In: Pio Baake and Rainald Borck (eds.) (2007)
Public Economics and Public Choice
Berlin and Heidelberg: Springer, 181-192.

10 The Rankings and Evaluations Mania

Bruno S. Frey[95]

University of Zurich,
CREMA – Center for Research in Economics, Management and the Arts

Beat Blankart is a quite extraordinary scholar. He has always pursued the kind of research he himself found important and has been perturbed remarkably little by current fads in his chosen field. He is a critical economist in the best sense, sometimes even a little whimsical – in any case he is far from being a run-of-the-mill economist. I therefore hope that he will agree with at least some of the ideas developed in this paper.

10.1 The Market and the Public Spheres

In recent years it has become a matter of course to introduce performance measurement in the public sector as a substitute for the market mechanism. Indeed, most people consider it absolutely *inevitable* and a logical consequence of pursuing a higher level of rationality in the public sector.

Yet this conclusion is bizarre in view of the fact that exactly those activities tend to be allocated to the public where output, or performance, is difficult to measure. The market does not work ("market failure"), or at least does not work particularly well, in the public sphere when elements of public goods, external effects and badly measurable output are dominant. It therefore is an odd idea to introduce output controls to the public sector. This seems to be warranted only when the government (wrongly) is en-

[95] I am grateful for the many helpful remarks by Margit Osterloh, Christine Benesch, Simon Lüchinger and Susanne Neckermann, and to Isabel Ellenberger for improving the English.

gaged in activities that could be performed by the market equally well, or even better. But in the genuinely public areas output control by its very nature does not work in a satisfactory way. For an alternative one has to turn to away from focusing solely on output and has to consider process and input controls.

Despite those fundamental theoretical problems universities and other academic institutions in German-speaking countries and beyond have introduced, or were forced to introduce, *rankings* and, even more broadly, *evaluations* of their activities. Rankings are part of evaluations, but cover many additional aspects. The negative consequences discussed here mainly refer to evaluations but some also to rankings. Currently, they are considered the *ne plus ultra* of any "rational" way of running such institutions, without considering any alternative whatsoever.

The flood of *rankings* is well visible for economics undertaken in Germany, Austria and Switzerland. One of the first ones was Bommer and Ursprung (1998), Eichenberger and Frey (2000), the rankings by the *Centrum für Hochschulentwicklung CHE* (Berghoff et al. 2002), and more recently the *Handelsblatt* ranking (September 18, 2006). Many of these rankings received considerable media attention and shape the perception of the general public and of political decision makers. And then there are the international rankings in which economists of German-speaking countries are listed such as the many different rankings published in the "Symposium on Evaluating Economics Research in Europe" published in the *Journal of the European Economic Association* in December 2003 and in the *RePEc* (www.repec.org) which every month presents rankings of 10,592 registered authors according to a large number of different criteria. The *Deutsche Wissenschaftsrat* took the next step and intends to establish a "super"-ranking for each discipline, sanctioned by its high prestige and official position. Presently, rankings are developed for sociology and chemistry, but it seems quite certain that such an effort will be expanded to all major subjects.

It cannot be denied that such rankings have *some* positive aspects. They are reasonably valid in the sense that the same scholars and institutions regularly are at the top of the list. But if this is really the case, what is the use of constantly repeating the exercise? The results provide little, if any, new information.

Another positive aspect may be that the results can act like a shock and may induce scholars and institutions to increase their efforts to undertake good research. But it is, of course, well known that such a shock evaporates rather quickly. The people concerned quickly get used to being positioned in rankings. Even more importantly, they quickly learn to react to them. In particular, they find ways and means to discount a bad ranking by attributing it to causes beyond their control. Once that has been achieved, rankings do not have much effect, if any, on performance rather life goes on as usual. Also, one has to consider whether the same positive effects on performance could not be reached by different, and more sustainable, measures. I will argue at the end of the paper that competition among scientific institutions and a careful selection of scholars are much more effective in improving performance.

There is also a surge of *evaluations* that flood academic institutions. Evaluations are understood here to be assessments for governments of *past performance by outside experts.* They are broader than rankings (but rankings are an essential part of evaluations) and more directly addressed to policy issues, most importantly the allocation of public resources. Today, evaluations are ubiquitous and are undertaken in ever shorter time intervals. Lately, *continuous* evaluations have become the craze. As a result we shall soon arrive at the point where every scholar, and every academic institution is evaluated all the time. Accordingly, a significant amount of material resources, manpower, attention and effort are invested by both the evaluators and the evaluatees. The latter have less and less time to do research, but rather have to spend more and more time to prepare for the time-consuming evaluations.

As in the case of ratings, evaluations may have a temporary beneficial shock effect. However, as evaluations increasingly become a normal part of a scholar's life, the shock tends to be overcome quickly. Also, it might lead to a "Hawthorne Effect" as individual scholars and academic institutions feel themselves attended to which may give them a sense of purpose and importance.

One of the main goals of evaluations in academia is a more efficient allocation of public funds: those institutions that are doing well are to receive more financial support, while those not doing well are to be given less funds or repudiated altogether. This may sound reasonable but is nevertheless a mistake. What has to be evaluated, of course, are the *marginal* effects of additional or reduced funds. It is well possible that a high ranked institution will not further improve its performance when receiving more resources. In some cases, for instance when the optimal size has been transgressed, performance may even weaken. Conversely, an institution ranked poorly may profit much from additional resources. People engaged in the by now sizeable "evaluation industry" will, of course, argue that they consider the expected changes in performance induced by a change in

funds. But where are the cases in which funds were taken from well-rated academic institutions and given to badly-ranked institutions based on the expected marginal effects (rather than on purely political reasons)?

I wish to argue that the *noxious effects* of rankings and evaluations are sizeable and that they tend to be *overlooked* and therefore these activities are undertaken *too often and in too large an extent.* I focus on aspects directly relevant for economics[96]. I don't want to discuss here the well-known shortcomings of publication and citation rankings such as whether all authors (or only the first author) are included; what kind of publications are considered (only narrowly defined economics journals or also publications in adjacent disciplines, publications in books etc); what language is counted (today normally only English, thus totally disregarding all the other languages in the world, including those spoken by far more people); how a particular academic institution is defined[97]; and what period is counted (life-time achievement or only the last few years or even months). Rather, I want to discuss some of the most important behavioural reactions to evaluations.

There is a wide spectrum of reactions induced by evaluations. It is often overlooked that these reactions do not pertain to evaluations as such but only occur if evaluations have important repercussions for the persons and institutions evaluated in terms of financial support and prospects for the future. As long as rankings and evaluations did not have many, or any, consequences academics considered them, at best, with some amusement, or often with outright scorn. With the rising importance of rankings and evaluations this has changed dramatically. It has become impossible *not* to participate in these exercises. If a scholar or institution did refuse, it would be charged of being afraid and in any case would quickly lose its academic status as it no longer appeared in the rankings.

It is useful to distinguish between the reactions of particular scholars and those of academic institutions but it can generally be said that many of the

[96] More general analyses are undertaken in Frey and Osterlohn (2006) and in Frey (2007a, b).

[97] This aspect should not be neglected, at least if one considers revealed behaviour. In the case of the Handelsblatt, the University of Munich ranked better than the University of Bonn. Hurt in their pride the Bonn economists decided that a Max Planck Institute was part of the University of Bonn so putting them ahead of the Munich economists. These then rightly argued that they could include the Ifo-Institute...One can well imagine further steps in this upward spiral. This is just one example of what behaviour is induced by evaluation exercises.

reactions generally neglected are *unproductive* from the point of view of scientific research.

10.2 Economists Evaluated

The reactions to a ranking or to an evaluation are strongly *asymmetric.* The consequences of these exercises for academia are therefore necessarily *distorted.* Persons faring well will have less incentive to react; they may simply enjoy their success. The situation is totally different for academics coming out unfavourably. They can resort to the following behaviours:

a) The results can be *put into doubt* and therewith the results *defined away.* There are virtually hundreds of arguments that prove how a particular ranking or evaluation is imperfect. Everyone who has only the slightest understanding of the ranking and evaluation techniques knows that they are subject to a large number of dubious assumptions and calculations. While the academics who have been badly ranked and evaluated may perhaps not be the greatest scholars, they certainly do have the capacity to pick on these shortcomings in ranking system. It may even be argued that they develop a special knowledge in that defensive activity as they can afford to do little else. But this is exactly what scientific research should not be about.

b) The rankings and evaluations may be *manipulated.* There are well known techniques on how to jack up the number of publications and citations. It is, for instance, not an accident that the number of persons given as the "authors" of a particular paper has strongly increased over the last few years. Decades ago, one author for a paper was the rule. Today two, or rather three authors have become normal, and the first papers with four and more authors have appeared. Of course, such a development can always be justified by reasons of content, but it is nevertheless remarkable that it is consonant with the effort to do well in rankings and evaluations[98]. Another reaction is to publish the same content with minor variations in several journals, and to break down the content to the smallest publishable unit.

[98] I am of course well aware that most current rankings take into account the number of authors. Notwithstanding, it is still better to be one of the co-authors than not to be an author at all.

Again, from the social point of view such efforts are unproductive and have nothing to do with producing good research.

c) *Political rent seeking* activities are undertaken in order to mitigate or to reverse the foreseeable damaging consequences of rankings and evaluations. Again, these activities do not contribute to advancing scientific knowledge but are "directly unproductive, profit-seeking (DUP) activities" (Bhagwati 1982).

d) Time and effort are redirected to other activities within academia such as *administrative and bureaucratic tasks*. If this led to more productive academics having more time available for teaching and research, it would be potentially beneficial. Alas, it is only too well known that all too often the result is an increase in bureaucracy affecting *all* members of an academic department in which case this reaction leads to an unproductive outcome.

e) The badly-ranked department members react by *actively seeking to block* the activities of its well-ranked members. This envy driven unproductive response is not unheard of in academic departments of German-speaking academic institutions.

f) The department members who perceive themselves to be unfairly ranked and evaluated respond by lapsing into *mental resignation* – while still occupying their positions and receiving their wages.

Several of these unproductive reactions to rankings and evaluations are not relevant if scholars who have been badly-ranked and evaluated can be forced to leave their positions. However, in most academic institutions there are many formal restrictions to dismissals, at least for scholars who have received tenure. Perhaps even more important is again the fact of asymmetric incentives.

Those who feel badly treated by the rankings and evaluations are greatly motivated and have the necessary time to oppose any effort to dismiss them. In many cases, the decision-makers foresee this resistance and make no effort to get rid of the unproductive members of a department. Alternatively, they are offered much money to make them leave voluntarily. While the latter seems to be an elegant solution it, of course, reduces the funds available for good research and teaching.

Those put at the top of the list in rankings and evaluations may also react in a way that is unfavourable for their own academic institution. Referring to their now "officially" sanctioned great performance they are motivated to ask for higher compensation. This makes the income distribution within

the department and university more unequal. There is (preliminary) empirical evidence that at least under some conditions a more unequal distribution reduces performance (Torgler et al. 2006). A move of the top people to other institutions inside or outside academia is beneficial for scientific research if these institutions act under competitive conditions. However, these conditions are far from being met in the German-speaking university system.

Evaluations have yet another disadvantage equally affecting everybody subjected to them. As far as they are perceived to be "controlling" by the evaluatees they tend to crowd out the internal motivation. But it is exactly this type of motivation, rather than the extrinsic one, which is fundamental to creative research (Amabile 1996; 1998). Indeed, it is well known that the great scholars were invariably motivated by an interest in science itself, and that the monetary gains going with it are secondary. As a result, the bureaucratic nature of evaluations tends to crowd out the work effort of the best scholars. Even *if* ranking and evaluation exercises were able to raise the average performance of economic institutions (for which there is no evidence), they hamper top performers. It may well be that this result is desired but it has little to do with the university as a place where the very best scholarly research is undertaken, and it brings up the question where this activity will be undertaken in the future.

10.3 Academic Institutions Evaluated

Universities and other scholarly institutions build of course on the performance of their members and, therefore, are directly affected by the damaging effects of evaluations on academics. However, there are some additional effects to be noted. Most importantly, rankings and evaluations are increasingly applied to academic institutions as a whole[99]. This disregards the fact that within these units there typically are huge differences in quality. Such an approach sends the wrong outside signals because it disregards these quality differences. For example, if a university as a whole is evaluated to be at the top, every faculty can claim to be part of this top ranking even if its actual performance is lacking. In contrast, if a university is evaluated to be average or less, an individual academic unit finds it very

[99] An important case is the designation of whole universities as "elite" institutions in Germany. This is an extreme case of a top-down approach to science undertaken by governments who seem to believe that good academic performance can be ordained, implicitly stating that money is the most important ingredient.

difficult to convince outsiders (e.g. in order to attract funds for research) that it does not actually share in this negative evaluation.

Some of the unproductive reactions to evaluations of individual scholars discussed above are strengthened at the institutional level. This applies in particular to the efforts to nullify or turn around an unfavourable evaluation. To the extent that a university's future depends on such an evaluation, there are very strong incentives to resort to unproductive political rent seeking. Universities know that politicians depend on their local constituency and will make great efforts to support them. There are many different arguments available to buttress their case. A convincing argument is always the general desire for a "just" distribution of government funds over space. Another one is the cartel formed by the universities and the local business communities that carries considerable weight especially if an "impact study" puts the prospective loss to the region in monetary terms.

10.4 What to Do?

The argument so far has been that the substantial and sizeable costs of rankings and evaluations have systematically been ignored. These are not, as often thought, the direct costs on the part of evaluators and evaluatees. While they are sizeable, they are partly reflected in direct monetary costs (notably on the part of the evaluators) as well as in the time and effort expended (which are often discussed among academics). The costs induced by the reactions of the evaluatees, however, are presumably much larger but nevertheless tend to be overlooked. The result is an overuse of rankings and evaluations that gravely damages the academic system. Many observers may well agree but argue that there is no alternative: how should government funds be allocated "rationally" if it is not known who is academically productive and who is not?

Unfortunately, the widespread and increasing use of academic evaluations is rarely seen in a broader perspective. Valid alternatives are therefore overlooked. But there are two institutional solutions that do not require ex post evaluations by external experts for the government.

The first solution establishes *competitive use of rankings of different academic units*. In such a setting the various departments have an incentive to attract those scholars who will make the greatest addition in the future performance of a university. Rankings still exist but are produced for the benefit of the various decision-makers in competition with each other

rather than for the information of the government. Care will be taken to produce rankings for the various areas of universities. For instance, there will be rankings for individuals deciding to take up their first year study, other rankings for graduate and post-graduate students, still other rankings for research in the different disciplines, fields and sub-fields, and in line with the globalization of science there will be international rankings. No effort will be made to establish *one "overall and official" ranking* of a discipline (such as endeavoured by the Deutsche Wissenschaftsrat). Moreover, the assessment of individual scholars will be directed to his or her expected future contribution rather than backwards as rankings and evaluations are in a government run university setting. As the present university systems in German-speaking countries are far from this desired setting, it is not further discussed here.

The second solution is possible within today's German-speaking university system. It relies on the idea of an appropriate *input control*. It is difficult or even impossible to effectively use process and output controls (for these terms see Frey and Osterloh 2006). The main emphasis is on a *good selection of scholars who are then essentially left to act at their own discretion*. The result to be expected is a wide variation in performance. Some scholars will excel under these conditions because they are left unbothered by bureaucracy. They can devote their effort and time to research instead of having to continually prepare for evaluations and react to them. At the same time some of the scholars will not perform well. They will exploit the discretion given to them, become lazy or engage in endeavours unrelated to their university position. The proportion of the well-performing type of scholars can be raised by a careful selection procedure including an intensive period of social integration into academia. This procedure allows universities to choose capable scholars with high intrinsic motivation for research and teaching.

This second solution is often considered to be naïve and outlandish. In any case it is contrary to the current notion of what makes people work efficiently. However, the continuous control of the performance exerted today in many corporations is not necessarily the best approach to reach excellence in the more creative areas such as science where people with particularly high intrinsic work motivation are needed. For that reason, the imminent introduction of performance pay in the academic system is doomed to failure – at least if original work is to be produced.

Today's general rejection in German-speaking countries of the second solution which is based on careful selection and social integration is surprising for two reasons. Firstly, the general tendency is to imitate the Ameri-

cans always and in all respects. But in this instance, one tries to raise the performance of academic institutions by extensively using rankings and evaluations from above. One fails to see that due to the competitive situation in which American universities find themselves, and the close association of the quality of scholars and of universities (Franck, Opitz 2006), they accord great importance to an extended selection process. The main goal is to find the persons best suited for a university position and to consider how he or she is likely to perform *in the future* - and then to trust that he or she will indeed perform well. It is understood that after careful selection and training one has to abstain from external evaluations regarding output and to some extent also regarding process control. Such a control approach to scientific research was emphasized by the famous President of Harvard University James Bryan Conant (Renn 2002):

> „There is only one proved method of assisting the advancement of pure science – that is picking men of genius, backing them heavily, and leaving them to direct themselves."
>
> (Letter to the New York Times, 13. August 1945).

This view is still part of the *Principles Governing Research at Harvard*, stating:[100]

„The primary means for controlling the quality of scholarly activities of this Faculty is through the rigorous academic standards applied in selection of its members".

The rejection of the approach based on a careful selection first and then allowing for the greatest possible freedom afterwards is also surprising in as much as it was prevalent in the German-speaking university system exactly while it was the dominant approach in the world. It can, of course, be argued that conditions have much changed since then and what was successful then need not be now. This is certainly true but I have tried to argue that the basic requirement for creative scholarship has remained the same, namely a good measure of discretion to exert one's intrinsic motivation for academic work.

10.5 Is a Change in Policy to Be Expected?

The general view that ideas which are working well in the market and private business, should also be adopted by the public sphere is still dominant; I therefore do not expect that the arguments proposed here have any

[100] See http://www.fas.harvard.edu/research/greybook/principles.html.

effect in the immediate future. Only slowly can the idea be entertained that the reverse transfer could also be of interest: private business can, in some respects, learn from government (Frey, Benz 2005). The one thing that can be done is to point out the many obvious shortcomings of an academic system relying on rankings and evaluations and related mechanisms such as performance pay in universities. The most grotesque cases of rankings and evaluations and their consequences can be publicized. This may slowly undermine the erroneous notion that what is (perhaps) good for business must be good for the public sphere, in particular universities.

References

Amabile T (1996) Creativity in Context: Update to the Social Psychology of Creativity. Boulder, Westview Press

Amabile T (1998) How to Kill Creativity. Harvard Business Review 76(5), pp 76-87

Berghoff S, Federkeil G, Giebisch P, Hachmeister CD, Müller-Böling D (2002) Das Forschungsranking deutscher Universitäten. Working Paper 40, Centrum für Hochschulentwicklung

Bhagwati JN (1982) Directly Unproductive, Profit-Seeking (DUP) Activities. Journal of Political Economy 90(5), pp 988-1002

Bommer R, Ursprung HW (1998) Spieglein, Spieglein an der Wand: Eine publikationsanalytische Erfassung der Forschungsleistungen volkswirtschaftlicher Fachbereiche in Deutschland, Osterreich und der Schweiz. Zeitschrift fuer Wirtschafts- und Sozialwissenschaften 118(1), pp 1-28

Eichenberger R, Frey BS (2000) Who's Who in Economics? Kyklos 53(4), pp 581-586

Franck E, Opitz C (2006) Incentive Structures for Professors in Germany and the United States: Implications for Cross-National Borrowing in Higher Education Reform. Comparative Education Review 50(4), pp 651-671

Frey BS (2007a) Evaluitis – eine neue Krankheit. Leviathan, forthcoming

Frey BS (2007b) Evaluierungen, Evaluierungen... Evaluitis. Perspektiven der Wirtschaftspolitik, forthcoming

Frey BS, Osterloh M (2006) Evaluations: Hidden Costs, Questionable Benefits, and Superior Alternatives. Working Paper 302, Institute for Empirical Research in Economics

Frey BS, Benz M (2006) Corporate Governance: What Can We Learn from Public Governance? Academy of Management Review, forthcoming

Renn J (2002) Challenges from the Past. Innovative Structures for Science and Contribution of the History of Science. In: Max Planck Forum 5, Innovative Structures in Basic Decision Research. Ringberg Symposium, 4.-7. Oktober 2000 in München, pp 25-36

Torgler B, Schmidt SL, Frey BS (2006) The Power of Positional Concerns: A Panel Analysis. Working Paper 2006-19, CREMA