 2008. "Citation Counts for Research Evaluation: Standards of Good Practice for Analyzing Bibliometric Data and Presenting and Interpreting Results." *Ethics in Science and Environmental Politics* 8:93–102.

Browman, H. I., and K. I. Stergiou. 2008. "Factors and Indices Are One Thing, Deciding Who Is Scholarly, Why They Are Scholarly, and the Relative Value of Their Scholarship Is Something Else Entirely." *Ethics in Science and Environmental Politics* 8:1–3.

Bush, V. 1945. *Science: The Endless Frontier*. Report to the president by Vannevar Bush, Director of the Office of Scientific Research and Development. Washington, DC: US Government Printing Office.

Butler, L. 2003. "Explaining Australia's Increased Share of ISI Publications—The Effects of a Funding Formula Based on Publication Counts." *Research Policy* 32:143–55.

Butler, L. 2007. "Assessing University Research: A Plea for a Balanced Approach." *Science and Public Policy* 34:565–74.

Campanario, J. M. 1996. "Using Citation Classics to Study the Incidence of Serendipity in Scientific Discovery." *Scientometrics* 37:3–24.

Campanario, J. M. 1998a. "Peer Review for Journals As It Stands Today, Part 1." *Science Communication* 19:181–211.

Campanario, J. M. 1998b. "Peer Review for Journals As It Stands Today, Part 2." *Science Communication* 19:277–306.

Campbell, D. T. 1957. "Factors Relevant to the Validity of Experiments in Social Settings." *Psychological Bulletin* 54:297–312.

Campbell, P. 2008. "Escape from the Impact Factor." *Ethics in Science and Environmental Politics* 8:5–7.

Chan, H. F., B. Frey, J. Gallus, and B. Torgler. 2013. "Does The John Bates Clark Medal Boost Subsequent Productivity and Citation Success?" *CREMA Working Paper* 2013-02. Accessed March 11, 2012. http://www.crema-research.ch/papers/2013-02.pdf.

Chen, T., K. Chung, I. Lin, and M. Lai. 2011. "The Unintended Consequences of Diabetes Mellitus Pay for Performance (P4P) Program in Taiwan: Are Patients with More Comorbidities of More Severe Conditions Likely To Be Excluded from the P4P Program?" *Health Services Research* 46:47–60.

Cicchetti, D. V. 1991. "The Reliability of Peer Review for Manuscript and Grant Submissions: A Cross-disciplinary Investigation." *Behavioral and Brain Sciences* 14:119–35.

Clark, Burton R. 1998. *Creating Entrepreneurial Universities: Organizational Pathways of Transformation*. Oxford, England: Pergamon Press.

Dasgupta, P., and P. A. David. 1994. "Toward a New Economics of Science." *Research Policy* 23:487–521.

Deci, E. L., R. Koestner, and R. M. Ryan. 1999. "A Meta-analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation." *Psychological Bulletin* 125:627–68.

Dogan, M. 1999. "Marginality." *Encyclopedia of Creativity* 2:179–84.

Donovan, C. 2007. "The Qualitative Future of Research Evaluation." *Science and Public Policy* 34:585–97.

DORA (San Francisco Declaration on Research Assessment). 2012. Accessed December 16, 2012. http://am.ascb.org/dora/files/SFDeclarationFINAL.pdf.

Eckberg, D. L. 1991. "When Nonreliability Indicates Solid Science." *Behavioral and Brain Science* 14:145–46.

Eijkenaar, F., M. Emmert, M. Scheppach, and O. Schöffski. 2013. "Effects of Pay for Performance in Healthcare: A Systematic Review of Systematic Reviews." *Health Policy* 110:115–30.

Eisenberger, R., and J. Cameron. 1996. "Detrimental Effects of Reward: Reality or Myth?" *American Psychologist* 51:1153–66.

Eisenhardt, K. M. 1985. "Control: Organizational and Economic Approaches." *Management Science* 31:134–49.

Espeland, W. N., and M. Sauder. 2007. "Rankings and Reactivity: How Public Measures Recreate Social Worlds." *American Journal of Sociology* 113:1–40.

Ethiraj, S. K., and D. Levinthal. 2009. "Hoping for A to Z while Rewarding Only A: Complex Organizations and Multiple Goals." *Organization Science* 20:4–21.

Evans, J. T., H. I. Nadjari, and S. A. Burchell. 1990. "Quotational and Reference Accuracy in Surgical Journals. A Continuing Peer Review Problem." *JAMA: Journal of the American Medical Association* 263:1353–54.

Falk, A., and M. Kosfeld. 2006. "The Hidden Cost of Control." *American Economic Review* 96:1611–30.

Fase, M. M. G. 2007. "For Examples of a Trompe-l'Oeil in Economics." *De Economist* 155:221–38.

Fehr, E., and K. M. Schmidt. 2004. "Fairness and Incentives in a Multi-task Principal-agent Model." *Scandinavian Journal of Economics* 106:453–74.

Figlio, D. N., and L. S. Getzler. 2002. "Accountability, Ability and Disability: Gaming the System." *NBER Working Paper No. W9307*. Accessed July 2, 2010. http://bear.warrington.ufl.edu/figlio/w9307.pdf.

Franzoni, C., G. Scellato, and P. Stephan. 2010. *Changing Incentives to Publish and the Consequences for Submission Patterns*. Working Paper. Torino, Italy: IP Finance Institute.

Freidson, E. 2001. *Professionalism. The Third Logic*. Chicago, IL: University of Chicago Press.

Frey, B. S. 1992. "Tertium Datur: Pricing, Regulating and Intrinsic Motivation." *Kyklos* 45:161–85.

Frey, B. S. 1997. *Not Just for the Money: An Economic Theory of Personal Motivation*. Cheltenham, England: Edward Elgar.

Frey, B. S. 2003. "Publishing as Prostitution?—Choosing between One's Own Ideas and Academic Success." *Public Choice* 116:205–23.

Frey, B. S. 2009. "Economists in the PITS." *International Review of Economics* 56:335–46.

Frey, B. S., and J. Gallus. 2013. "Awards Are a Special Kind of Signal." CREMA Working Paper 2014-04. Accessed March 11, 2014. http://www.crema-research.ch/papers/2014-04.pdf.

Frey, B. S., F. Homberg, and M. Osterloh. 2013. "Organizational Control Systems and Pay-for-performance in the Public Service." *Organization Studies* 34: 949–72.

Frey, B. S., and R. Jegen. 2001. "Motivation Crowding Theory." *Journal of Economic Surveys* 15:589–611.

Frey, B. S., and S. Neckermann. 2008. "Awards—A View from Psychological Economics." *Journal of Psychology* 216:198–208.

Frey, B. S., and M. Osterloh. 2010. "Motivate People with Prizes." *Nature* 465:871.

Frey, B. S., and K. Rost. 2010. "Do Rankings Reflect Research Quality?" *Journal of Applied Economics* XIII:1–38.

Frost, J., M. Osterloh, and A. Weibel. 2010. "Governing Knowledge Work: Transactional and Transformational Solutions." *Organizational Dynamics* 39: 126–36.

Fuyuno, I., and D. Cyranoski. 2006. "Cash for Papers: Putting a Premium on Publication." *Nature* 441:792.

Gagné, M., and E. L. Deci. 2005. "Self–determination Theory and Work Motivation." *Journal of Organizational Behavior* 26:331–62.

Gans, J. S., and G. B. Shepherd. 1994. "How Are the Mighty Fallen: Rejected Classic Articles by Leading Economists." *Journal of Economic Perspectives* 8:165–79.

Garfield, E. 1972. "Citation Analysis as a Tool in Journal Evaluation." *Science* 178: 471–79.

Garfield, E. 1997. "Editors are Justified in Asking Authors to Cite Equivalent References from Same Journal." *British Medical Journal* 314:1765.

Gillies, D. 2005. "Hempelian and Kuhnian Approaches in the Philosophy of Medicine: The Semmelweis Case." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:159–81.

Gillies, D. 2008. *How Should Research Be Organised?* London, England: College Publication King's College.

Gioia, D. A., and K. G. Corley. 2002. "Being Good versus Looking Good: Business School Rankings and the Circean Transformation from Substance to Image." *Academy of Management Learning and Education* 1:107–20.

Gneezy, U., and A. Rustichini. 2000. "Pay Enough or Don't Pay at All." *Quarterly Journal of Economics* 115:791–810.

Goodall, A. H. 2009. "Highly Cited Leaders and the Performance of Research Universities." *Research Policy* 38:1070–92.

Goodhart, C. 1975. "Monetary Relationships: A New Form of Threadneedle Street." *Papers in Monetary Economics 1*, Reserve Bank of Australia.

Hargreaves Heap, S. P. 2002. "Making British Universities Accountable." In *Science Bought and Sold: Essays in the Economics of Science*, edited by P. Mirowski and E.-M. Sent, 387–411. Chicago, IL: University of Chicago Press.

Hayek, F. A. 1991. *The Fatal Conceit: The Errors of Socialism*. Chicago, IL: The University of Chicago Press.

Heilig, J. V., and L. Darling-Hammond. 2008. "Accountability Texas-style: The Progress and Learning of Urban Minority Students in a High-stakes Testing Context." *Educational Evaluation and Policy Analysis* 30:75–110.

Helbing, D., and S. Balietti. 2011. *How to Create an Innovation Accelerator*. Accessed March 11, 2014. http://arxiv.org/abs/1011.3794v3.

Hennessey, B. A., and T. M. Amabile. 1998. "Reward, Intrinsic Motivation and Creativity." *American Psychologist* 53:647–75.

Hirsch, J. E. 2005. "An Index to Quantify an Individual's Scientific Research Output." *Proceedings of the National Academy of Sciences* 102:16569–72.

Holcombe, R. G. 2004. "The National Research Council Ranking of Research Universities: Its Impact on Research in Economics." *Econ Journal Watch* 1: 498–514.

Holmstrom, B. P., and P. Milgrom. 1991. "Multitask Principal-agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization* 7:24–52.

Horrobin, D. F. 1996. "Peer Review of Grant Applications: A Harbinger for Mediocrity in Clinical Research?" *Lancet* 348:293–95.

Hudson, J. 2013. "Ranking Journals." *The Economic Journal* 123:F202–22.

Hudson, J., and D. N. Laband. 2013. "Using and Interpreting Journal Rankings: Introduction." *The Economic Journal* 123:F199–201.

Janis, I. L. 1972. *Victims of Groupthink: A Psychological Study of Foreign-policy Decisions and Fiascos*. Boston, MA: Houghton Mifflin.

Jarwal, S. D., A. M. Brion, and M. L. King. 2009. "Measuring Research Quality Using the Journal Impact Factor, Citations and 'Ranked Journals': Blunt Instruments or Inspired Metrics?" *Journal of Higher Education Policy and Management* 31:289–300.

Kay, J. 2010. *Obliquity. Why Our Goals Are Best Achieved Indirectly*. London, England: Profile Books.

Kleinert, S., and R. Horton. 2014. "How Should Medical Science Change." *The Lancet* 383:197–98. Accessed March 11, 2014. http://dx.doi.org/10.1016/S0140-6736(13)62678-1.

Kotiaho, J. S., J. L. Tomkin, and L. W. Simmons. 1999. "Unfamiliar Citations Breed Mistakes." *Nature* 400:307.

Kriegeskorte, N. 2012. "Open Evaluation: A Vision for Entirely Transparent Post-publication Peer Review and Rating for Science." *Frontiers in Computational Neuroscience* 6:1–18.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Laband, D. N. 2013. "On the Use and Abuse of Economics Journal Rankings." *The Economic Journal* 123:F223–54.

Laband, D. N., and R. D. Tollison. 2003. "Dry Holes in Economic Research." *Kyklos* 56:161–74.

Lalo, F., and R. Mosseri. 2009. "Bibliometric Evaluation of Individual Researchers: Not Even Right. Not Even Wrong!" *Europhysics News* 40:26–29.

Lane, J. 2010. "'Let's Make Science Metrics More Scientific." *Nature* 464:488–89.

Lawrence, P. A. 2003. "The Politics of Publication—Authors, Reviewers, and Editors Must Act to Protect the Quality of Research." *Nature* 422:259–61.

Lazear, E. P., and K. L. Shaw. 2007. "Personnel Economics: The Economist's View of Human Resources." *Journal of Economic Perspectives* 21:91–114.

Lee, F. S. 2007. "The Research Assessment Exercise, the State and the Dominance of Mainstream Economics in British Universities." *Cambridge Journal of Economics* 31:309–25.

Lindenberg, S. 2001. "Intrinsic Motivation in a New Light." *Kyklos* 54:317–42.

Locke, E. A., and G. P. Latham. 2009. "Has Goal Setting Gone Wild, or Have Its Attackers Abandoned Good Scholarship?" *Academy of Management Perspectives* 21:17–23.

Lucas, R. E. 1976. "Econometric Policy Evaluation: A Critique." In *Carnegie-Rochester Conference Series on Public Policy. The Phillips Curve and Labor Markets*, edited by K. Brunner and A. H. Meltzer, 19–46. New York: North Holland.

Mahoney, M. J. 1977. "Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System." *Cognitive Therapy Research* 1:161–75.

Maier-Leibnitz, H. 1989. "The Measurement of Quality and Reputation in the World of Learning." *Minerva* 27:483–504.

Merton, R. K. 1968. "The Matthew Effect in Science." *Science* 159:56–63.

Merton, R. K. 1973. *The Sociology of Science: Theoretical and Empirical Investigation*. Chicago, IL: University of Chicago Press.

Moed, H. F. 2007. "The Future of Research Evaluation Rests with an Intelligent Combination of Advanced Metrics and Transparent Peer Review." *Science and Public Policy* 34:575–83.

Monastersky, R. 2005. "The Number That's Devouring Science." *Chronicle of Higher Education* 52:A12.

Nelson, R. R. 1959. "The Simple Economics of Basic Scientific Research." *Journal of Political Economy* 67:297–306.

Nelson, R. R. 2004. "The Market Economy, and the Scientific Commons." *Research Policy* 33:455–71.

Nichols, S. L., G. V. Glass, and D. C. Berliner. 2006. "High-stakes Testing and Student Achievement: Does Accountability Pressure Increase Student Learning? " *Education Policy Analysis Archives* 14. Accessed February 15, 2014. http://epaa.asu.edu/epaa/v14n11/.

Ordonez, L. D., M. E. Schweitzer, A. D. Galinsky, and M. H. Bazerman. 2009. "Goals Gone Wild: The Systematic Side Effects of Overprescribing Goal Setting." *Academy of Management Perspectives* 23:6–16.

Osterloh, M. 2010. "Governance by Numbers. Does It Really Work in Research?" *Analyse und Kritik* 32:267–83.

Osterloh, M., and B. S. Frey. 2000. "Motivation, Knowledge Transfer, and Organizational Forms." *Organization Science* 11:538–50.

Osterloh, M., and B. S. Frey. 2009. "Are More and Better Indicators the Solution? Comment to William Starbuck." *Scandinavian Journal of Management* 25: 225–27.

Oswald, A. J. 2007. "An Examination of the Reliability of Prestigious Scholarly Journals: Evidence and Implications for Decision-makers." *Economica* 74: 21–31.

Ouchi, W. G. 1979. "A Conceptual Framework for the Design of Organizational Control Mechanisms." *Management Science* 25:833–48.

Parsons, T. 1968. "Professions." *International Encyclopedia of the Social Sciences* 12:536–47.

Perrin, B. 1998. "Effective use and misuse of performance measurement." *American Journal of Evaluation*, 19: 367–79.

Polanyi, M. 1962. "The Republic of Science: Its Political and Economic Theory." *Minerva* 1:54–73. Reprinted in Polanyi, M. 1969. *From Knowing and Being*, 49–72. Chicago, IL: University of Chicago Press. Re-reprinted in Mirowski, P., and E. M. Sent, eds. 2002. *Science Bought and Sold. Essays in the Economics of Science*, 465–85. Chicago, IL: The University of Chicago Press.

Porter, Theodore M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Posner, R. A. 2010. "From the New Institutional Economics to Organization Economics: With Applications to Corporate Governance, Government Agencies, and Legal Institutions." *Journal of Institutional Economics* 6:1–37.

Power, Michael. 2005. "The Theory of the Audit Explosion." In *The Oxford Handbook of Public Management*, edited by E. Ferlie, L. E. Lynn, and C. Pollitt, 326–44. Oxford, England: Oxford University Press.

Prichard, C., and H. Willmott. 1997. "Just How Managed is the McUniversity?" *Organization Studies* 18:287–316.

RAE (Research Assessment Exercise). 2008. Accessed March 11, 2014. http://www.rae.ac.uk/.

Roach, M., and H. Sauermann. 2010. "A Taste for Science? PhD Scientists' Academic Orientation and Self-selection into Research Career in Industry." *Research Policy* 39:422–34.

Rothwell, P. M., and C. N. Martyn. 2000. "Reproducibility of Peer Review in Clinical Neuroscience. Is Agreement between Reviewers Any Greater than Would Be Expected by Chance Alone?" *Brain* 123:1964–69.

Sauder, M., and W. N. Espeland. 2009. " The Discipline of Rankings: Tight Coupling and Organizational Change." *American Sociological Review* 74:63–82.

Schmickl, C., and A. Kieser. 2008. "How Much Do Specialists Have to Learn from Each Other When They Jointly Develop Radical product Innovations?" *Research Policy* 37:473–49.

Schweitzer, M. E., L. Ordonez, and B. Douma. 2004. "Goal Setting as a Motivator of Unethical Behavior." *Academy of Management Journal* 47:422–32.

Seglen, P. O. 1997. "Why the Impact Factor of Journals Should Not Be Used for Evaluating Research." *British Medical Journal* 314:498–502.

Sgroi, D., and A. J. Oswald. 2013. "How Should Peer-review Panels Behave?" *The Economic Journal* 123:F255–78.

Simkin, M. V., and V. P. Roychowdhury. 2005. "Copied Citations Create Renowned Papers?" *Annals Improbable Research* 11:24–27.

Simonton, D. K. 2004. *Creativity in Science. Chance, Logic, Genius, and Zeitgeist*. Cambridge, England: Cambridge University Press.

Singh, G., K. M. Haddad, and S. Chow. 2007. "Are Articles in "Top" Management Journals Necessarily of Higher Quality?" *Journal of Management Inquiry* 16: 319–31.

Smith, R. 1997. "Journal Accused of Manipulating Impact Factor." *British Medical Journal* 314:463.

Sokal, A. D. 1996. "A Physicist Experiments with Cultural Studies." *Lingua Franca*, May/June, pp. 62–64. Accessed March 11, 2014. http://www.physics.nyu.edu/faculty/sokal/lingua_franca_v4/lingua_franca_v4.html.

Spangenberg, J. F. A., R. Starmans, Y. W. Bally, B. Breemhaar, and F. J. N. Nijhuis. 1990. "Prediction of Scientific Performance in Clinical Medicine." *Research Policy* 19:239–55.

Starbuck, W. H. 2005. "How Much Better Are the Most Prestigious Journals? The Statistics of Academic Publication." *Organization Science* 16:180–200.

Starbuck, W. H. 2006. *The Production of Knowledge. The Challenge of Social Science Research*. Oxford, England: Oxford University Press.

Stensaker, B., and M. Benner. 2013. "Doomed to be Entrepreneurial: Institutional Transformation or Institutional Lock-ins of "New Universities?" *Minerva* 51: 399–416.

Stephan, P. E. 1996. "The Economics of Science." *Journal of Economic Literature* 34:1199–235.

Stephan, P. E. 2008. "Science and the University: Challenges for Future Research." *CESifo Economic Studies* 54:313–24.

Stern, S. 2004. "Do Scientists Pay To Be Scientists?" *Management Science* 50: 835–53.

Strathern, M. 1966. "From Improvement to Enhancement: An Anthropological Comment on the Audit Culture." *Cambridge Anthropology* 19:1–21.

Swanson, E. 2004. "Publishing in the Majors: A Comparison of Accounting, Finance, Management, and Marketing." *Contemporary Accounting Research* 21:223–25.

Taylor, M., P. Perakakis, and V. Trachana. 2008. "The Siege of Science." *Ethics in Science and Environmental Politics* 8:17–40.

The Economist. 2013a. *How Science goes Wrong*. October 19–25, 11.

The Economist. 2013b. *Trouble at the Lab. Scientists Like to Think of Science as Self-correcting. To an Alarming Degree, It Is Not*. October 19–25, 21–24.

The Guardian. 2013. *Nobel Winner Declares Boycott of Top Science Journals*. December 9. http://www.theguardian.com/science/2013/dec/09/nobel-winner-boycott-science-journals.

Thurner, S., and R. Hanel. 2011. "Peer-review in a World with Rational Scientists: Towards Selection of the Average." *The European Physical Journal B* 84:707–11.

Tsang, E. W. K., and B. S. Frey. 2007. "The As-is Journal Review Process: Let Authors Own Their Ideas." *Academy of Management Learning and Education* 6:128–36.

Ursprung, H. W., and M. Zimmer. 2006. "Who is the "Platz–Hirsch" of the German Economics Profession? A Citation Analysis." *Jahrbücher für Nationalökonomie und Statistik* 227:187–202.

van Fleet, D., A. McWilliams, and D. S. Siegel. 2000. "A Theoretical and Empirical Analysis of Journal Rankings: The Case of Formal Lists." *Journal of Management* 26:839–61.

van Raan, A. F. J. 2005. "Fatal Attraction: Conceptual and Methodological Problems in the Rankings of Universities by Bibliometric Methods." *Scientometrics* 62:133–43.

Weibel, A., K. Rost, and M. Osterloh. 2010. "Pay for Performance in the Public Sector–Benefits and (Hidden) Costs." *Journal of Public Administration Research and Theory* 20:387–412.

Weingart, P. 2005. "Impact of Bibliometrics upon the Science System: Inadverted Consequences?" *Scientometrics* 62:117–31.

Whitley, Richard. 2011. "Changing Governance and Authority Relations in the Public Sciences." *Minerva* 49:359–85.

Woelert, P. 2013. "The Economy of Memory": Publications, Citations, and the Paradox of Effective Research Governance." *Minerva* 51:341–62.

Worrell, D. 2009. "Assessing Business Scholarship: The Difficulties in Moving beyond the Rigor-relevance Paradigm Trap." *Academy of Management Learning* 8:127–30.

Author Biographies

Margit Osterloh is a professor (em.) at the University of Zurich, Switzerland professor at the Zeppelin University Friedrichshafen, Germany, and research director at CREMA (Center for Research in Economics, Management and the Arts), Switzerland. Her research areas are Organization, Corporate Governance, and Research Governance.

Bruno S. Frey is a professor at the Zeppelin University Friedrichshafen, Germany and research director at CREMA (Center for Research in Economics, Management, and the Arts), Switzerland. He was formerly a distinguished professor of Behavioral Science at the University of Warwick. His research areas are political economy and behavioral economics.