Rankings Games

Bruno S. Frey and Margit Osterloh

August 2011
Rankings Games

by

Bruno S. Frey and Margit Osterloh

Abstract

Research rankings based on publications and citations today dominate governance of academia. Yet they have unintended side effects on individual scholars and academic institutions and can be counterproductive. They induce a substitution of the “taste for science” by a “taste for publication”. We suggest as alternatives careful selection and socialization of scholars, supplemented by periodic self-evaluations and awards. Neither should rankings be a basis for the distributions of funds within universities. Rather, qualified individual scholars should be supported by basic funds to be able to engage in new and unconventional research topics and methods.

(94 words)

Keywords: Academic governance, rankings, motivation, selection, socialization.

JEL Classification: A10, D02, H83, L23, M50

Corresponding Author: Bruno S. Frey, University of Zurich, Wilfriedstrasse 6, CH-8032 Zurich, Switzerland. Email-address: bruno.frey@econ.uzh.ch and University of Warwick, UK.
Today, academic rankings based on publications and citations dominate research governance in academia. Moreover, they serve as the basis for assessing the performance and impact of scholars, faculties, and universities for three purposes.

First, they are widely used to make decisions on the hiring, tenure, and income of scholars. In many countries, recent reforms have increasingly linked scholars’ salaries to the number of publications they have in international journals. Some universities, for example, in Australia, China, and Korea, provide cash bonuses for publications in key journals in order to raise their position in international and national rankings (Fuyuno and Cyranoski, 2006; Franzoni et al., 2010). The assumption is that such measures lead to more and better publications in highly ranked journals.

Second, academic rankings supposedly give the public a transparent picture of scholarly activity. A common view of academic rankings is that they make scientific merits visible to politicians, public officials, deans, university administrators, and journalists, people who have no special knowledge of the field (see e.g. Worrell, 2009).

Third, academic rankings make universities more accountable for their use of public money. They may help to allocate resources more efficiently according to indicators that measure past performance in an objective way.

However, in recent times, academic rankings have come under scrutiny. A lively discussion about the quality of academic rankings is taking place (e.g. Butler, 2007; Adler and Harzing, 2009; Albers, 2009). This discussion focuses on the method used to determine academic rankings and the tools available to improve it. That more and better indicators are needed to enhance the quality of rankings is taken as a given (e.g. Starbuck 2009; Lane 2010). Only in a few cases is it asked whether rankings may produce unintended negative side effects, even if the indicators for research quality are
perfect (e.g. Espeland and Sauder, 2007; Osterloh, 2010). Consequently, whether there are viable alternatives to academic rankings as an instrument for academic governance remains open.

This paper discusses two issues. We first evaluate the advantages and disadvantages of rankings, in particular their unintended consequences. Secondly, we ask whether there are alternatives to academic rankings as the main instrument of academic governance.

1. **Advantages of Rankings**

Academic rankings are based on peer reviews, which 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, which constitutes 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, 2006). In contrast, with peer reviews, the quality of a piece of research is rapidly identified.

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 et al., 2009; Bornmann and Daniel 2009). As such, empirical findings have disclosed the following problems.

1. The extent to which reviewing reports conform to each other is low. The correlation between the judgments of two peers falls between 0.09 and 0.5 (Starbuck, 2005). In clinical neuroscience, it was found that the correlations among reviewers’ recommendations ‘was little greater than would be expected by chance alone’ (Rothwell and Martyn, 2000, p.1964). It is important that the correlation is higher for papers
rejected than for papers accepted (Cicchetti, 1991). This means that peer reviewers are better able to identify academic low performers than excellent research (Moed, 2007).

2. The correlation of a particular reviewer’s evaluation with quality, as measured by later citations of the manuscript reviewed, is also quite low; it lies between 0.25 and 0.30 (Starbuck, 2006, pp. 83–4).

3. Documentation exists showing that highly ranked journals rejected many papers that later were awarded distinguished prizes, including the Nobel Prize (Gans and Shepherd, 1994; Campanario, 1996; Lawrence, 2003).

4. Fourth, reviewers find methodological shortcomings in 71% percent of papers that contradict mainstream thinking compared to only 25% that support the mainstream (Mahoney, 1977).

5. In spite of the fact that resources to improve the screening of papers have risen substantially, it has been shown that the percentage of ‘dry holes’ (i.e. articles in refereed journals that have never been cited) in economic research during 1974 to 1996 has remained constant (Laband and Tollison, 2003).

6. It is sometimes not in the interest of irrational or selfish reviewers to accept certain research or to give advice on how to improve it. These reviewers might reject papers that threaten their previous work (Lawrence, 2003), or that draw attention to competing ideas. In a simulation, the results showed that unless the number of rational reviewers (as well as the number of unreliable and uninformed reviewers) is kept well below 30 percent, peer review is no better than a coin toss (Thurner and Hanel, 2010).

7. In addition, editors sometimes are prone to serious errors. A famous example is the ‘Social Text’ affair: The physicist Alan D. Sokal published an article in a (non-refereed) special issue of the journal Social Text, which he wrote as a parody. The editors did not realize that the article was a hoax and published it as a serious scholarly article (Sokal, 1996).
For these reasons, the quality and credibility of peer reviews has become the subject of much scrutiny. As a result, rankings have become popular. Compared to qualitative peer reviews, they promise considerable advantages (e.g. Abramo et al., 2009).

1. Rankings are based on more than three or four evaluations typical of qualitative approaches. Through statistical aggregation, individual reviewers’ biases may be balanced.

2. The influence of the old boys’ network may be avoided. An instrument is provided to dismantle unfounded claims to fame. Rankings can serve as fruitful, exogenous shocks to some schools and make them care more about the reactions of the public (Khurana, 2007, p.337).

3. The data to construct a ranking are easily available from publications lists or data sources like the Web of Science.

4. Rankings are cheaper than pure qualitative reviews, at least in terms of time. They admit updates and rapid inter-temporal comparisons.

5. Rankings promise to facilitate the comparison between a large number of scholars or institutions.

6. They promise to give non-experts like research administrators, politicians, journalists, and students an easy-to-use device to evaluate the standing of the research.

   Two fortunate consequences arise. First, politicians and administrators who assess scholarly quality only from outside may get an objective measure to allocate resources. Second, public attention for research outcomes is raised, which enhances the willingness to spend money.

2. Technical and Methodological Problems with Rankings

   Though rankings may counterbalance some problems of qualitative peer reviews, in recent times, it has become clear that rankings have disadvantages of their
own (Butler, 2007; Donovan, 2007; Adler et al., 2008; Adler and Harzing, 2009). Until now, mainly technical and methodological problems were highlighted concerning the aggregation process of publications and citations as well as the use of impact factors of journals (van Raan, 2005).

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 et al., 2008). It is unlikely that there is an equal distribution of the errors. Kotiaho et al. (1999) found that names from unfamiliar languages lead to a geographical bias against non-English speaking countries. It also has been shown that small changes in measurement techniques and classifications can have large effects on 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 been widely discussed recently (Lawrence, 2002, 2003; Frey, 2003, 2009; Adler et al., 2008; Adler and Harzing, 2009).

- First, there are selection problems. Usually only journal articles are selected for incorporation in the rankings, although books or proceedings may contribute considerably to scholarly work. Other difficulties include the low representation of small research fields, non-English papers, 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) estimated that 70% to 90% of the papers cited had not been read by the person doing the citing. Consequently, incorrect citations are endemic. Third, using the impact factor of a journal as a proxy for the quality of a particular article leads
to substantial misclassification. Many top articles are published in non-top tier journals, and many articles in top tier journals generate very few citations in management research (Starbuck, 2005; Singh et al., 2007), economics (Laband and Tollison, 2003; Oswald, 2007), and science (Seglen, 1997; Campbell, 2008). A study of the International Mathematical Union concluded that the use of impact factors could be ‘breathtakingly naïve’ because it leads to large error probabilities (Adler et al., 2008, p.14). Fourth, there are difficulties comparing citations and impact factors between disciplines and even between sub-disciplines (Bornman et al., 2008).

Such 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, p.72). However, there remain important unintended negative side effects of rankings (lack of heterogeneity and unintended behavioural and motivational side effects), which would exist even if the rankings did work perfectly.

2.1. Unintended Side Effect of Rankings: Lack of Heterogeneity

All kinds of rankings are by definition one-dimensional. They press the multifacetedness, heterogeneity, and ambiguity of scholarly endeavours into a simple order (Fase, 2007). Such an order is easy to understand by the public like football leagues or hit parades. However, in contrast to such endeavours, scholarly work is characterized by controversial disputes, which 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 (Heintz, 2008). This contradicts the idea of research as institutionalized scepticism (Merton, 1973). In contrast to rankings, peer reviews, though they have many shortcomings, produce a great heterogeneity of scientific content and views. This heterogeneity fuels scholarly debates. Moreover, peer reviews discourage scholars with unorthodox views less than
rankings. If rejected by the reviewers of one journal, the reviewers of another equivalent journal might accept the article; a student who unsuccessfully applies to one university often is successful when applying to another university with a similar reputation. Mostly overlooked are the disadvantages of centralized and one-dimensional rankings as compared to the decentralized evaluation process by peer reviews. It may be of special importance during radical innovations or paradigm shifts according to Kuhn (1962), as one-dimensional rankings serve as the motivation to invest in improving one’s position in the rankings instead of improving the intended performance. This has two more unintended side effects that are interrelated: strategic behavioural reactions and the dampening of intrinsically motivated curiosity.

2.2. *Unintended Side Effect of Rankings: Behavioural Reactions*

Unintended behavioural reactions become stronger as more indicators become politically important – and this is surely the case for academic rankings today. They consist of the so-called reactive measures (Campbell, 1957), which are caused by the fact that people change their behaviour strategically in reaction to being observed or measured, in particular if the measurement is not accepted voluntarily (Espeland and Sauder, 2007). Reactivity threatens the validity of measures according to the saying: ‘When a measure becomes a target, it ceases to be a good measure’ (Strathern, 1966, p.4). This effect is closely related to Goodhart’s Law (1975) in monetary policy and to the Lucas Critique (1976) in econometric modelling. It takes place at the level of individual scholars as well as at the level of institutions.

2.2.1. *Reactions by individual scholars*

Reactivity on the level of individual scholars may take the form of goal displacement or counterstrategies to ‘beat the system’. Goal displacement means that people maximize indicators that are easy to measure and disregard features that are hard to measure (Merton, 1940; 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 et al., 2004; Fehr and Schmidt, 2004; Ordonez et al., 2009).\(^1\) In academia, examples can be found (e.g. the ‘slicing strategy’) whereby scholars divide their research results into a ‘least publishable unit’ by breaking the results into as many papers 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 the quality as measured by relative citation rates decreased. Further, there is evidence showing that academics with the highest score in publication rankings score only modestly in a ranking based on their contributions in editorial boards (Rost and Frey, 2011).

Counterstrategies are more difficult to observe. They consist of altering behaviour itself to ‘game the system’. Examples discussed in public service as reactions to evaluations from outside are: chronically ill patients are excluded in healthcare, teachers exclude bad students from tests (for empirical evidence in the United States, see Figlio and Getzler, 2002), or lower quality students are put in special classes that are not included in the measurement sample (Gioia and Corley, 2002). In academia, examples include scholars who distort their results to please, or at least not to oppose, prospective referees. Bedeian (2003) found evidence that no less than 25% of authors

\(^1\) Locke and Latham (2009) in a rejoinder provide counterevidence to Ordonez et al. (2009). They argue that goal setting has no negative effects. However, they disregard that goal setting may well work for simple but not for complex tasks within an organization. For the latter case, see Ethiraj and Levinthal (2009).
revised their manuscripts according to the suggestions of the referee although they knew that the change was incorrect. To be accepted, authors cite possible reviewers because the latter are prone to judge papers that approvingly cite their work more favourably (Lawrence, 2003, p.260). Frey (2003) calls this behaviour 'academic prostitution'. Even worse, to be accepted in refereed journals and to meet the expectations of their peers – many of whom consist of mainstream scholars – authors may be discouraged from conducting and submitting creative and unorthodox research (Horrobin, 1996; Prichard and Willmott, 1997; Armstrong, 1997; Gillies, 2008).

2.2.2. Reactions by institutions

Reactivity on the institutional level takes several forms. To the extent that rankings are used as a measure to allocate resources and positions, they create a lock-in effect. One cannot controvert this effect at the level of individuals or single institutions. 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 follow the rules and to play the game. For example, in several countries, highly cited scientists are hired in order to raise publication and citation records. Such stars are highly paid although they often have little involvement with the respective university (Brook, 2003; Stephan, 2008). Another example is editors who encourage authors to cite their respective journals in order to raise their impact rankings (Garfield, 1997; Smith, 1997; Monastersky, 2005).

2 Such problems of sabotage in tournaments have been extensively discussed in personnel economics (see Lazear and Shaw, 2007).
A negative walling-off effect sets in. Scholars themselves 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). 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 neighbouring fields (van Fleet et al., 2000; Swanson, 2004). However, this practice in turn reinforces specialization and furthers a walling-off effect between disciplines and sub-disciplines. By using output indicators instead of communicating on the contents, the knowledge in the various fields becomes increasingly disconnected. This hampers the ability to create radical innovations that often cross disciplinary borders (Amabile et al., 1996; Dogan, 1999).

Research is in danger of being increasingly homogenized. Research endeavours tend to lose the diversity that is necessary for a creative research environment. For economics, Great Britain provides an example: The share of heterodox, not strictly neoclassical, economics sank drastically since the ranking of departments became based mainly on citation counts. Heterodox journals have become less attractive for researchers due to their smaller impact factor when compared to mainstream journals (Lee, 2007; see also Gioia and Corley, 2002; Holcombe, 2004).

The establishment of new research areas is inhibited. In Great Britain, the Research Assessment Exercise has discouraged research with uncertain outcomes, and it has encouraged projects with quick payoffs (Hargreaves Heap, 2002).

2.3. Unintended Side Effect of Rankings: Motivational Reactions

The behavioural effects of reactivity are enforced by the motivational consequences of the dependence on rankings: the decrease of intrinsically motivated
curiosity. This kind of motivation generally is acknowledged to be of decisive importance in research (Spangenberg et al., 1990; Stephan, 1996; Amabile, 1996, 1998; Simonton, 2004). In academia, a special motivation called ‘taste for science’ exists (Merton, 1973; Dasgupta and David, 1994; Stephan, 1996; Sorenson and Fleming, 2004; Roach and Sauermann, 2010), which 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. They value the possibility of following their own scientific goals more than financial rewards (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 ways. Firstly, it leads to a useful self-selection effect of creative researchers. Secondly, autonomy is the most important precondition for intrinsic motivation, which in turn is required for creative research (Amabile et al., 1996; Amabile, 1998; Mudambi et al., 2007).

Rankings can negatively affect the motivation of scholars, in particular when high-ranking positions are promised monetary rewards. They may crowd out intrinsically motivated curiosity. In psychology and behavioural or 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 et al., 1999; Gneezy

---

3 For the importance of low monetary incentives for a selection of intrinsically motivated employees, see Lazear and Shaw (2007).
and Rustichini, 2000; Gagné and Deci, 2005; Falk and Kosfeld, 2006; Ordonez et al., 2009). Though until now, there is no direct empirical evidence on a crowding-out effect by rankings, numerous findings in other fields suggest that rankings 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 is in danger of losing importance. It is substituted by extrinsic incentives (Kruglansky, 1975; Lindenberg 2001). That is, the ‘taste for publication’ replaces the ‘taste for science’. Consequently, dysfunctional reactions like goal displacement and counterstrategies are enforced because they are not constrained by intrinsic preferences. The inducement to ‘game the system’ in an instrumental way may get the upper hand.

3. Proposals Made to Overcome the Unintended Side Effects of Rankings

There exist several proposals to overcome the unintended side effect of rankings. The first one takes aim at the lack of heterogeneity of rankings. The suggestion is to use a number of rankings because their results differ markedly (e.g. Adler and Harzing, 2009), in particular with respect to the ranking of individuals (Frey and Rost, 2010). There even could be an argument that the number of rankings should be augmented to the extent that each individual one loses its importance (Osterloh and Frey, 2009).

The second proposal aims at the informed use of the rankings by experts. It has been argued that non-experts should not use rankings as ready-to-go indicators (van

---

4 The crowding-out effect sometimes is contested, for example, by Eisenberger and Cameron (1996), Gerhart and Rynes (2003), Locke and Latham (2009). However, the empirical evidence for complex tasks and actors intrinsically motivated in the first place is strong (Deci et al., 1999; Weibel et al., 2010). For a survey of the empirical evidence, see Frey and Jegen (2001).
Raan, 2005; Bornmann et al., 2008). Therefore, the development of standards of good practice for the analysis, interpretation, and presentation of rankings is critical.

The third proposal suggests a combination of qualitative peer reviews and rankings, or so-called ‘informed peer reviews’. To do this, the advantages and the disadvantages of the two ranking methods must be balanced so that the rankings can be put into context (Butler, 2007; Moed, 2007).

All three proposals to some extent may mitigate the problems associated with rankings. However, first they induce high costs not only on the side of the evaluators (in particular in the case of the first proposal) but also on the side of those being evaluated. This will cause a trade-off between accountability and performance (Bouckaert and Peters, 2002; Dubnik, 2005). Instead of improving performance through accountability, too much energy and time is being consumed in reporting, negotiating, reframing, and presenting performance indicators, all of which distracts from the performance that is desired. Second, the responsible use of rankings as a handy instrument for politicians, public officials, administrators, journalists, and other non-experts is considerably reduced. Applying these proposals means that an accountability paradox arises: The more clear-cut indicators are used to make universities accountable to the public, the less accountability will actually occur. Well applied accountability of universities towards the public means to communicate that scholarly work has to be evaluated in a way taking into account that diversity and discourse are essential elements of scholarly research.

4. **Policy Implications for Academic Governance**

To overcome the problems discussed, we refer to insights from managerial control theory (e.g., Ouchi, 1977, 1979; Eisenhardt, 1985; Schreyögg and Steinmann, 1987). 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 non-experts. However, as discussed, 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 cause-effect relationships or 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. If neither output control nor process control works sufficiently well because of the complexity or the ambiguity of the tasks being evaluated, input control has to be applied.\textsuperscript{5} This kind of control is usually applied when easy to measure outputs are not available or processes are not precisely observable. Input control is based on careful selection, socialization, and placement of the candidates. It intends to make sure that individuals have internalized norms and to maintain professional standards. Input control takes place inside professional groups, such as life-tenured judges (e.g. Benz and Frey, 2007; Posner, 2010). Once a candidate has passed the input control that candidate becomes a member of a profession. Autonomy is curtailed only by professional norms that are confirmed by institutionalized rituals.

In the case of research governance, input control means that aspiring scholars should be carefully socialized and selected by peers to prove that they have mastered the state of the art, have preferences according to the ‘taste for science’ (Merton, 1973), and are able to direct themselves. Those who have an ‘entrance ticket’ to the republic of science after having passed a rigorous input control can be given much autonomy to foster their creativity and intrinsically motivated curiosity. This includes the provision of basic funds to provide a certain degree of independence after having passed the entrance barriers (Gillies, 2008; Horrobin, 1996).

\textsuperscript{5} Ouchi (1979) calls this kind of control ‘clan control’.
Input control is part of the ‘Principles Governing Research at Harvard’, which 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’. Input control has empirically proven to be successful also in R&D organizations of industrial companies (Abernethy and Brownell, 1997). This is in accordance with empirical findings in psychological economics, which show that on average intrinsically motivated people do not shirk when given autonomy (Frey, 1992; Gneezy and Rustichini, 2000; Fong and Tosi, 2007). Instead, they raise their efforts when they perceive that they are trusted (Osterloh and Frey, 2000; Falk and Kosfeld, 2006; Frost et al., 2010).

Input control has advantages and disadvantages. The advantages first consist of downplaying the unfortunate ranking games while inducing young scholars to learn the professional standards of their discipline under the assistance of peers. Second, although input control still requires process and output control in the form of informed peer evaluations, this applies during limited time periods, namely during situations of status passage. Thus, input control draws consequences from the fact that peer control, as well as rankings, are problematic in assessing academic quality. Third, input control is a decentralized form of peer evaluation, for example, when submitting papers or applying for jobs. It supports the heterogeneity of scholarly views central to the scientific communication process. Fourth, input control is a kind of ‘informed peer review’ that is able to use output indicators in an exploratory way.

---

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 their funds. This disadvantage is lowered when the selection process is conducted rigorously. Second, input control is in danger of being submitted to groupthink and cronyism (Janis, 1972). This danger can be mitigated by fostering the diversity of scholarly approaches within the relevant peer group. Third, the public as well as university administrators do not have an easy to comprehend picture of scholarly activities, as is the case for output control based on rankings. People outside the scholarly community have to acknowledge that evaluating scholarly activities is a particularly difficult task.

As a result, accountability to the public has a different meaning than providing clear-cut output indicators. Instead, the criteria of admission for becoming a scholarly member of the ‘republic of science’ have to be communicated and reported thoroughly. The quality of a university depends on how carefully and rigidly it follows the criteria of selection and socialization. For example, a broad and international selection of reviewing peers is necessary in order to avoid cronyism. This procedure corresponds to the characteristics of an acceptable legal process, where the judges make decisions according to transparent and comprehensible rules, although the content of a decision is sometimes hard to follow.

To compensate for the disadvantages of input control, two measures are worth considering that use elements of both output control and process control. The first

---

7 As a result, universities leaders like presidents, vice chancellors, and deans should consist of accomplished scholars. In contrast to pure managers, top scholars have a better understanding of the research process. Goodall (2009) showed that for a panel of 55 research universities, a university’s research performance is improved after an accomplished scholar has been hired as president.
measure consists of periodic self-evaluations including external evaluators. Here, 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 non-experts (Frey and Osterloh, 2010), thus providing an overall evaluation, which avoids any manipulation of particular metrics (Frey and Neckermann, 2008).

As the empirical evidence shows, both measures, though being partly extrinsic motivators, do not crowd out intrinsic motivation (Neckermann et al., 2010). They match motivational preconditions of the ‘taste of science’, which consist in the first place of peer recognition and the granting of autonomy (Stern, 2004; Roach and Sauermann, 2010). They motivate well even for those who do not actually win such an award.\footnote{The money attached to awards is less important than the reputation of the award giving institution, see Frey and Neckermann (2008).}

5. Conclusions

Academic rankings have major disadvantages that tend to be disregarded or downplayed both in the literature and in practice. We argue that input control in the form of rigorous selection and socialization should play a major role in research governance, supplemented by periodic self-evaluation and awards. Input control consists of informed peer reviews, which encompasses rankings but only during a limited time of a scholar’s career. Therefore, input control draws consequences from the fact that peer reviews as well as rankings are highly problematic when evaluating academic quality. The main disadvantage of input control is that the public does not receive an easy to comprehend picture of scholarly activity. However, it should be part
of the accountability to the public of research organizations to communicate that scholarly work has to be evaluated in a special way.

Our discussion also suggests that rankings should not be a basis for the distribution of research funds within universities. In addition, the ability to acquire outside funds should not be considered a valuable output measure (as it presently is in many countries). Instead, the ‘taste for science’ of the individual researchers should be supported by giving every qualified researcher basic funds. This allows the researcher to engage in new, unconventional, and so far not yet accepted research topics and methods. Outside funds then can be used to supplement basic funds without any pressure to adapt to ‘normal science’.

All this does not signal a return to the old system of academic oligarchy or the old boys’ network as long as the heterogeneity of scholarly approaches is maintained and the rules and procedures of input control are conducted with utmost care. Failures with the selection of members of the ‘republic of science’ cannot be compensated by rankings.

Bruno S. Frey is a Distinguished Professor of Behavioural Science, Warwick Business School, University of Warwick, and Professor of Economics, University of Zurich. Margit Osterloh is a Professor of Management, Warwick Business School, University of Warwick and Professor of Business Administration (em.) at the University of Zürich
References


Adler, R. Ewing, J. and Taylor, P. (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).


Browman, H.I. and Stergiou, K.I. (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*, vol. 8, pp. 1–3.


