COMPARING COMPARISONS: ON RANKINGS AND ACCOUNTING IN HOSPITALS AND UNIVERSITIES

Sarah de Rijcke, Iris Wallenburg, Paul Wouters, Roland Bal

INTRODUCTION

Comparisons have become ubiquitous in the management of quality in social domains that were previously governed by professional elites. Often these comparisons are framed as ‘transparency instruments’ (Hazelkorn 2011b: 41) which come in a variety of forms:

- informative guides for prospective ‘clients’ (e.g. students or patients);
- accreditation procedures to certify the legitimacy of a particular organisation to act as a university or hospital;
- benchmarking as a way to ‘compare like with like’ or to check the compliance with formalised norms and standards;
- regular formalised evaluations or assessments of the activity, quality, and impact of the products’ analysing and processes;
- classification systems to develop sharper profiles of the institute and its components; and
- rankings of universities and hospitals to see who is ‘best’ in the comparison according to a particular set of measures.
Driven by the aim of transparency and user empowerment in public and private services, comparisons have become a crucial instrument in contemporary capitalism. Typically, they are supposed to ‘enable consumers to choose’ between telephone companies, universities, power companies, schools, or hospitals. These types of comparison are characteristically framed in terms of ‘improving quality’ (both in terms of service organisations and individual service workers) and ‘empowering consumers’ (in regards to how the visibility of performance enables consumers to make a deliberate choice about what service to choose for and by whom) (Shore and Wright 1999). Otherwise, comparisons are framed for organisations to ‘publicly account’ for their performance, and regulators take measures based on comparative analyses. Comparing is therefore not only an everyday practice; it is also a highly specialised activity that has become ubiquitous in many contemporary forms of governance.

In this chapter, we focus on one comparative technique: ranking. A key difference between ranking and other comparative techniques is that it presents an ordered list of an entity according to how it scores on a particular (set of) indicator(s), often starting with ‘the best’. A ranking is easier to grasp than a more complex comparative technique such as benchmarking or a multi-dimensional classification or assessment system. It is also easier to misunderstand. The indicators used for ranking may not measure the qualities the ranking is supposed to capture. Moreover, the entities they measure may not have similar profiles. In this sense, to what extent are the current rankings of universities and hospitals (which are the focus of this chapter) comparing ‘like with like’? These criticisms have not diminished the power of rankings – in fact, it is quite the contrary. Though rankings often start with a heterogeneous set of organisations, they are able to make very different entities comparable. As an instance of a comparative technology, a ranking produces certain realities and identities; it creates that which is compared (see Mol 2002). In this sense, rankings are indeed a member of the family of comparative technologies: comparability is their outcome, as well as their foundation.

Below we examine how rankings came into being, how they are done, and what they do by comparing ranking systems for, and within, universities and
comparing comparisons

hospitals. On top of this, we seek to explore our own ways of comparing ranking practices. As the introduction to this book explains, we wish to question the epistemic position of comparative methodologies in the social sciences. We neither believe that they are inherently superior to the in-depth study of one particular case, nor do we agree with the outright rejection of comparative sociology as reductionist. Rather, we are interested in the kinds of effects that comparisons produce – both the comparative technologies of rankings systems, and our own comparative analysis of the ranking of (and within) universities and hospitals.

This chapter presents our comparative layers in the form of a triple jump. First, we describe the emergence and development of rankings in universities and hospitals. In this part of our chapter, we zoom in on some of the differences made by these rankings after their introduction in the past two decades. We describe rankings as tools for governance that revolve mostly around competition and ‘commensuration’ – social mechanisms through which highly diverse entities (countries, institutions, people) are rendered measurable and comparable through quantitative means (Espeland and Stevens 1998). Second, by drawing on a comparison between university and hospital ranking practices, we analyse how ranking contributes to making organisations auditable and comparable. We examine some of the differences and similarities of ranking practices in universities and hospitals by focusing on three themes that emerged from our comparative work: 1) the ambivalence of ranking; 2) the performativity of ranking; and 3) coordinating ranking practices. We link these themes to the literature on comparisons. Third, we reflect on how we as analysts have ‘practised comparison’ by expounding on how we enacted ranking in our comparative work (see Urry and Law 2006), and on some of the costs and benefits involved in our comparative endeavour.

The chapter is based on two distinct research projects in which we analyse contemporary practices of university and hospital ranking in the Netherlands. Our cooperation was triggered by similarities we observed in approach and empirical material. The projects share a theoretical focus in the sense that both projects zoom in on the enactment (or the daily work of ‘doing’) (see Mol
2002) rankings) in ‘real’ organisational practices. The modest empirical research done by others thus far mainly focuses on higher management levels, and/or on large institutional infrastructures. Instead, we analyse hospital and university ranking practices from a whole-organisation perspective. Both projects look at how rankings translate, purify, and simplify heterogeneity into ordered lists of comparable units, and the kinds of realities that come into being through these ranking practices. Among other things, we are interested in the constitutive effects of ranking (see Dahler-Larsen 2012) and the kinds of ordering mechanisms (Felt 2009) that ranking brings about on multiple organisational levels – ranging from the managers’ office and the offices of coding staff to the lab benches and hospital beds.

Both research designs were comparative, and both projects performed ethnographic work in three places. Sarah de Rijcke (SdR) and Paul Wouters (PW) conducted research with three biomedical research groups (a lab, a group of medical statisticians, and a clinical research group) in a Dutch university medical centre. Their project focused on the implications of research assessment and ranking on biomedical knowledge production. The rationale for having three places of investigation in this project emerged in part from institutionalised distinctions between basic, translational, and clinical research at the centre. These boundaries not only related to differences in epistemic cultures, but they were also quite literally felt in terms of institutional architecture (the laboratories were, for instance, located in a separate building). We took it as part of the ethnographic work to analyse the enactment of these epistemic and material differences in the research practices under study.

For each group, we held interviews with researchers at different career stages – with technicians, group leaders, heads of departments, and quality managers. In addition, we performed observations during work-in-progress meetings, seminars, appraisal meetings, and interactions with companies and other stakeholders. SdR had full-time access for a month at each group. The access was granted by the research groups and the dean of the institute, who is also one of the drivers behind a project on ‘systemic failure in medical research’ at the Netherlands Organisation for Health Research and Development
Comparing Comparisons

(ZonMW) (Groen 2013). This ZonMW project was triggered (among other things) by a perceived increase in publication pressure in biomedicine, a lack of interdisciplinary cooperation, and a responsiveness to the ‘societal relevance’ of research (Ibid. 1).

In the study on hospital rankings, Iris Wallenburg (IW) and Roland Bal (RB) compared three Dutch hospitals.¹ Hospital selection was done by looking at similar sized hospitals but in different competitive environments, due to an expectation that the level of competition hospitals find themselves in would influence the way rankings affect hospitals. More competitive regions show higher levels of tight coupling, emphasising strong hierarchical coordination between managerial and professional departments (whereas loosely coupled organisations allow for more professional autonomy).

For this project, interviews with a multitude of actors implicated in hospital ranking were held, and observations were performed in meetings of quality of care committees, meetings of hospital managers with outside actors (like insurers and regulators), and during administrative work in the hospital (both in clinical settings and information departments). As rankings are increasingly being used in the governance of Dutch hospitals (and policy actors expect much from them), the aim of this project was to get a better understanding of how ranking affects hospital organisations and care practices.²

In most rankings, the work of commensuration and classification is black boxed. The same holds for a lot of comparative research in the social sciences; for example, when it is assumed that comparative research designs by definition provide more robust forms of knowledge. As well as analysing the comparative effects of rankings, our aim for this chapter is to be more open about how we enacted comparison, leaving room for reflections on our own classification work. That is, we mobilise observations in one project to discuss findings in the other, and vice versa. This way of creating ‘rapport’ (Stengers 2011, in Akrich and Rabeharisoa, this volume) played an important role in drawing conclusions from our respective field notes. As a result, comparison turned out to be far messier – both in ranking practices and in our own work – than is often assumed. We think that this is not a deficit but a consequence of the grounded nature of all comparative practices.
HOP: THE GROWTH OF RANKING IN UNIVERSITIES AND HOSPITALS

How did it come to be that we now inhabit a world in which rankings seem to be inevitable? Sociologists and anthropologists contextualise the popularity of ranking as a broader manifestation of audit processes in an increasingly wide variety of societal sectors and professional fields. Audit processes now range from the online rating of movies, books, and restaurants, to assessing professional performance in sectors such as healthcare and higher education. Today, there are virtually no areas in which professionals are not – in one form or another – invited to respond to regular assessment exercises. The rise of performance-based funding schemes is one of the driving forces behind the increased interest in university and hospital rankings. Some studies suggest that shrinking governmental research funding from the 1980s onwards has resulted in ‘academic capitalism’ (see Slaughter and Lesly 1997). By now, universities have set up special organisational units and devised specific policy measures in response to ranking systems. Recent studies point to the normalising and disciplining powers associated with ranking and to the response to ‘reputational risk’ as explanations for organisational change (Burrows 2012; Espeland and Sauder 2007; Power et al. 2009; Sauder and Espeland 2009).

The first university rankings were published at the beginning of this millennium (Hazelkorn 2011), roughly two decades after the first signs of an unprecedented growth of evaluation institutions and procedures emerged in the 1980s. This growth was due to (among other things) the increased economic and social role of science and technology; an increase in the scale of research institutes; a general move towards formal evaluation of professional work; and limitations and costs of peer review procedures. Today, universities routinely monitor the publication of national and international league tables, and promote their position on websites, in newsletters, and in advertisements that target new students and staff. There is a growing emphasis on ‘reputation management’, and the use of quantitative performance indicators in quality assessment policies is steadily increasing. In short, quantitative indicators of science and technology
are applied from the most ‘macro’ to the most ‘micro’ by governing bodies and agents across those levels.

In Dutch healthcare, performance measurement was introduced at roughly the same time. Healthcare had routinely been characterised by a ‘closed shop model’ in which physicians decided upon the processes of healthcare delivery – only being accountable to knowledgeable colleague physicians (Harrison and McDonald 2008). However, since the 1990s, medical professional autonomy has gradually eroded, and professional self-regulating principles have intermingled with principles of performance management (Wallenburg et al. 2012; Waring 2007). Generally, the rise of performance management is a consequence of two intertwining developments. First is a shift in healthcare policy. Due to the introduction of New Public Management policies in the early 1990s, healthcare workers are increasingly being held accountable for the care delivered, and as part of this are obliged to provide insight into their work and performance. In the Netherlands, the shift to this ideal of transparency became even more prominent with the introduction of the system of regulated competition in the mid-2000s (Bal and Zuiderent-Jerak 2011).

The second main development is a shift in the regulation of professional work. The medical profession has been confronted with a sharp decline of public trust in medical expertise and ‘medicine’s good work’ (Freidson 2001; Dixon-Woods et al. 2011). Together with the growing specialisation within medicine, the introduction of information technologies, and the regulation of working hours, medical work has become increasingly ‘normalised’ and regulated (Nettleton et al. 2008; Wallenburg et al. 2013). The comparison of hospital performance by ranking has been made possible as part of (and due to) these changes in healthcare governance. In the Netherlands, the most well-known are the rankings of the popular newspaper Algemeen Dagblad, and the weekly magazine, Elsevier. Yet in the past few years, healthcare insurers, patient organisations, and social entrepreneurs have created many more rankings. These rankings are, amongst others, based on patient experiences, and on the organisation and the outcomes of care – for instance, as seen in mortality rates for specific diseases or the percentage of pressure ulcers (Jerak-Zuiderent and Bal 2011).
Though ranking has become increasingly important in public service sectors, not much research has been done as of yet on the ways in which these comparative processes affect these sectors. The little evidence that exists tends to focus on universities. Here, an interest in ranking indeed seems mainly driven by a competition in which universities are being made comparable on the basis of ‘quality’ and ‘impact’. One of the most obvious manifestations of the increased popularity of ranking practices is apparent in the way universities have started to routinely monitor the publication of global league tables, and in how they advertise their position in these tables on websites, in newsletters, and in advertisements that target new students and staff. This responsiveness is telling of the importance ascribed to rankings, though the rationales behind their construction tend to be disregarded. All measurements are of course preceded by decisions pertaining to the object(s) and focus of measurement. Certain factors are labelled as relevant in this categorisation process, and others as less relevant (or even irrelevant). Decisions will be made pertaining to the parameters of the categories that will be taken into account. These decisions fundamentally shape the subsequent measurements.

First of all, every form of ranking is based on data about a limited number of features which are subsequently made measurable. Global university rankings, for example, will focus on the 1,000 or so universities that are visible at the international level while ignoring the other 16,000, because these only play a role at the local level. They also tend to focus on research performance since this can more easily be made comparable at the international level (by way of citation analysis). Hospital rankings, in their turn, are criticised for using quickly changing performance indicators that underlie the rankings, rendering it difficult for hospitals to meet the criteria that are being set. Moreover, they are targeted at easily measurable aspects of care such as mortality, while ignoring other aspects – which are sometimes deemed more important – like diagnostic accuracy or empathy for patients. At the same time, the kinds of parameters that are used fundamentally shape the outcomes. Some rankings strongly favour large universities or ones with long-established reputations, or they give more weight to publications in certain types of journals (e.g. *Nature, Science*), thereby implicitly leaning towards the devaluation of certain types of research (e.g.
humanities, social sciences). In addition, composite rankings like the Shanghai University ranking or the Dutch hospital rankings will merge different aspects of university performance (e.g. research, teaching, valorisation, social impact) or hospital performance (e.g. mortality, infection rates) into one number. How this composite number is calculated is rather arbitrary, not always transparent, and changes over time. It is therefore unclear as to what extent a change in position has on an actual change in performance, or if it should be ascribed to an insignificant fluctuation. In addition, even individual outliers can cause seemingly robust improvements of the performance of universities or hospitals.

**As discussed above, rankings in hospitals and universities show some similar characteristics in that they are embedded in, and form an infrastructure for, competition on a ‘market’ for public services (in this case, healthcare and research/education). They are also developed against a background of increasing demands for accountability of elite professions, and tap into a neoliberal agenda where auditing is seen as a practice of soft regulation.**

One of the goals of our own comparison was to gain more insight into how rankings are enacted in day-to-day university and hospital practices (more than would have been possible without the comparison). We therefore started to compare some of the outcomes of our research projects. In comparing our findings, we developed three themes that we further discuss below: 1) the ambivalence of rankings; 2) the performativity of rankings; and 3) coordinating ranking practices.

**The Ambivalence of Ranking**

In the ranking literature, the adequacy of indicators or their composites is generally critiqued (see Jacobs et al. 2005; van Dishoeck et al. 2011; Marginson...
This literature questions the validity of rankings, and argues that using rankings in regulation and funding decisions should generally be avoided. Similarly, in our case studies, rankings induced ambivalent responses. Regarding hospital rankings, our respondents commonly felt that rankings are not important for the ways in which organisations operate. They stressed the lack of validity and the volatility of rankings, arguing that they are ‘lotteries’ and are unpredictable. It was argued that rankings do not appear to have any consequences in terms of patient choice or insurer commissioning of care. Other respondents, particularly physicians, pointed out that indicator-based performance measurement directs attention to measurable agendas, while the most difficult agendas with regard to performance in the domain of quality and safety are related to the tacit expertise of professionals (such as diagnostic interpretations).

In our fieldwork in universities, we noticed that an interest in global university rankings and the development of adaptive strategies seemed mainly relevant at the level of deans and other research managers. The researchers who acted as informants seemed mainly involved with their own performance and that of their colleagues. As noted in the dialogue below, their commitment with their institution’s ranking scores was rather low:

**SdR**: Do you think that university rankings affect trust relations between different academic medical hospitals?

**Professor** [switches to 3rd person plural]: Well, when they go up in the ranking, this leads to a celebratory announcement on our intranet. It is much like what happens when the Dutch national football team wins an important match’ (4 November 2012).

Perhaps this lack of commitment to institutional ranking scores makes sense if we take into account how academic careers currently take shape. Job insecurity may trigger researchers (as members of an increasingly flexible ‘workforce’) to be more committed to pursuing the next step in their career than to the performance of the organisation they are affiliated with. Similarly, in hospitals, medical specialists were usually more interested in their relations to the same speciality
in other organisations than to the performance of the hospital in which they work. Yet, at the same time, it was felt that rankings had to be taken seriously, as they were one of the drivers of the increasingly important reputation of hospitals and universities. For example, during an interview, the quality manager from hospital A pointed towards the limited usefulness of rankings, and continued her argument by saying

[w]e still want to end high. When we dropped from the top 25 to place 60, that wasn’t liked much (12 November 2012).

During interviews, informal conversations, and meetings, rankings were often criticised. Yet we noted this did not impede managers and practitioners in actively engaging with ranking practices:

During one of our [IW and RB] interviews, a hospital administrator criticised current measurement policies that, according to him, did not reflect reality: ‘According to the numbers we were a kind of “death hospital”, but it all depends on how you measure mortality rates’. According to this administrator, hospitals are heavily disciplined by ranking policies, clearly objecting to the practice and even lecturing us on Foucault’s notions of discipline. After having said this, he turns to a pile of papers on his desk, showing us the figures of the performances of the different hospital wards: ‘You see, ward Z did an excellent job, they will have cake and a picture with me on the intranet next Monday! [smiling] They love it if we celebrate good performance’ (Observation notes, 22 June 2012).

SdR and PW made a somewhat similar observation when they discussed the possibilities for ethnographic research in a large medical research centre:

The present and future dean, the director of research, and a quality manager hosted the meeting. On the basis of our research proposal, all of us reflected on adverse effects of the increasing use of quantitative performance indicators. Until the quality manager received an e-mail, which he read on his
iPad, containing bibliometric data on the centre’s performance (as measured through citation analysis). Compared to the year before, the institute had ‘gone up’ on all bibliometric indicators, a fact he immediately shared with the present dean by handing him the iPad. The dean was quick to ask whether ‘his’ institute was now ‘first’, and began to fantasise about presenting a list of the scores of all competing medical research centres at his farewell party a couple of weeks later (Observation notes, 1 March 2012).

This meeting to establish researchers’ access (along with the meeting with the CEO in one of the hospitals) revealed how indicators and rankings were both criticised and embraced, depending on the specific ‘partial connections’ that were made (Strathern 2000). Although our respondents express their experiences of how rankings may be critiqued in a variety of ways, they cannot escape from rankings. Essentially, rankings are actively used to enhance organisational performance. In the above extract from our fieldwork at the university medical centre, a complex set of bibliometric measures was translated into a ‘simple’ ranking by one of the deans. This ‘responsive’ or ‘implied’ ranking practice affords this dean strategic use of a more intuitive comparison between organisations than would be possible if he were to draw on the entire bibliometric assemblage that was presented to him by the quality manager. The hospital manager, in turn, used the ranking to encourage nursing wards to enhance their performance. Nursing wards that performed well were displayed on the intranet, with the CEO being part of them. Yet in the same ‘moment’, ranking also induced a deep criticism and a feeling of discomfort, as these managers (both also professionals) were highly sceptical about the practices underlying the rankings.

As researchers, this two-sided picture of resistance and engagement surprised us. Although our ethnographic methodology does not allow for strict causal analysis, we did wonder why this experience was so strong. Rankings seem to perform comparison in ways that tie in with deeply embedded cultural notions of performance and competition (i.e. ‘who is the best’ is a notion which already starts at pre-school). The tighter coupling between perceived performance and distribution of resources since the 1980s has further strengthened the effects of rankings. In addition, ‘to rank’ comes naturally for professions that have become
highly competitive amongst themselves. Many researchers and medical specialists seem driven by the urge to outperform their colleagues, and the ranking mechanisms we encountered make this quite visible. We suggest that it is this simple visibility of an ordered list that makes rankings stand out compared to more complex forms of comparison.

To us researchers, the excerpts above were exemplars of what Marilyn Strathern has pointed out as ‘ethnographic moments’ – a relation that joins the understood (i.e. what is analysed at the moment of observation) to the need to understand (i.e. what is observed at the moment of analysis) (Strathern 1980; Mol 2011). In our situations, with more or less overlapping observations in both research projects, these ‘moments’ acted as points of recognition and of shared surprise and feelings of discomfort. It was stimulating to have long repeated debates about the ambivalence of ranking and about its contesting values and ethics (e.g. ‘good practice is much more complex and thus not easily measurable’ versus ‘we want to be outstanding, and measurement and ranking may help to achieve this’) and all other ambivalences involved. It made us realise that criticising ranking is (too) easy, just like understanding ranking as a practice of ‘gaming’ or creating a well-cut and organised world next to a much more fuzzy world of professional work. Instead, ranking involves both conflicting ideas and ways of acting; it is both order and mess, and these go hand in hand.

**Performativity of Ranking**

In the sociological literature, university rankings are associated with competition at the level of the entire organisation. We also noticed how rankings were used to police work processes and enhance organisational performance. In one of the hospitals, for instance, the clinical pathway for breast cancer treatment was revised to reduce waiting times for surgery in order to obtain higher scores on the national ranking of best hospitals. However, our comparative approach revealed that rankings enacted organisational practices that went beyond this competitive aspect. Ranking was also used to create or enhance group identity (e.g. by displaying the medical research centre as the best), to reform routine
practices, and to encourage individual workers to excel. Within the translation of rankings to everyday work processes, the complex metrics that underlies a ranking were reconfigured to simplified lists of factors that made performance measurable and comparable. The presence of such ‘implied’ ranking practices points to the performativity of ranking – that is, a reinforcing loop that redefines organisations, individuals, and also projects, in terms of a ranking:

Professor D is preparing a funding application for Cardiovascular Onderzoek Nederland (CVON) as one of the Principal Investigators (PI) in a larger consortium; his institute is the prospective ‘coordinating group’. Funding schemes such as this one explicitly use the ‘H-index’ as an indicator for the aptness and ‘proven track record’ of PIs (CVON call 2012, p. 5). The calculation of this index is relatively simple: an H-index of X means that the researcher has published X articles that have each been cited at least X times. Consortia need at least one PI with a high H-index for a proposal to be eligible for funding at all. These PI’s can take part in maximum one proposal a year. Combined, these two criteria create a lot of lobbying and a ‘run’ on PI’s with a high H-index (Observation notes, pre-clinical research group, September 2012).

In this frame of comparison, an increasingly competitive funding landscape forms the background for situations in which project proposals are being made comparable on the basis of how the PIs in the consortia score on the H-index. This ratio of published articles/number of citations homogenises, simplifies, and enables funding agencies to rank consortia and prioritise proposals by ‘enlisting’ PIs via their proven track record (as expressed in their H-index). Recognising the importance of the H-index encourages scientists to use it even if this is not officially required. One of our informants mentioned taking part in a ‘grant proposal preparation class’ at his institute. The institutional research policy advisor (and host of the workshop) advised participants to mention their H-index if they felt it could help them stand out in the ranking of research proposals and researchers during the prioritisation work done by review committees. These ways in which researchers and policy officers act strategically with performance
indicators are not (yet) part of the official evaluation criteria. However, these actors play an active role in these evaluation systems. The scientific system is highly competitive. Therefore, it is in the interest of institutions and researchers if high-scoring individuals actively display their scores, thereby fuelling the further development of indicator-based research assessment (Wouters 2014).

Similarly, hospitals tend to focus their quality policies on those areas which are important for their score on rankings. For this purpose, hospitals benchmark themselves on underlying performance indicators, taking particular notice of hospitals in their direct environment. Programmes that focus both on registration work and quality improvement are set up especially in those areas where hospitals score relatively low. For instance, one of the hospitals we studied did rather poorly on the performance indicator for malnutrition. The quality manager was requested to investigate the causes of the low score. It appeared that the hospital failed in measuring the nutrition status of elderly patients at their fourth day of admission, as was required by the performance indicator. Subsequently, nutrition assistants were trained to conduct these measurements. From the interviews with healthcare professionals, it appeared that this reorganising of care goes against areas of care that are not represented in the rankings, or that are not made measurable. Similarly, as discussed above in the example of the pressure ulcer scores, ranking practices became embedded in the hospitals to stimulate professionals to do registration work.

Again, we see similar processes in universities and hospitals. In both institutions, strategic behaviour is induced through the reputation game of the ranking practices. Ranking practices direct focus and activities, even though – in the same ‘moment’ – they also elicit criticism. The organisations under study seemed increasingly embedded in responsive ranking practices. On a number of levels, the organisations and individuals within them defined themselves in terms of a ranking (for instance, when they tried to get a handle on more complex comparative mechanisms at play).

When comparing our findings, we had room (both intellectually and methodologically) to move back and forth between our joint discussions and our research sites. In developing this collective comparative space, the performativity of rankings became a much-debated topic. We ended up constructing
similarities that related, for example, to an equal emphasis on ‘good scores’ and the policing of existing organisational and working routines to enhance these scores. Simultaneously, our focus on similarities also enacted and underscored differences. For instance, in the academic context we found that performance was increasingly related to individual performance (think about the H-index mentioned above). While IW and RB translated this outcome to the hospital setting – attempting (perhaps expecting) to discern a similar shift to individual performance in medical work – they did not find this result. Discussing our ethnographic moments, surprises, and expectations helped create new lenses to reconsider our data and research fields. In the end, transparency about individual performance seemed in conflict with medicine’s emphasis on socialisation and moral protection of physician-colleagues (Bosk 2003 (1979); Wallenburg et al. 2013).

**Coordinating Ranking Practices**

On a strategic level, rankings seem to affect university and hospital policies in a number of ways (Marginson 2012; Rauhvargers 2011; 2013). These include the bringing in of new types of knowledge like reputation management and marketing; the realignment of administrative processes and the development of new types of research/care and accountability processes; the ‘buying of CVs’ to enhance measured performance; and the responsive ranking practices that our informants resorted to in order to make sense of more complex comparative strategies for the purpose of performance measurements. These and other feedback mechanisms co-define how researchers, healthcare professionals, and policymakers operationalise the notion of ‘high quality’. Interestingly, these feedback mechanisms may result in strategic behaviour which potentially undermines the validity of the performance indicators that ranking practices are based on at large. This is not merely the result of top-down criteria that ‘trickle down’ to local research practices. Interactions from the bottom up between people involved at different levels within the organisation are equally relevant. Let us take a closer look at the fieldwork in the hospital sector.
The hospitals studied went through great changes in terms of the organisation of administrative processes. What is euphemistically called an ‘uitvraag’ (an information demand by an external party) sometimes involved many months of work for the quality and information departments in collecting information from different sources in the hospital. This administrative work entailed many ‘investments in form’ to bring the information together (Thévenot 1984), including the involvement of health professionals to collect and register indicator information and the standardisation of care processes to enable data collection. Apart from information guiding the treatment process, health professionals had to collect data on all kinds of scores necessary for performance indicators. Nurses, for example, had to do risk assessment for pressure ulcers, delirium, and malnourishment, and had to regularly check whether a patient was in pain. The hospitals we studied had all installed different methods to make sure registration of care was actually done. These included building indicators in the electronic patient record, disciplining professionals by publishing information on registration, and ‘policing’ professionals to make sure registration was actually done. However, as one of the doctors we interviewed indicated, such policing is sometimes hardly possible:

The urologist argues that every time new measures come up ‘the hospital board wants us to participate in that’. He goes on to say ‘I just give the desired scores. Taking a biopt in one day is impossible, but I just indicate that we do it nonetheless. I don’t spend more than five minutes on this. It is uncontrollable’ (12 December 2012).

While hospital administrators and quality managers aim to standardise healthcare processes (and with that, the collection of data as much as possible), health professionals work around the system by enacting a form of ‘pragmatic compliance’. They just ‘tick a box’, as the urologist above points out. In this way, he complies with the demands put on him through the rankings while not changing his work practices, as he is aware that not ticking the box might have more serious consequences, if only to be publicly displayed within the hospital to be a ‘good’ or ‘bad’ performing ward (as was the case with the pressure
ulcer scores of nurses). This does not mean, however, that indicators are not taken seriously. The same urologist participated in several working groups both within the hospital, and in his medical association, developing performance indicators and related policies. A surgeon stressed the increasing importance of performance indicators developed by the medical professional associations:

SURGEON: The Netherlands Association of Surgeons (Nederlandse Vereniging van Heelkunde [NVvH]) possesses a complication registry. For an honest registration it’s crucial that this information is not made public […] Two years ago, Hospital X [a neighbouring hospital] had many reoperations for colon surgery. A delegation of the NVvH visited the surgeons. This all went quite harmoniously, you know, they came to see what happened and how things could be improved.

IW: Yet, I can imagine that such a visit says something; they aren’t there for nothing.

SURGEON: Of course, of course, they [the surgeons of hospital X] were fed up. They knew something was wrong. They had to act (2 November 2013).

The excerpts of both the surgeon and the urologist reveal medicine’s ambivalence towards performance indicators and rankings. Physicians feel a certain resistance towards external monitoring, but at the same time are driven by an interest in legitimising and developing their professional work. This results in what Levay and Waks (2009) have pointed out as ‘soft autonomy’, which combines professional internalisation of originally non-professional auditing practices with maintaining professional control over evaluation criteria. However, whereas Levay and Waks (along with other scholars studying changing professionalism) have emphasised the medical profession’s creative capabilities to capture external attempts to regulate their work (e.g. Waring 2007; Currie et al. 2012; Kuhlmann 2008), our research shows that despite the incorporation of performance indicators by the medical profession, they also act as a tin opener, elucidating the working routines of health professionals – making these visible, comparable, and negotiable.
We observed a similar dynamic in the university context. As part of the fieldwork at a cell biology laboratory, we were granted access to the yearly appraisal of one of the four group leaders with the head of the department. The meeting was held a couple of months after the institutional research assessment (held every six years) had taken place, and when the institute was in the middle of processing the results. The international committee performing the evaluation had followed procedures laid out in the Dutch ‘Standard Evaluation Protocol’ (SEP), and had used ‘informed peer review’ (see Colwell et al. 2012). This is a system in which peer review provides the overall framework for evaluation, but statistical data and citation indicators play a specific, often obligatory, role. Heads of departments are being held accountable on the basis of these assessments of their groups. Some managers use the numerical information to help make decisions about departmental research priorities, the use of lab space, and the distribution of other material and financial resources. In the yearly appraisal, the group leaders seemed well aware of these numerically driven decision-making processes. However, as seen below, Professor P’s own presentation in the appraisal, for instance, was also saturated with other indicators (particularly the number of articles and the Journal Impact Factor):

Professor (P): We have published nearly fifty articles, that means nearly one a week, and this is for the entire section; it is really unbelievable, two of them are really breakthrough papers. When I go somewhere […] they have all read it; it attracts a lot of attention […] I am currently working with [two Chinese postdocs] on a couple of very good papers. We will be able to send them to top [i.e. high impact factor] journals […]

Head of the department (H): I know you’re charmed with the Chinese, they score high, but they do leave afterwards.
P: Yes, but they do not need much supervision; I see them briefly during the weekend.
H: But there will be polarisation in your group if not everyone can live up to that level.
P: Yes, but what do you want? We score ‘very good’, not ‘excellent’.
[Here he refers to scores on the institutional evaluation]

H: You would have scored excellent if the past two years would have been taken into account [in the bibliometric analysis].

P: This did happen with other departments!

H: No, we stuck to that rule. If other departments were sloppy they were reproached for that […] These numbers are slow; it takes a long time before you get above a ‘2’.

[Here, H points his finger at one of the indicators in the bibliometric report, the group’s ‘Mean Normalised Citation Score’ (MNCS). The bibliometric analysis uses a relatively long citation window of five years and did not include the last two years. Calculation of the Journal Impact Factor is ‘faster’ because it is done on the basis of a two-year citation window.]

P: I am interested in excellence. If the assessment procedures do not match the work done, things will become difficult (Observation notes, 26 September 2013).

In the yearly appraisal, two indicators are drawn on to arrive at: 1) an ‘implied’ ranking of the professor’s group compared to other groups in the department, and 2) a reputational ranking of the journals that the group targets as outlets for their articles. In the former case, it is striking how a complex assemblage of indicators that forms the basis of the institutional evaluation is simplified, and now only revolves around this one indicator (the MNCS). In the appraisal, two ranking practices come together: institutional (via the MNCS) and disciplinary (via the Journal Impact Factor). Again, the measures act as ‘tin openers’. For instance, they enable the group leader to make a point about the other ranking game he is involved in (a comparative practice within molecular cell biology in which excellence is measured mainly through the impact factor). The professor celebrates the performance of his Chinese postdocs who have succeeded in publishing a ‘Nature, Cell or Science’ (NCS) paper. The reputation of his group is defined in part through the reputational ranking of the journals its members publish in. A reputational ranking of journals makes perfect sense for the PI, because it also helps him make decisions about managing his group (e.g. through the amount of ‘work’ they have done in relation to their performance),
or about how to ‘rank’ his employees in relation to where to allocate specific resources. Such ranking practices seem to form a routine part of peer review in his discipline. These processes take place on a global scale and are shaped in interactions between thousands of labs.

Institutional-level rankings are not that relevant in this process. As such, the excerpt above nicely reveals differences in ‘accountability repertoires’ (see Moreira 2005). That is, the indicators also enable the head of the department to caution the professor about his leadership style (which he draws in as a corollary of the reputational ranking of journals’ dynamics and says is creating pressure in the PI’s group). Accountability repertoires appear in different forms and pursue different goals. They are thus all considered valid and important. The actors involved need to make sense of these various co-existing repertoires, and attempt to find ways to combine them.

These different repertoires were also visible in the hospital context. Here, too, rankings and indicators not only operated through processes of self-control, but they were also opened up to new types of interactions (including with other actors), making negotiations on professional work possible. As one care manager noted,

[a]nd the other thing is that we of course use [rankings and indicators] as a management tool to get through to medical specialists […] [to get towards] particular improvement practices in the care process that have to be done. Step one, the ranking enters [the hospital]. In the following […] [For example], cardiologists score badly. As a consequence I go and see the medical manager, or do sometimes even visit the whole group of specialists, and I tell them:

‘Guys, this is really going badly here.’

Then they would tell me: ‘The numbers are not correct.’

Then we first look at the numbers together which they delivered […]

And I tell them: ‘This number was delivered, and you signed it. How come they are not correct nevertheless? What is the reason?’

Then they say that the case mix is […] different.5

Then you check this out.
Then you tell them: ‘From the benchmark it seems that that is not the case’ [i.e. incorrect case mix].

Then you approach the core and say: ‘Guys, you still score low, we took away variability, and now we have to discuss what we can do in our organisation, in our work process, in our medical policy, in our care process in order to make sure that there are better outcomes next time.’

But then, nevertheless, the ranking is for me still an instrument in order to effect change. Rankings are not a goal in themselves (2 May 2013).

In their internal use, indicators, and the ranking practices they support, increase the power for executives because they open up the primary process of care or knowledge creation for strategic criteria. In other words, they serve as tactical means to enable managers to negotiate and shape performance improvement agendas with professionals and researchers. Managers thus act as the ones undertaking the comparative work. Although health care practitioners and scientists conduct comparative work as well, in the end the managers brought the collective comparative work together and were accountable to external regulators assessing their organisation (whether these were the health care inspectorate and health insurers in the hospital case, or the heads of department in the university case). This also encouraged practitioners to streamline work processes or reduce the number of medical complications (as in the excerpt above). Therefore, ranking may act as a ‘tin opener’, but it also induces a new coordinating role to managers.

Considering our own comparative work, we (unconsciously) enacted a third comparative strategy. Besides studying shared research moments and seeing our research projects through the lenses of the other, we also grappled with the heterogeneity of ranking practices and observed how these are coordinated in everyday organisational work. Much more than trying to learn from ‘the other research project’, here we brought our findings together and considered them as one pool of data. We searched for relevant lines in this data that taught us about how ranking practices are coordinated and how this ‘coordination work’ (Mol 2002) influences debates about accountability and evaluation of ‘good practice’.
JUMP: ENACTING COMPARISONS

In this final part of our ‘hop-skip-jump’ approach, we discuss some elements of the ‘production process’ of the comparison between the hospital and university rankings we described above. By analysing how we approached the comparison and what it brought us, we aim to contribute to recent theoretical work on comparative methods in qualitative social science (see Niewöhner and Scheffer 2010). Above, we have ‘practised comparison’ by conceiving hospital and academic ranking practices through the lens of the other, and by subsequently searching for connections. These lenses helped us to understand the dominance of ‘measurability’ within the organisations we studied. We also described the heterogeneity of ranking practices and how actors have to work to align the different ways in which they are enlisted.

Comparison often entails ‘commensuration’ (see Espeland and Stevens 1998), and that was also the case in our own comparative practice. For example, although hospitals and universities are more like nodes in networks than ‘organisations’, in a confined sense (Clegg, Kornberger, and Rhodes 2005), we created a correspondence between them by approaching hospitals and universities as bounded entities. In doing so, we revealed some of the politics of tabulation and differentiation intrinsic to rankings, and zoomed in on their particular enactments in university and hospital contexts. One of the intellectual driving forces behind our comparison was that we were slightly dissatisfied with the crudeness of some recent analyses that point to the normalising and disciplining effects of rankings (see Power et al. 2009; Espeland and Sauder 2007). It was our ambition to come up with a more differentiated understanding of the workings of rankings. It is not enough to explain the popularity of ranking by pointing to an increasing drive for ‘competition’ in neo-liberal ‘audit societies’. Rather, we found the importance of competition to be an emergent property of highly situated ranking practices. The main purpose of both of our ethnographically driven research designs was to render visible the enactment or daily work of ‘doing’ rankings in real organisational practices, from an in-depth, whole-organisation perspective. Among other things, our comparison showed that rankings tend to evoke ambivalences as a result of their ‘decentredness’. That is, they are held
together through their fluidity; ranking is an effective comparative technology precisely because it is responsive, flexible, and capable of engaging multiple worlds (see De Laet and Mol 2000).

The performativity of our ethnographic mode of comparison was also visible in how we enacted a classical comparison between professional and managerial work, often employed in the analysis of quantitative comparative techniques (Triantafillou 2007; Sauder and Espeland 2009). As analysts, we differentiated between the types of work connected to research and care on the one hand, and organising work on the other, by studying the interaction between the types of activities that we saw, for example, in the practice of ‘pragmatic compliance’. However, we also noted that ranking became part of professional practices themselves, thus opening up new ways of interactions between the different types of work, which in a way transcended distinctions which were often made (also by us) between organisations, epistemic cultures, and work practices. For example, the observation of the meeting between the professor and his boss showed the intricate intertwining between managerial and research work, where the question of who is comparing whom, or to what effect, is no longer obvious.

How did we involve ourselves in this comparative process? We sat together (a lot) to share fieldwork experiences in offices and conveniently located teashops; we engaged with the respective material from the two projects by exchanging draft texts, and sat down together again to discuss similarities and differences. Importantly, our comparison was shaped through a combined background in STS. We drew from a shared reservoir of sociological and anthropological literature on classification, governance, quantification, and accountability. Clearly, this shared background also shaped our own classifications and the categories we drew up in combining the empirical material. Our background in STS is, for instance, very visible in our description of the ‘performativity of rankings’ section (above). This particular classification was certainly influenced by a ‘turn to performativity’ in STS – a mode of analysis and description that has been used to counter representationalist world views (see Pickering 1995) by demonstrating how descriptions, theories, and models become involved in the constitution of research objects they set out to represent.
Comparing Comparisons

But our training in STS is not the only reason for wanting to problematise ranking practices. Our comparative analysis (and an input into our analysis) was certainly also driven by our own mixed reactions to being ranked. So, whereas in our academic environments our performance gets measured through our publications, we also share a commitment to engaged research (Wouters and Beaulieu 2006; Bal and Mastboom 2007) and try to contribute to discussions in the Netherlands on care and research systems by giving lectures, participating in public debates, and writing publications in Dutch that are less visible in ranking practices. In addition, we are implicated in ranking practices in a more direct sense. We relate to the fieldwork material as researchers do to their empirical ‘data’, but in the case of the university rankings there is also another relationship. SdR and PW work at the Centre for Science and Technology Studies (CWTS), a research institute that not only hosts researchers who critically examine the impact of evaluation on knowledge production, but that also produces bibliometric analyses, including the ‘Leiden Ranking’. As such, the ethnographic work could have onto-political purposes in questioning how our colleagues practice bibliometrics, and how certain norms about knowledge production and ‘excellence’ are inscribed into citation databases and enlisted in rankings. The statistical experience of our colleagues rests on a great – yet positivistically inclined – sensitivity to category construction and classification. Questions of how ‘users’ are interpellated in bibliometric analyses, for instance, are not part of their acknowledged spectrum of analytic challenges. But like Stockelova (this volume), we find it unproductive to simply rebel against these prevailing frames of reference. Instead, we look for opportunities to carefully reinforce certain frames and challenge others. One opening we have is that there is an increasing need in the field for ethically responsible metrics, and for handles on how to generate productive feedback with ‘users’ about the ‘misuse’ of bibliometrics. We recently contributed to these discussions in an opinion piece for one of the leading information science journals (De Rijcke and Rushforth, forthcoming) at dedicated workshops and plenary sessions at scientometric conferences that we co-organised. The great asset of being located at one of the leading scientometric centres is that both practices (the scientometric and the ethnographic) are forced to interrogate each other. We
expect that this will lead to a better form of scientometrics (in terms of political as well as intellectual goals) and to a more informed ethnographic sensitivity.

IW and RB also work in a place heavily infused with benchmarking and cost-effectiveness research, and they collaborate in those kinds of projects. The original performance indicators of the Healthcare Inspectorate on which some of the rankings of hospitals are based were, for example, designed at the institute (Berg et al. 2005). Moreover, RB regularly sits on governmental committees discussing performance management systems in health care, and is involved to some extent in the ranking business. By being part of ranking practices in these diverse ways, and being attuned to STS types of analyses of quantification and commensuration, the projects we engaged in are in a way attempts to reflect on our own work and experiences. It made us aware of the pragmatic use of such comparative techniques on the one hand, and critically aware of their problematic nature on the other. This refrained us from becoming critical in a classical sociological sense of rankings, which – we like to think – has allowed us to do a more symmetrical analysis of them. However, as the above arguments on the consequences of our own analysis shows (i.e. bounding organisations and differentiating between managerial and professional work), such symmetry also came with a cost: in order to perform our symmetrical analysis, we had to engage in commensuration ourselves.

Our ambitions for this chapter are of course relative to the overlapping space we created between the findings of our individual projects. By being explicit about our approaches, we wanted to ‘thicken our ethnographic explication’ (Niewöhner and Scheffer 2010: 10) by searching for concrete interactions in ‘noncoherent practices’ (Mol 2011) – both at the level of the ethnographic material, and at the level of our own cooperation. We worked out this overlapping space in an attempt to follow unfolding relations in the situations under study where rankings make differences – without arranging them into an ordered list.
ACKNOWLEDGEMENTS

The authors are greatly indebted to the insights and support of the participants in our fieldwork. We would also like to thank the other authors of this volume for their valuable interactions at the workshops in London. Finally, we express gratitude to the editors for their professional guidance; and to Madeleine Akrich, Thomas Franssen, Michael Guggenheim, and Alexander Rushforth for their useful feedback on earlier versions of this chapter.

NOTES

1. A third researcher, Julia Quartz (JQ), was later added to the team, but after we started working on this publication. See Quartz et al. (2013) for a full account of the hospital study.
2. The research was funded by the Netherlands Organisation for Health Research and Development (ZonMw).
3. There are international differences in the introduction of rankings and other comparative techniques; in the UK, hospital rankings were introduced in 1983, while in the US the first rankings appeared in the early 1990s. In the Netherlands, the first ranking was published in 2004 (see Pollitt et al. 2010 for a comparative analysis of Dutch and English hospital ranking systems).
   See p. 5. In this comparative framework, Prof D is expected to be highly competitive; his H-index was 91 at the time of applying for funding [E-mail from one of his postdocs, 17 September 2012, in which the postdoc used Google Scholar for the calculation of the H-index]. This is quite high in his own field (basic research). The number will certainly stand out when compared to more clinically oriented medical scientists in the funding scheme.
5. Case mix refers to the characteristics of patients treated on the ward in terms of age, sex, co-morbidities, and the like, which might affect outcomes of clinical work.
7. We borrowed this sentence from Madeleine Akrich’s review of an earlier version of this chapter.
BIBLIOGRAPHY

Akrich, M., and V. Rabharisoa, ‘Pulling Oneself Out of the Traps of Comparison: An Auto-ethnography of a European Project’, this volume


Bal, R., and F. Mastboom, ‘Engaging with Technologies in Practice: Travelling the North-west Passage’, *Science as Culture*, 16.3 (2007), 253–266


Currie, G., R. Dingwall, M. Kitchener, and J. Waring, ‘Let’s Dance: Organization Studies, Medical Sociology and Health Policy’, *Social Science and Medicine*, 74.3 (2012), 273–280


de Rijcke, S., and A.D. Rushforth, ‘To Intervene, or Not to Intervene, is that the Question? On the Role of Scientometrics in Research Evaluation’, *Journal of the Association for Information Science and Technology*, forthcoming


Hirsch, J., ‘An Index to Quantify an Individual’s Scientific Research Output’, *PNAS*, 102.46 (2005), 16569–16572 
Jacobs, R., M. Goddard, and P. C. Smith, ‘How Robust are Hospital Rankings Based on Composite Performance Measures?’, *Medical Care*, 43.12 (2005), 1177–84 

Stöckelová, T., ‘Frame Against the Grain: Asymmetries, Interference, and the Politics of EU Comparison’, this volume


