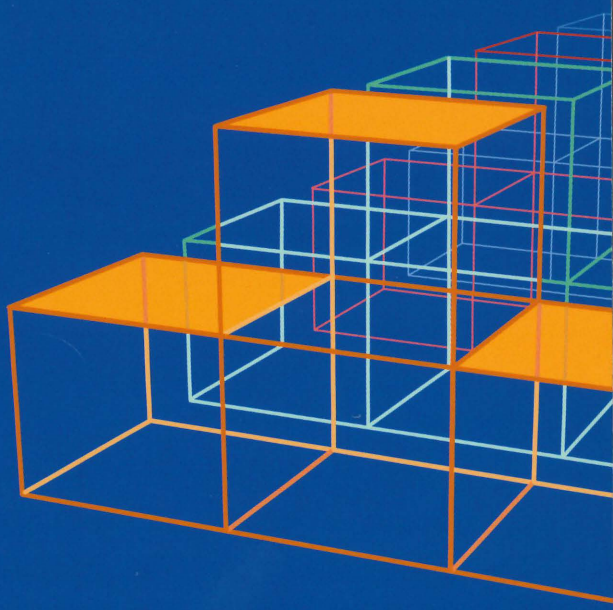
EDITIONS DE L'UNIVERSITE DE BRUXELLES



EDITED BY
CATHERINE DEHON,
DIRK JACOBS
AND CATHERINE VERMANDELE

Ranking universities

éducation

Ranking universities

Edited by
Catherine Dehon, Dirk Jacobs, Catherine Vermandele

ISBN 978-2-8004-1441-6

Avenue Paul Héger 26
1000 Bruxelles (Belgique)
EDITIONS@.ulb.be
http://www.editions-ulb.be

Imprimé en Belgique

Acknowledgements

This book results from an international conference organised in December 2007 at the Université Libre de Bruxelles on the topic of "Ranking and Research Assessment in Higher Education". This conference took place within the framework of the European PhD in Socio-Economic and Statistical Studies, under the presidency of Jean-Jacques Droesbeke of the Université Libre de Bruxelles. It was organized by Catherine Dehon, Mathias Dewatripont, Dirk Jacobs and Catherine Vermandele. We sincerely wish to extend our thanks to all those who made the conference and the resulting publication possible.

In particular we would like to thank Pascal Delwit, Bram De Rock, Mathias Dewatripont, Philippe Emplit, André Sapir, Françoise Thys-Clément and two anonymous referees for their insightful comments and stimulating discussions.

We wish to acknowledge the support of the Interuniversity Attraction Pole (IAP) Programme of the Belgian Federal Science Policy Office (BELSPO) "Higher Education and Research : Organization, Market Interaction and Overall Impact in the Knowledge-Based Era". We are also grateful to the Fonds de la Recherche Scientifique, the Université Libre de Bruxelles and the Faculté des Sciences Sociales et Politiques / Solvay Brussels School of Economics and Management for their financial support.

Finally, we are grateful to Michele Mat, Editor-in-chief of the Editions de l'Université de Bruxelles, and Marjorie Gassner for her advice.

Foreword

An international conference entitled "Ranking and research assessment in higher education" took place in Brussels at the Université Libre de Bruxelles in December 2007. Some highly important questions were debated during the two days with leading international experts, a number of which were invited to give scientific contributions to the present book.

As appeal, dangers, merits and future challenges of ranking systems are discussed in depth in the introduction made by Dehon *et al*. hereafter, I will focus this preface on research assessment.

Research assessment has been gradually introduced this last decade in European universities and is presently being developed in many other countries. The question of research evaluation is closely related to two hot topics: the ranking of universities and higher institutions, and their funding.

Indeed, most ranking systems take into account some elements of research performance, and there is an increasing tendency in many countries to link part of the funding of an institution to research output scores. This raises the crucial question of what is the best way of assessing the research performance of an institution.

Evaluation instruments are indeed numerous: they are based on peer review or on metrics, focus either on individuals or on research groups, deal with ex-ante or ex-post assessments, consider only research output production or take into account its quality, … All types of evaluation methods have their own respective advantages and limitations as well as typical bias.

One should keep in mind that the generic goal of any evaluation process is to provide useful feedback to a wide audience. With regards to research evaluation, the targeted audience may be a university's research management, national authorities, funding agencies, sponsors, …

The evaluation process should obviously be conducted in a different way depending on the targeted audience, on the goals, and on the level of governance.

While university management might be interested in an effective policy instrument aimed at remediation and research quality improvement, for which there appears to be a common agreement on the need for peer reviews, the focus is gener-

ally only put on output parameters (publications/citations or other) when ranking or funding are concerned. This raises the question of whether it makes sense to attribute funding either exclusively or even in part based on research. It is important to consider the diversity of a university's missions and its characteristics in order to avoid being confined to a single normative framework.

The use of quantitative indicators in international comparisons and national allocation models makes quantitative research output scores a major issue for universities, even when they are aware that these do not necessarily reflect the quality of their research. Several questions have yet to be answered, such as:

— Should the framework for research assessment and funding make greater use of quantitative information?
— How should we handle the much needed development of a method for producing bibliometric quality indicators? Are indicators such as the impact factor, the citation index, the h-index or the Crown-indicator the most appropriate? What are their possible biases? What restrictions should be put on their use?
— Can a metrics-based system of assessment be used for all subjects, including arts, humanities and social sciences?
— How to assess the relationship between an evaluation and its impact (on various time-scales)?
— While transversal research should be encouraged to generate cross-fertilization, can it be evaluated properly using only metrics?

Considering the importance of these questions and the potential impact of the current and future evaluation assessments on the development of the higher education institutions, it is essential that academic actors as well as evaluation agencies share their expertise and concerns on these topics within the frame of such symposiums.

Prof. Véronique HALLOIN
Secrétaire générale du Fonds de la Recherche Scientifique-FNRS

Ranking and research assessment in higher education: current and future challenges

Catherine Dehon, Dirk Jacobs and Catherine Vermandele

Summary

University rankings are "hot". Some universities, policy makers and journalists seem to take them quite seriously. At the same time, however, they are fiercely criticized. The best known worldwide rankings tend, for instance, to have a strong anglo-saxon bias and tend to give insufficient valorisation to human sciences. Are improvements and alternatives possible? Should universities care about rankings and let them influence their practices? Parallel to international rankings, research assessments have become increasingly important in several countries. What are the current practices of research evaluation? What are the challenges, obstacles and advantages? How should one assess the quality of research in a fair and balanced way? These and other questions were debated with leading experts at an international conference entitled "Ranking and research assessment in higher education" held in Brussels at the Université Libre de Bruxelles on 12 and 13 December 2007. Looking back on what was a very enriching conference, we decided to invite a number of contributors to bundle their papers into one volume. We are confident that it will prove to be a useful instrument for informed debate, not only at our university, but also in the broader Belgian and European academic landscape, on the topical issues of ranking and research assessment.

1. The appeal, merits and dangers of ranking

All kinds of ranking endeavours appear regularly in the press and the general public seems to love them. There seems to be some kind of fatal attraction to competition and excellence made visible in numbers. A ranking tries to summarize by some statistics quite complex patterns of behaviour. These days, the media bombard us with all kinds of rankings with different degrees of relevance and quality, be it the *Migrant Integration Policy Index*, the latest news from the *Guinness book of world records*, the *Forbes lists* of the world's richest people or new listings of the best – or even worst – dressed women of the planet. You find them in all sorts of fields from sports (*FIFA/Coca Cola World Ranking* in football, *ATP ranking* of tennis players, …), to culture (*Top of the Pops* for pop and rock music, *Oscar Academy Awards*,

Golden Globe Awards and other *Césars* for the movie world, …) to academic life (*Academic Ranking of World Universities* from Shanghai Jiao Tong University, *CHE Research Ranking*, *European business schools ranking* of the *Financial Times*, ...). It is this last category of ranking exercises which will be of interest to us in this book.

If a ranking is merely a chart made for amusement, like deciding who is the most handsome man in the world, it is a fairly innocent competition. However, as rankings can make or break reputations, they are often to be taken rather seriously. It matters quite a bit whether your restaurant has one star, two stars or no star according to famous restaurant guides like the *Michelin*.

In the field of higher education, ranking is not a matter of frivolity – or at least should not be one. Ranking is closely linked to the topic of evaluation. In the academic world there is a strong tradition of evaluation which is taken very seriously. Traditionally it is done through procedures of peer review of scientific publications. However, increasingly, academic evaluations are being undertaken all over Europe in addition to the traditional (publication related) peer review procedures. They help to evaluate academic performance, not only on the level of individual researchers, but of entire research centers, departments and even universities. In their contribution to the third section of this book, Roel Bennink, coordinator of *Quality Assessment Netherlands Universities*, and Seamus Hegarty, one of his evaluators, shed light on the Dutch experience with research evaluation of this kind. The Netherlands has quite a bit of experience in the matter which already dates back to 1992.

Parallel to this increasing culture of accountability and evaluation, a number of widely covered worldwide rankings of universities have been appearing. They give us an idea of the strength of universities on a global scale but are often undertaken with far less methodological scrutiny than one would wish for.

There is no doubt that, whether we like it or not, evaluations and rankings are rapidly gaining importance in the field of higher education and are here to stay. They are legitimized by referring to the increased competition in recruiting students and staff in the wake of European efforts towards alignment of universities through the Bologna process and in light of wider patterns of globalization in the academic field. Although in May 2006 the *International Ranking Expert Group* established 14 criteria – mainly pertaining to transparence and methodology (see the *Berlin Principles on Ranking of Higher Education Institutions*) – which should be respected when classifying in the field of higher education, one cannot ignore the fact that most of the current ranking efforts still have serious flaws despite this effort towards regulation and higher quality standards. The most widely cited rankings often suffer from a severe bias towards the "big" universities since sheer volume of production is rewarded more than productivity in the sense of efficiency. They also have a bias in favouring institutions from the Anglo-Saxon world – since scientific publications in journals published in English are the main reference – and favour those universities which excel in exact and biomedical sciences (rather than in human and social sciences) – since they are better represented in the databases of the scientific publications that are being used.

Given the widespread appeal of ranking exercises, not least among policy makers, it is important that the scientific world takes an interest in the debate on

ranking of higher education institutions, and more particularly universities. On the one hand to critically assess the different procedures which are currently being used to create these ranking systems and pinpoint their weaknesses, on the other hand to help identify good methodological practices for evaluating universities in order to be able to make reasonable comparisons.

Just as international profiling agencies make credit and financial ratings of companies and financial institutions (Standard & Poor Rating System, Fitch System, …) allow them to rapidly evaluate their risk level, it is now perhaps necessary to undertake a global analysis of the methodologies used in ranking endeavours in the academic field to filter out the most relevant indicators for a suitable evaluation system. Indeed, fundamental questions have to be taken a lot more seriously: the precise definition of studied variables, the validity and implications of methods of normalization which are being used, the vulnerability of ranking due to the uncertainty of one or several indicators, the issue of statistical inference, the acknowledgment of different potential biases, etc. These are all issues which need to be addressed.

2. What is wrong with current ranking efforts?

Let us just consider three topics where current rankings seem to be failing. A first crucial question that needs to be addressed pertains to seemingly obvious issues of operationalization and research design: the definition of the object of study ("what is a university or a higher education institute?") and the variables to be monitored. The question "what is a university?" is not a stupid one in this context, since the institutional structure of the higher education landscape can differ greatly from one country to another. Countries like France, for instance, which have important research centers (that function semi-autonomously from universities) risk to be penalized if these are not counted as higher education institutes. Then there is the question of what we want to measure exactly and subsequently on what basis we want to rank. Are we to focus on indicators to assess the quality of research or the quality of education? Is due attention given to the innovative nature of research and usefulness of research for society? Does it matter if the university "gives back to society" or is instead locked up in an ivory tower? To what extent is the internationalization of research teams, staff and students taken into account? Does the university care about equal opportunities and gender issues? Do we merely look at the crude production level of a university in terms of number of publications or do we equally pay attention to its efficiency (what is being produced with what kind of resources)? Although of huge importance, these basic questions have remained largely ignored or under investigated.

Scientific common sense tells us that it is impossible to construct one simple indicator which can adequately reflect the different strengths and weaknesses of universities in doing research, offering education and "giving back to the community". Indeed, a good indicator has to be consistently constructed in a stable way and measure exactly what we want to measure and this seems to be impossible here. From the moment you start combining different indicators which aim to measure completely different types of behavior, decisions have to be made about weighting of dimensions. Such weighting procedures will never be neutral. Indeed, opinions

will vary on the importance of all kinds of academic activities. We therefore think it would be reasonable and preferable to construct separate scales and develop specific indicators for each of the three main missions of the university, at least: research (the construction of knowledge), education (the diffusion of knowledge), and service to society (the valorization and use of knowledge). Most probably, for each of these main missions, further sub-dimensions have to be distinguished. In such an exercise specialized bodies of knowledge and in-depth theoretical reflection on the issue – as for instance offered in the body of work brought together by Dewatripont, Thys-Clément and Wilkin (2001, 2002, 2008) – has a crucial role to play. One should then look for appropriate proxy indicators for each of the sub-dimensions of the three main missions of academia.

A second issue pertains to methods of data analysis. How do we properly investigate the relations which exist between different indicators and variables and construct appropriate models to analyze them? Currently mainly – if not only – simple linear relations are taken into account. Is this the best strategy? In addition the issue of normalization of variables in light of comparison must be studied very closely. In our opinion, the choice of normalization strategy must be neutral with regard to ranking, which is hardly the case in current rankings. Taking for instance the ranking of the *Times Higher Education Supplement* and comparing the ranking of 2006 (based on a normalization strategy awarding 100 points to the best performer in each variable) and the ranking of 2007 (using a normalization strategy based on the "z-score", that is, a normalization of variables by centering them in relation to the average result and dividing them by the standard deviation), we observe important changes in the ranking of institutions while we know there have not been important changes in their performance record. This is, of course, unacceptable. Not to mention that statistical distributions which are associated to research related variables are highly asymmetric and thus unsuitable for "z-score" transformations – which are more appropriate in the context of, for instance, exam scores where a normal distribution will be more common. In his contribution to the second section of this volume, Philippe Vincke, after having shortly discussed the characteristics of two of the most well known university rankings (i.e. the Shanghai ARWU Ranking and the Times Higher Education Ranking), demonstrates the vulnerability of rankings linked to the choice of the normalisation procedure.

As far as the issue of measurement of volume of production in opposition to the level of productivity and efficiency is concerned, it is widely assumed that the stakeholders in the academic world (students, researchers, professors, political authorities, …) are mainly interested in the prestige of universities and hence primarily look at the sheer volume of production and its impact. However, this strategy adopted by most ranking efforts sometimes leads to clustering with little relevance for policy makers (Hazelkorn, 2007). Policy makers should in fact be much more interested in getting "value for money". Furthermore, it is of course legitimate to question the idea that "bigger is better" (Mohrman, 2007). Faced with such criticisms, the *Academic Ranking of World Universities 2007* of Shanghai Jiao Tong University has recently introduced a sixth variable (*Size*, since 2008 called *PCP, per capita performance*)

which relates the performance of the five other criteria used[1] to the size of the institutions, with a weighting of ten percent. Of course, this strategy – using a ten percent weight – does not solve the fundamental problem at all. One could even say that this attempt to find a compromise between productivity and efficiency even obscures the issue at stake. We call for more clarity here. Production level and efficiency are in some cases interrelated and certainly complementary variables, but we still need a separate indicator for each of them. We need to measure production levels first and then create an indicator of efficiency by controlling volume of production for size of the universities (in terms of researchers, professors, students and funding) and assess them separately.

Indeed, once good measurements of production levels, productivity and efficiency have been established, the choice of the use of either the indicator for the volume of production or the productivity / efficiency indicator will depend on the stakeholder. A student may be mainly interested in the prestige of a big university if she has no financial limits and no mobility constraints. These are only the happy few; a student faced with more constraints might perhaps find productivity – and hence relative prestige – much more important as a decision criterion. Some students might prefer a small university in which personal interactions with the staff are easier to establish – rather than the reputation of the university – or simply opt for the university which is closest to the place where he or she lives. Likewise, a researcher or professor will in most cases probably take into account both the indicator for volume of production as the efficiency indicator for a given university – and not in the least a set of personal considerations with regard to the expected quality of life – when making decisions about moving to another city or country (or not doing so). Furthermore, political authorities wanting to finance their universities on the basis of "excellence" criteria, would in our opinion make a mistake if they based their decisions merely on crude production levels, which depend heavily on size. Instead, they should at least also look at efficiency as a funding criterion and reward *good practices,* unless they want to end up with a highly skewed two-tiered system. Is the very unequal system of financing of universities currently being used in the United Kingdom – heavily rewarding productivity and thus in practice mainly stimulating Oxford and Cambridge – adequate? Such questions bring us to another debate than the one on ranking in the strict sense, but if the answer is yes, then at least the practical and political consequences of focussing on production volume should be clear for everyone (i.e. a strong incentive for all talent to flock together to one place and second class status to be given to other universities).

Finally and perhaps even most importantly, ranking (and evaluation) should take the diversity of tasks of universities and the diversity across disciplines into account. As we have already stated, a crucial component for establishing a valid and reliable ranking is the issue of proper operationalization of indicators – and suitable data collection procedures – relating to all three central missions of the university: research,

[1] (1) Alumni of an institution winning Nobel Prizes and Fields Medals; (2) Staff of an institution winning Nobel Prizes and Fields Medals; (3) Highly cited researchers in 21 broad subject categories, (4) Research Output Articles published in *Nature* and *Science*; (5) Articles indexed in *Science Citation Index-expanded*, and *Social Science Citation Index*.

education, and service to society. As far as research is concerned, methodologies are fairly well developed and two distinct – partly competing – strategies coexist: bibliometrics and peer review. Several indicators are suggested in the literature (number of articles, number of citations, Hirsch-index, number of citations per article, etc.) but there is no privileged indicator that stands out by receiving support from all stakeholders. Furthermore, what works in one academic discipline, does not necessarily make sense in another field of science. A composite index seems appropriate here. In their contribution to section 5 of this book, which focuses on bibliometric indicators, Wolfgang Glänzel and Koen Debackere suggest a number of alternative strategies for taking into account the diversity of and within higher education institutes.

Interestingly, several studies have shown that there are fairly good correlations to be observed between bibliometric measures and peer review results (Rons, 2008; Williams, 2007) but that they tend to vary significantly according to the research domain under scrutiny. A specialised analysis is hence necessary within the university system to take into account the diversity of academic disciplines. Several economists (for instance Coupé, 2003; Combes and Linnemer, 2003) have suggested methods to classify (their) economics departments. In section 6 of this volume, Michel Lubrano examines the results of different bibliometric mesures in trying to identify significant differences between economics departments. One could imagine similar specialised exercises for other scientific disciplines.

As far as education is concerned, potential criteria are abundant and a theoretical approach is indispensable to construct proxies for the different dimensions linked to educational practices. In our opinion, equal opportunities (and hence the diversity of the student population in terms of class, gender and ethnic background) are certainly an important aspect to take into account. Let us note that some tentative efforts have been undertaken to equally start monitoring and analysing the extent to which the university fulfils its third mission of rendering services to the society (Montesinos, 2007).

3. Challenges for the future

Although the Shanghai ranking is fiercely criticized – notably because of the choice of indicators and the way there are combined –, it does have the merit of being one of the more transparent ranking systems to assess the quality of research in higher education institutions at world level. Taking as a given that, for the time being, there does not seem to be a better instrument – with the same universal appeal – to measure the quality of universities, Aghion and his colleagues use the Shanghai ranking in their comparative analysis of higher education systems in their contribution to section 7 of this volume. The Shanghai ranking may be problematic, it can nevertheless help us out in launching the debate on performance levels of universities. Measuring the reputation or quality of a higher education institute is surely rife with challenges. Crude indicators like those used in the Shanghai ranking can, nevertheless, have some heuristic value. Let us, for the sake of argument, compare it with another blunt instrument: the gross domestic product (GDP) as a measure of the value of all goods and services produced within a country during a certain period of time. Most people are well aware of the limitations of an indicator like GDP: it does,

for instance, not take into account domestic work, voluntary work, irregular work or hidden social and ecological costs. Nevertheless, GDP is a widely used indicator in the scientific literature and it would be difficult to imagine that economists, political scientists, social geographers and sociologists would simply stop using it altogether because of the inherent conceptual problems and issues of measurement error. At the same time we should not forget about the basic flaws and limitations of crude measures like GDP. The same holds for the use of crude indicators like those of the Shanghai ranking. So if Aghion and colleagues use it as a heuristic tool to develop their argument, we cannot ignore the limitations – for instance that it only can claim to measure the research dimension in an assessment of quality of higher education institutions – but do not have to throw away the baby with the bathing water. Clearly, in the long run, there is need for more sophisticated and more valid measures of such a complex issue as the quality of a higher education institute. We need transparent, reliable and genuinely comparable data to do cross-national evaluations.

A number of recent initiatives suggest we are moving in the right direction. In 2006 the *Programme for Institutional Management in Higher Education* (IMHE) of the Organisation for Economic Co-operation and Development (OECD) and the *International Association of Universities* (IAU) proposed to embark on a study of the positive and negative effects of rankings on strategic and administrative decision making by universities (Hazelkorn, 2007). One of the fears is that rankings would lead to exaggerated forms of competition. Efforts to compare might degenerate into a process in which diversity becomes the victim of pressures towards convergence around a single normative framework, embodied by certain American universities which currently top the rankings. In a second phase this project started to analyse adaptive behaviour in three countries (Germany, Australia and Japan). The OECD did not stop short there. It has now started an international assessment program aiming to evaluate the outcomes of higher education policies (AHELO – *Assessment of Higher Education Learning Outcomes*). Due to the democratisation of higher education the number of students has doubled during the last ten years. Inevitably, this has consequences for public spending and calls for a procedure of quality assurance. AHELO aims to measure educational achievement at the university level. A pilot study is being undertaken at the moment of the writing of this introduction and its results are expected to be available at the end of 2009. The aim is to verify whether it is possible to test university performance levels using criteria which are independent of linguistic, cultural and geographic factors. A test will be offered to students of the first cycle of academic education in order to assess their general skills (critical reflection, analytical reasoning, etc.) and specific competences linked to their disciplines (engineering and economics have been proposed as the test disciplines in the feasibility study). If achievable, AHELO might well grow out to be the follow up study of the famous PISA studies (Programme for International Student Assessment) which monitor competencies of 15-year-old pupils around the world on a three yearly basis. AHELO will then be able to assess performance levels of the average student in a country, or even of the average student of particular higher education institutes in a country. However, it will not be able to function as a direct test of the quality of teaching at the university level. Indeed, higher education institutes have different

selection criteria at entry and differential policies with regard to equal opportunities for disadvantaged groups, two important factors which could have an impact on overall results. Furthermore, the composition of the student population can be very different with regard to the typical socio-economic profile (and this might not only vary from region to region but even from discipline to discipline), another factor which might impact on educational outcomes. Conscious of these limitations, the OECD will study how one can try and take the blurring effects of the quality of prior education and selection criteria at entry into account. The statistical challenge will be huge, but it would indeed be quite an achievement to try and measure the added value of higher education institutes in terms of educational performance and teaching efficiency. The OECD is equally concerned about the multiple dimensions and diversified tasks of universities which should be taken into consideration in a quality assessment. As a result, a number of contextual criteria will be taken into account such as the quality of equipment and infrastructure, the international openness, the diversity of outlet possibilities, the quality of research and so on. Interestingly, this set of criteria will be based on the methodology proposed by the German Center of Higher Education Development (CHE). The work of CHE, making use of a multidimensional analysis of universities, is presented in the fourth section of this volume in the contributions by Gero Federkeil and Sonja Berghoff.

Another interesting new development are the plans of the European Commission's Directorate General for Education and Culture. They have launched a call in December 2008 for an assessment exercise along the lines proposed by CHE, that is, a multidimensional analysis in which the complexity of universities can be addressed. This exercise will also need comparable, reliable and transparent data. Without having a ranking as the prime objective, there is a wish for a comparative evaluation of higher education institutes in order to show the strengths and weaknesses but also the diversity of the academic institutional field. Contrary to the OECD project in which the primary focus is on education, this European project wants to take into account all missions of the university in its assessment.

It goes without saying that a qualitative higher education system is of crucial importance for the society at large. From the perspective of equal opportunities it is essential to be able to combine the creation of an elite which excels in terms of research with a mass educational system of the highest possible quality. Europe has an interest in breaking the hegemonic position of the United States in most of the scientific fields and assuring a large degree of liberty and diversity both in educational as in research related matters. It can help if students, decision makers and stakeholders get a clearer picture of the current state of the academic field, based on reliable and publicly available data of greater quality that the current set of global rankings.

4. What this book offers

We have asked a number of leading experts who participated in an international conference we organized on the subject at the *Université Libre de Bruxelles* in December 2007 to shed some light on some of the methodological issues with regard to evaluation and ranking in the world of higher education. Let us briefly present each of these contributions, a number of which we have already referred

to earlier in this introductory chapter. Philippe Vincke, the Rector of the *Université Libre de Bruxelles,* at our conference and in his opening address for the academic year 2007-2008 severely criticized the *Shanghai ranking* and the ranking undertaken by the *Times Higher Education Supplement*. He pinpoints the fragility of rankings due to the insufficient normalisation strategies which are being used.

Roel D. Bennink, coordinator of *Quality Assurance Netherlands Universities (QANU)*, presents the long track record of Dutch universities with regard to evaluation. A systematic external evaluation procedure was put in place in 1992 and was evaluated itself in 2003. Interestingly, in contrast to the UK system, the procedure has no direct consequences for the financing of universities. One of the peer evaluators participating in the QANU-system, Seamus Hegarty, helps us to understand as a prime witness what the different challenges are in undertaking such a difficult task as a scientific evaluation.

In order to solve a number of problems and address a number of criticisms with regard to other rankings, the *Center of Higher Education Development* (CHE) has since 1998 made a number of multivariate ranking systems available. Gero Federkeil and Sonja Berghoff present the experience of the CHE Research Ranking of German Universities. Gero Ferderkeil focuses on evaluation at the University level, and Sonja Berghoff presents rankings which covered no less than 16 academic disciplines.

In their contribution, Wolfgang Glänzel and Koenraad Debackere (2008) stress the complexity of universities and the difficulty of correctly using bibliometrics in assessing the quality of research at universities. They propose three strategies to take into account the biases linked to differences across disciplines or differential structures of universities: clustering of similar universities, a breakdown by field and standardization of indicators.

Michel Lubrano, in a more technical contribution, presents us with models to measure the quality of research in economics departments across Europe. Based on his analysis, Lubrano proposes test procedures which should allow us to verify whether observed differences between departments are statistically significant.

We conclude with a contribution by Aghion, Dewatripont, Hoxby, Mas-Colell and Sapir who reflect on the issue of financing and autonomy of universities in Europe. Indeed, the sudden enthusiasm for university rankings has in recent years also triggered a more general reflection and debate on (the lack of) European excellence in research. The issues of finance and autonomy help us to explain quite a bit of the gap between American and European universities in the rankings. Of course, while the study documents a correlation between the Shanghai ranking on the one hand, and financing per student as well as measures of autonomy on the other, one has to be cautious. This is only a first step in the understanding of university performance.

Clearly, there is still quite some scientific (and policy) work to be done in order to construct a set of appropriate indicators for ranking and research assessment. As we have argued, these will have to do justice to the complex reality and the different tasks of universities, will need to avoid arbitrary methodological choices and will have to empower different actors of the academic world to make adequate comparisons at the world level, or at least at the European level, possible. We have not reached that stage just yet, but prospects for the future are promising.

References

Assessment of Higher Education Learning Outcomes (AHELO): www.oecd.org/edu/ahelo.

Berlin Principles on Ranking of Higher Education Institutions. [Retrieved June 2006], www.che.de/downloads/Berlin_Principles_IREG_534.pdf, *International Ranking Expert Group* 2006.

Combes P.-Ph., L. Linnemer (2003), "Where Are the Economists Who Publish? Publication Concentration and Rankings in Europe Based on Cumulative Publications", *Journal of the European Economic Association*, 1(6), pp. 1250-1308.

Coupe T. (2003), "Revealed Performances: Worldwide Rankings of Economists and Economics Departments, 1990-2000", *Journal of the European Economic Association*, 1(6), pp. 1309-1345.

Dewatripont M., F. Thys-Clément and L. Wilkin (2001), *The Strategic Analysis of Universities: Microeconomic and Management Perspectives*, Editions de l'Université de Bruxelles, Brussels.

Dewatripont M., F. Thys-Clément and L. Wilkin (2002), *European Universities: Change and Convergence*, Editions de l'Université de Bruxelles, Brussels.

Dewatripont M., F. Thys-Clément and L. Wilkin (2008). *Higher education in a globalized world: governance, competition and performance*, Editions de l'Université de Bruxelles, Brussels.

Hazelkorn E. (2007), "Learning to Live with League Tables and Ranking: The Experience of Institutional Leaders", *International Ranking Expert Group*.

Hazelkorn E. (2007), "L'impact du classement des établissements sur la prise de décision dans l'enseignement supérieur", *Politique et gestion de l'enseignement supérieur*, OCDE, vol. 19(2), pp. 95-122.

Marginson S. and M. Van der Wende (2007), "Globalisation and Higher Education", OCDE, EDU/WKP 3.

Mohrman K. (2007), "Educational Exchanges: What World-Class Universities Should Not Adopt from US Higher Education", WCU.

Montesinos P. (2007), "Third Mission Ranking for World-Class Universities: Beyond Teaching and research", WCU.

Rons N., A. De Bruyn and J. Cornelis (2008), "Research evaluation per discipline: a peer-review method and its outcomes", *Research Evaluation*, 17, pp. 45-57.

THES-QS 2006, http://www.topuniversities.com/worlduniversityrankings/results/2006.

THES-QS 2007, http://www.topuniversities.com/worlduniversityrankings/results/2007.

Williams R. (2007), "Peer opinion and performances measure? Measuring the institutional standing of Australian universities", *International Ranking Expert Group*.

University rankings

Philippe VINCKE

Summary

Over the past few years, there have been significant developments in higher education which are highlighted by the following two major features:

a) A significant increase in the number of policymakers involved in higher education and research: new generously funded universities are appearing in emergent countries; closer to us, private institutes are being created, prestigious universities are setting up branches all over the world, university degree programmes are proposed on the Internet, …

b) Increasing student and teacher-researcher mobility, undertaken directly by the public authorities of every country.

Students and researchers are now intent on being informed of the quality of higher education institutions where they may one day study or with whom they will be collaborating. This legitimate concern goes hand in hand with an increasing demand from the public that universities justify the means at their disposal and be accountable to those who fund or subsidise them.

KEYWORDS. Ranking, sensitivity analysis, standardization, universities.

1. Introduction

Research assessment is a tradition in the scientific world. Researchers are used to discussing their findings with their peers. Articles to be published in scientific journals or presented at international conferences are first submitted to and reviewed by other specialists who may accept, edit or reject them. Researchers generally know who the specialists in their own field are and where the top teams are located. Even without any ranking in the press, quality assessment is part and parcel of a teacher-researcher's life, at least in the case of his own research activity. At this stage, the same cannot be said of teaching activities.

The novelty – and this is a direct effect of the recent growth in the supply of higher education and the increasing mobility of students and researchers – is increased competition between universities to attract the best students or researchers

and of course, to secure funding. This has led to comparisons between institutions, hit-parades and rankings.

Two types of rankings regularly make the headlines in the media (which is reluctant, however, to seriously analyse the methodology used): that of the Shanghai Jiao Tong University (called the "Shanghai ranking") and the yearly ranking published in the *Times Higher Education Supplement* (hereafter referred to as the "*Times* ranking").

Universities can no longer ignore these rankings. A good position in a ranking has become an argument for promotion, the calling card of the university, the *sine qua non* condition for valuable international collaboration. These rankings are therefore going to influence the global university landscape at an increasing rate (see Hazelkorn, 2007a).

Any assessment is the result of a comparison and at this stage, a distinction must be drawn between two radically different situations:

— *First situation*: goals or standards are set *a priori* and one's activities (one's achievements) are compared to these goals or standards. In this case, we are in a context of *absolute assessment*, even if the results of the assessment depend on standards that were set *a priori*. For example, take safety standards for cars. If you compare the features of your car with these standards, you may consider that, as far as safety is concerned, your car is very safe, fairly safe, or not very safe.

— *Second situation*: your performance is compared with that of your colleagues or competitors. We are now in a context of *relative assessment*. When comparing the safety of your car to that of your neighbour's, you may consider that your vehicle is more reliable, as reliable as or less reliable than your neighbour's. Such a conclusion does not, however, tell you anything about its level of absolute safety that is, a comparison to a system of standards. Your car may be more reliable than your neighbour's and, at the same time, not be very safe. Conversely, it may be less reliable, but very safe.

The Equis accreditation system[1], which assigns an international label of quality to Business Schools, is based on absolute assessment. It compares the Schools' parameters to predetermined standards, which is not the case for the Shanghai university rankings and the *Times* ranking, which are based on relative assessment. These methods will now be analysed.

2. The Shanghai ranking

A. Presentation of the Shanghai ranking

Depending on the university under consideration, the Shanghai ranking is based on 4, 5 or 6 criteria.

The first criterion used to measure the quality of education of a university is the number of alumni who were awarded a Nobel Prize or a Fields Medal. An alumnus is

[1] The Equis accreditation is delivered by a private association, the European Foundation for Management Development (EMFD), see the webpage: http://www.efmd.org/.

defined as a person who holds at least one degree from the university being assessed; such a person counts as one unit if the degree was obtained after 1990, as 0.9 if it was obtained between 1980 and 1990, etc.; it counts as 0.1 if the degree was obtained between 1900 and 1910.

Two criteria measure the quality of the teaching staff of a university:

— The number of Nobel Prizes and Fields Medals awarded to the academic staff teaching at the university under review; here too, the weight decreases with seniority and a complex fraction system weights the results if the laureate was active in several universities simultaneously and/or if several laureates shared the same prize;
— The number of teachers – researchers of a university who are among the 250 most cited authors for a given period of time (for the 2006 ranking, the 1981-2003 period was retained) within 21 major subject categories (this will be developed below). The data for this criterion come from a commercial database (the *Essential Science Indicators database*), distributed by Thomson Scientific, a firm based in Philadelphia. For lack of specific information on the methodology used to draw up this data, it has been impossible to reconstruct it and hence to check it out.

The fourth criterion is the number of articles from the university under review published in the journals *Nature* and *Science* within the last 5 years, with a weight system for co-authored publications. Since these two science journals favour exact or laboratory sciences, this criterion is not taken into account if the university is not active in these fields: its weight is then redistributed among the other criteria. This is the case, for example, of the London School of Economics, an institution of renown which specializes in humanities and social sciences.

The fifth criterion is the number of articles from the university under review listed, for a given period, in the *Science Citation Index* and the *Social Science Citation Index*. These two listings, published by Thomson Scientific, establish statistics on the number of papers cited in other papers and, therefore, to a certain extent, on the fame or impact of scientific publications.

In the first version of the Shanghai ranking, only those 5 criteria were taken into account (reduced to 4 for social science institutions). The authors then realised that their ranking was strongly correlated to the size of the universities: all of the criteria were expressed in terms of absolute numbers (number of awards and prizes, citations, publications…) independently of the number of researchers working in each university.

An attempt was made to correct this flaw by adding a sixth criterion defined as follows: for each university the scores for the first 5 (or 4) criteria are summed up and then divided by the number of teachers – researchers in the university under consideration. The result obtained is the score for the sixth criterion for that university. However, as the authors did not have data available for all of the universities, this additional criterion was not taken into account for all universities. Then a number of universities are assessed on the basis of 4 criteria, others on 5 and others on 6.

For each of these 4, 5 or 6 criteria the highest-ranked university is given a score of 100; the others are given a score which is a mark out of 100, obtained using a simple rule of 3.

The global score of a university is the weighted sum of the scores obtained for the different criteria. Each criterion is worth 25 % of the final result when there are 4 criteria, and is worth 20 % when there are 5. When there are 6 criteria, the first one (the number of Nobel prizes awarded to alumni) and the sixth (introduced to reduce the size effect) are worth 10 % each, while the other four are each worth 20 %.

It must be added that the authors claim they corrected some scores when an anomaly was noticed. However, no clear explanation was given as to what they considered as an anomaly or how they proceeded to make these corrections.

B. What comments can be made on the Shanghai ranking?

The authors state the following: "In fact, we do not consider ourselves specialists in scientometry or bibliometrics. However, university rankings increasingly have an influence on the development of global higher education. This is why we hope that competent people will assist beginners".

In other words, the authors admit that they are not experts. They are, in a way, sending a message to those who are indeed competent to tackle the problem seriously.

From his *curriculum vitae*, the main author is a Chemistry professor specialized in polymers who, in 2002, suddenly stopped his scientific activity and turned to the ranking of universities. One of the most embarrassing aspects of his work lies in the grey areas of his methodology (a few of these have already been mentioned), which do not allow the reader to reconstruct the ranking obtained, and hence, to verify it: a shortcoming that needs to be put right if the authors hope to ever see their work recognized as scientific.

This publication started out as a patriotic endeavour, of which the aim was to raise the level of *research* in Chinese universities by encouraging emulation with universities in other countries. This can be seen in the wording of the criteria, since comparisons focus on scientific production, i.e. research. Teaching, student training or cost of studies, for example, are not considered. Moreover, the authors make no comments on what they consider as a relevant use of their results or on how to interpret them.

However, a close examination of the data used and how they were processed prompts us to interpret them with utmost care.

The criteria on which the Shanghai ranking is based will first be examined.

The first two criteria are related to Nobel Prizes and Fields Medals awarded to alumni and researchers. Determining which university should receive an award is not a simple task. Scientists, particularly at this level, often move from one institution to another during their career, and a prize often rewards research that was carried out many years before.

This difficulty has led to inextricable situations. The two major universities of Berlin (the Frei Universität and Humboldt Universität) were excluded from the Shanghai ranking, because it was impossible to determine to which of them

Albert Einstein's Nobel prize should be awarded: awarding the prize to one or the other would have modified its ranking by more than one hundred places in the final score.

Other criteria include figures for citations of authors or papers. Literature abounds on the advantages and drawbacks of citation indices and bibliometric analyses (Moed, 2005). Without going into the details of these studies, which would warrant a presentation in itself, it should be noted that, given the current situation, these tools appear to be relevant in the subject categories that have integrated them in their operating mode (in particular life and health sciences). At this stage, however, these same tools are much less relevant in other fields (see Hicks, 2004).

Overall, experts in bibliometrics agree that some fields, such as engineering science, environmental sciences, social sciences, law and humanities cannot be dealt with satisfactorily using the traditional techniques of bibliometric analysis. Exclusive use of these techniques is therefore likely to introduce a significant bias when comparing universities, depending on which subject categories are represented.

As mentioned earlier, one of the criteria is based on the number of researchers included in the 250 most cited authors within the 21 subject categories retained by Thomson Scientific. Here is a quick overview of their list: Mathematics, Physics, Chemistry, Biology – Biochemistry, Computer science, Geoscience, Space science, Engineering, Materials science, Agriculture, Environment, Clinical medicine, Veterinary medicine, Pharmacology, molecular and genetic Biology, Microbiology, Immunology, Neuroscience. 18 out of 21 subject categories have just been cited and not one branch of Humanities has been mentioned. The last 3 subject categories are Psychology – Psychiatry, Economics and Management, Social Sciences. Philosophy, Arts, History, Archeology, Law, Political Science are not even listed…

A huge imbalance between the different fields is revealed when the number of journals published in each field is examined. The methodology used does not take into account the fact that the volume of scientific production can vary considerably between fields and, therefore, may favour some universities over others, depending on the subject categories they cover.

It also disregards the specificity or cultural aspects of the different fields. For example, in some subject categories, papers are often co-authored by 6, 7 or 10 researchers, whereas in others single-authored papers are a must. Therefore, taking into account the number of times that authors are cited in lists could bias the assessment.

Two criteria relate to the number of publications in *Nature* and *Science* or indexed in the Thomson Scientific database. It must be said that in the case of co-authoring, the three main authors count respectively as 1, 1/2, 1/4 (which in many subject categories does not make sense) and as $1/10^{th}$ for the others.

In other words, the greater the number of authors, the more positive the paper's contribution to the assessment of the source university: by extending this reasoning even further, it could be suggested that, from now on, all colleagues from one university co-author all of the papers produced at this university.

Finally, despite adding a sixth criterion, the size factor continues to play a significant part (Mohrman, 2007), except, of course, for the few universities, which are

regularly, awarded Nobel Prizes. To move up in the Shanghai ranking, the Belgian universities would simply have to merge. This, in fact, is what some French universities have decided to do: the President of the Université Pierre et Marie Curie (Paris 6) recently mentioned that merging with the Université Paris-Sud (Paris 11) would put them close to the eighth world place, without having to make any changes to their educational and research policies.

No mention is made by the authors of the Shanghai ranking of the respective weighting of the criteria, yet this has a major effect on the final ranking. Their Web site invites visitors to set the weights themselves, without any reference being made to the scale units used or the standardisation mode, which is contrary to the basic principles of this type of work.

Notwithstanding the comments on the selected criteria and the bias they cause, one may wonder about the accuracy of the numerical data used.

Minor material errors in large scale bibliometric analyses which range from spelling mistakes in the names of authors, to errors in the input of their affiliation, the issue numbers or the pages of the related journals, cannot be avoided.

A study published in 2002 in the journal *Nature* estimated at 30 % the global error due to this material dross (Moed, 2002). This observation does not apply only to the Shanghai ranking.

The authors of this ranking claim that their data is accurate within a 2 % margin of error, without supporting this assertion in any way. Moreover, no mention is made of any potential error or inaccuracy in the presentation of their results. Yet, 2 %, for example, distinguishes the 45th place from the 60th in the Shanghai ranking. And 30 % – the estimated error in the *Nature* study – distinguishes the 12th from the 100th place in this same ranking.

The affiliation of authors of scientific publications to universities is also a problematic criterion. Many authors do not name the institution to which they belong accurately. For example, "Université Libre de Bruxelles" (ULB) is the official name of the French-speaking university at Brussels, yet many professors of this university refer to it by various names especially – and this occurs frequently – when writing in English: University of Brussels, Free University of Brussels, Brussels University, …

This was verified using a small sample of eminent professors from this university. The correct name "Université Libre de Bruxelles" was cited in less than 10 % of their scientific production. This phenomenon causes a strong bias in favour of Anglo-Saxon universities.

Sometimes, instead of citing his University, the author of a paper will mention his Faculty or his Department, and – for the specific case of ULB – addresses ranging from "avenue Roosevelt" to "boulevard du Triomphe", or "route de Lennik" in Anderlecht or even "rue des professeurs Jeener et Brachet" in Gosselies. The individual or rather the software which processes the data at Thomson Scientific or in Shanghai, is unaware that these different addresses refer to the same university.

Along the same lines, medical research is particularly ill-accounted for: a major portion of this research is carried out in university hospitals and does not always appear in universities' scientific publications. It was recently found that the publica-

tions from the university hospital associated to ULB, "hôpital Erasme", had been linked, to the Erasmus Universiteit of Rotterdam in the Thomson Scientific databases.

Finally, to conclude this analysis of the Shanghai ranking, a technical aspect in itself calls for the greatest care when processing the results of this ranking.

C. Standardisation problem

The Shanghai ranking uses a 0 to 100 standardisation of scales, attributing a 100 point mark for each criterion to the university which ranks first for the given criterion, and by applying a rule of three to the other universities' scores. Each university then receives, as a global score, the weighted average of its results. This technique can be illustrated by a numerical example.

	0,4 C1	**0,4 C2**	**0,2 C3**
a	2,000	500	5
b	1,360	440	10
c	1,600	375	10

This table shows the assessment of three universities a, b and c for three criteria. For the first criterion, the assessed values are respectively 2,000, 1,360 et 1,600 (the figures indicate, for example, the number of researchers). For the second criterion, the assessed values are respectively 500, 440 and 375 (for example, the number of PhDs over a given period of time). For the third criterion, the assessed values are 5, 10 and 10 (this could be the number of prestigious awards and prizes). The fractions above this table represent the weighting for the criteria: 0.4, 0.4 and 0.2.

The first step consists in standardising the criteria. For the first criterion, the highest assessed value is 2,000. It is brought down to 100, by dividing it by 20. By proceeding in the same way with the other elements of this first column, the result is 68 for b, and 80 for c. For the second criterion, the highest assessed value is 500. It is brought down to 100 by dividing it by 5. Proceeding in the same way with the other elements of this second column, the result is 88 for b, and 75 for c. For the third criterion, the assessed values must be multiplied by 10 to raise the highest value to 100, which yields 50 for a, and 100 for b and c respectively.

	0,4 C1	**0,4 C2**	**0,2 C3**
a	100	100	50
b	68	88	100
c	80	75	100

The weighted averages of the universities are then calculated. For a: $100 \times 0.4 + 100 \times 0.4 + 50 \times 0.2 = 90$. In the same way, we obtain 82.4 for b and 82 for c.

Now assume that the assessed value for a is modified for the first criterion, replacing 2,000 by 1,700. Nothing else changes, and the following table is obtained:

	0,4 C1	0,4 C2	0,2 C3
a	1,700	500	5
b	1,360	440	10
c	1,600	375	10

The standardisation of the first criterion is thus carried out by dividing the assessed value by 17 instead of 20 (b: 80, c: 94). This does not affect the other two criteria, and the standardized table now reads as follows:

	0,4 C1	0,4 C2	0,2 C3
a	100	100	50
b	80	88	100
c	94	75	100

Computation of the weighted averages leads to 90 for a, 87.2 for b and 87.6 for c.

	Before	After
a	90	90
b	82.4	87.2
c	82	87.6

Comparing the scores before and after the changes made to the table reveals that the scores for b and c are now closer to those of a, which is normal since the performance of a for the first criterion decreased. However – and this is more problematic – the order of the scores for b and c has been modified. University b ranked better than c before the modification. This is no longer the case, although the performances of b and c remain unchanged.

This is an example of what could be called a pernicious effect of standardisation as practiced in this approach. Modifying the performance of a university on one criterion can drastically change the ranking of the other universities, all things being equal otherwise.

This type of consideration requires that the conclusions drawn from the Shanghai ranking be put into perspective; yet it is not mentioned in the literature devoted to this particular ranking. It may well be that the authors themselves are not aware of this phenomenon and, generally (since only one example was selected) of the hidden properties or pernicious effects of their methodology.

From these comments, it seems clear that the scientific nature of the approach taken by the authors of the Shanghai ranking ought to be seriously questioned.

3. The *Times* ranking

The *Times Higher Education Supplement* ranking is published by a private firm, a subsidiary of the News International Publishers Limited, a company which publishes the *Times* and *Sunday Times*. The methodology is based in part on a survey

of the renown of universities. Several hundred scientists from different countries are invited to list the universities they consider to be the best in the parts of the world for which they feel they are competent. Recently, this was completed by another survey conducted on a number of employers.

The authors of the ranking extract the universities' scores for the first criterion from the results of this double survey. This will count for 50 % in the final score. However, nothing is known about the actual calculation carried out to translate the findings of the survey into scores for this criterion, nor about how the surveys of the previous years are taken into account.

The other 4 criteria are:

— the impact, in terms of citations, of university researchers (with a weight of 20 %),
— the student/teacher ratio (with a weight of 20 %),
— the percentage of foreign students (with a weight of 5 %),
— the percentage of foreign teachers (with a weight of 5 %).

The first of these criteria, namely the impact in terms of citations, is here again, drawn from the Thomson Scientific databases. Incomplete information as to how this impact is attained does not allow us to reconstruct and verify calculations.

The data for the other 3 criteria are provided by the universities themselves, from an on-line questionnaire to be completed.

Until 2006 the standardisation to 100 of each scale and the aggregation through a weighted mean follow the Shanghai ranking method. But in 2007, they have decided to replace the normalization strategy awarding 100 points to the best performers on each variable by the "z-scores" method. For each criterion (variable), the empirical mean and the empirical standard deviation are computed. Then in order to construct the standardized variable, the value taken by each institution on one variable is subtracted by the associated mean and divided by the associated standard error. The z-score indicates how far the institution deviates from the mean value using as unit the standard deviation.

A number of comments can be made about the *Times* ranking. To begin with, a significant part of the ranking is based on recommendations formulated by "experts". According to the authors of this ranking, the scientific world is familiar with the "peer review" system and their methodology is apparently in line with this system.

This argument, however, seems rather superficial. Here, scientists are not asked to assess a scientific paper or a research project in line with their fields of expertise or even the *curriculum vitae* of a colleague in their field (in which they truly are the experts). They are asked to give an enlightened and clear opinion on the performance of tens or hundreds of universities considered globally and in their own complexity: so their reply, if they choose to reply, might be a vague extrapolation of what they already know or echo a rumour they have heard or a piece of information read in the press.

It is as if an oenologist were asked to assess the quality of a number of restaurants not only for their wine list, which is indeed his area of expertise, but for

their cuisine, the originality of the dishes, the reception, the service, … without ever having set foot in most of them.

The methodology followed to conduct this survey raises a number of questions. How were the "experts" recruited? What is the assessment protocol? What was the response rate to the survey? What is the profile of the respondents? What is the profile of those who did not reply? What is the distribution of the replies? How were conflicting responses dealt with? Can the accuracy of the conclusions of the survey be assessed? What credit can a "free-thinking" reader give to a survey in which none of this information is available?

Anthony Van Raan (2005) of the University of Leiden recently calculated that the correlation between the scientists' replies to the questions of the *Times* and a conventional bibliometric analysis was of the order of 0.005, i.e. equivalent to 0. So, what does the *Times* measure? For many, it provides, at best, information on the competence of the individuals surveyed.

For the second criterion, namely the impact in terms of citations, please refer to the comments made above on the Shanghai ranking, on the difficulty to gather reliable data and on the numerous material errors that occur in this type of exercise.

For the last three criteria, the data are provided by the universities themselves in reply to a questionnaire. The questions asked are nonetheless far from clear and the terms and wording can lead to various interpretations. To such an extent that in 2007, in the French-speaking community, the universities decided to agree on a common interpretation of the questions asked. Of course, there is no way of telling whether other universities around the world have adopted the same interpretation.

Experience has also shown that, even when the data is provided by the universities themselves, errors still occur. For example, the data used by the *Times* in 2006 were entirely wrong for ULB, and this was unfortunately detrimental. The errors, acknowledged by the ranking officials, set this university back 100 places between 2005 and 2006.

According to the authors of the *Times*, a small percentage difference in the overall score may not be significant. However, barely 8 % separate the 100th university from the 200th and 80 % of universities rank within a 20 point interval out of 100 (20 out of 100 was, according to a study published in *Nature*, less than the margin of error accounted for cumulated material input errors and inaccuracies in the data provided).

It is unfortunate that the authors of the ranking, and the journalists who report the results, do not qualify their statements.

4. General comments

Additional general comments can also apply to most rankings currently available in the media or in specialised literature.

First of all, above and beyond issues of methodology or technical aspects, university rankings raise fundamental questions that must be addressed. Five such issues are listed below.

1. How is a university defined? Should we pool "complete" and "incomplete" universities, American or Asian private business schools and European public uni-

versities? What about the French *Grandes Ecoles*, the Max Planck Institute, the laboratories of the CNRS (Centre for Scientific Research)? How can we compare organisations with institutional and cultural traditions that are totally distinct? How do we define exactly what we would like to compare and rank?

2. How can the "quality" of a university be defined? The *Shanghai* and *Times* rankings seem to imply that the quality of a university is an objective reality, clear in everyone's mind, readily measurable (see for this problem Dehon *et al*., 2009). Can one always and confidently assert that a university is of better quality than another? Is this a sensible question? What does "better quality" mean? Can the quality of a university be measured in the same way as the length of a table?

At the top of the list are some universities of a quality no one would think of questioning, e.g. Harvard, Stanford, Yale, MIT, Cambridge, Oxford (universities whose financial means, admittedly, match their ambitions), but this does not prove that their full ranking reflects an objective reality. In fact, there is a peculiar sentence in an editorial by the authors of the *Times* ranking: "Nothing indicates that a university which ranks well in our tables is better than a less well-ranked university". If this is the case, what do the other remarks made by these same authors mean?

3. Is there an absolute model for a good university, a model that all the institutions in the world should attempt to match? Is it desirable that there be only one such good university model?

4. What aspects of a university should be taken into account when assessing its quality? Research? Teaching? Research and teaching? How efficiently it is run? Its ability to obtain external funding? Its contracts? Patents? Spin-offs? Its international dimension? Its participation in regional economic development? Lifelong learning? Its cultural dimension? Its social role? The remuneration of its teaching staff? The cost of studies? The quality of its student accommodation? The wealth of its library? Whether admission is selective or not?

Note that the quality of teaching is hardly represented in the two rankings mentioned here and that nearly all of the other aspects mentioned above are practically non-existent.

For students who want to choose where they are going to study, would it not be more useful to compare, for each subject category, the programmes offered by several universities?

For university authorities, would it not be better to highlight the strong and weak points of their institution by comparing the latter, criterion by criterion, with those institutions having similar objectives?

These remarks now lead us to the fifth fundamental question: what use can be made of these rankings? Very few authors of rankings ever spell out their objectives. However, how can a scientific method with the aim to compare and rank universities be defined, if the objective and the use to be made of the results have not been defined *a priori*?

In my opinion, any attempt at ranking should first answer these fundamental questions. However, aspects related to methodology must also be taken into consideration. Five of these are listed below:

1. Assuming that the aspects to be taken into account have been selected, how can they be measured? With what tools? How can these measures be carried out and the data collected for the thousands of universities around the world, or even for only a few hundred of them, within a reasonable timeframe, especially if the ranking is published on a yearly basis? How can the reliability and accuracy of the data collected in this way be assured? What would a 10, 20 or 30 point difference or a 10, 20 or 50 place difference in the final score mean? Here too, the authors of the rankings remain silent on the statistical meaning of their results.
2. As mentioned earlier, a particular difficulty arises when assessing scientific activities in Arts and humanities. On the one hand, bibliometric analyses, which have become common in some subject categories are, today at least, unsuitable for most branches of Arts and humanities. On the other hand, a survey conducted among scientists stands a greater risk of bias because of their choice, to the extent that there are "schools" of thought with diametrically opposed views on certain types of research. This is perhaps even more true for Arts and humanities than for exact sciences. Assessing scientific activity in Arts and humanities remains a problem that needs revisiting.
3. The rankings we have discussed only assess the output of universities, without ever considering their input or the context in which they have to operate. The available budgets, constraints in terms of student admission (some universities are open to all, others are very selective), constraints in terms of registration fees, remuneration of the teaching staff, human resources, none of this is taken into account.
4. The authors all admit that there is a bias in favour of Anglo-Saxon universities. This cannot be denied, in most subject areas English is naturally the language of communication among researchers. Recent studies have shown that for German- and French-speaking universities, this bias could result in a high percentage of their scientific production being underestimated.
5. Assuming that all the required data were collected and their reliability guaranteed, is it legitimate to compress this mass of information into a mark out of 100 given to every university in the world?

What does this mark out of 100 really mean when it is supposed to integrate as diverse aspects as research, pedagogy, student/teacher ratio, quality of campus life… and at the same time all of the subject categories offered at a university (sciences, medicine, art history, law, economics, psychology…)? Does the concept of an average mark still mean anything in such a context?

Finally, once again it is important to draw the reader's attention to a purely technical aspect, common to most ranking methods, but with far from trivial effect. The performance of universities for the various criteria is aggregated using a weighted average.

First of all, a weighted average deletes all information on the strong and weak points of universities, and thus sets a university scoring well on all criteria and another university having serious weaknesses for some criteria that are compensated for by excellent aspects for other criteria on the same footing.

From this point of view, presenting and comparing university profiles (that is, a vector of their performance) would be far more instructive than an aggregated score

which does not measure any objective reality. It is true that, if a university improves on all of the criteria, its weighted average will also rise, but the converse is not true. A rise in the weighted average for a given university does not mean off hand that this university has improved. It does mean that it is improving in certain aspects, but it could also have regressed in others. The overall quality of a university is too complex a concept for it to be measured accurately by a single number (Dewatripont *et al*., 2002 and 2008).

The authors of the rankings may be unaware of this, but choosing the weighted average as an aggregation technique implies a political choice.

This can be illustrated by a numerical example. Assume three universities a, b and c were assessed on two scales by means of points ranging between 0 and 100 as follows:

	C1	C2
a	41	97
b	100	38
c	68	68

One policy could be to prefer a university rating very high on one criterion, even if it is rated very low on the other. This policy would lead to retaining university a or university b, depending on the most weighted criterion. Another policy could be to prefer a university that is highly rated on the two criteria considered, that is, a university without weak points or shortcomings. In this case, university c could be selected. The technical tool used to designate the "best" university should take both policies into account.

A simple calculation, which has not been reproduced here[2], shows that this is far from being the case: it is mathematically impossible for university c to rank first if the weighted average is the chosen aggregation technique, whatever the weights given to the criteria. Selecting the aggregation technique is therefore an implicit political choice.

5. Conclusions

Most university authorities claim that rankings as reported by the press are disputable. Nonetheless, whether we like it or not, the public-at-large and the academic world consider these rankings as representative of "the" true quality of universities. Opinions can always be discussed, not a mark out of 100 which is "objective" information.

Even if the rankings do not reflect reality today, they will do so tomorrow, because they will have become the reference that universities turn to for their policy; it is a fact that a university which does not integrate ranking criteria in its governance stands a good chance of becoming a second rate university (see Hazelkorn, 2007b).

[2] Equal weights for the two criteria leads to an ex-aequo between a and b (with a score of 69). The score of a is evidently larger than 69 when more weight is given to the second criteria, and it is the inverse for b. But for c the score is constant and equal to 68.

It must nevertheless be borne in mind that too many assessment and ranking procedures can also gradually limit the necessary freedom that is indispensable for creating new knowledge. There is already serious competition between researchers and it is healthy that competition also exists between universities, but one must be wary of excesses.

In some Chinese universities today, researchers receive a bonus for their contribution to the good ranking of their university. In the contract signed by the President of the University of Arizona, a clause stipulates that he will receive $10,000 if he improves the ranking of his university.

What will happen when universities compete to attract Nobel Prize winners or the most prolific writers by spending millions of dollars or euros, as football clubs do for the superstars of the sport?

We must pay close attention to the fact that all universities around the world could choose to adapt their behaviour to the most popular rankings, with the adverse effects this can have. At the next meeting of the Board of Administration, members could move that the least profitable Faculties in terms of ranking all be closed. Such a measure would have an immediate beneficial impact on our university's ranking, in the same way as restructuring often leads to a rise in a company's share prices on the stock market a few days after the announcement of job cuts.

It must be remembered that publishing a ranking modifies the reality that it is supposed to measure. Assessing universities on an annual basis could lead to trends, to gradually aligning all of the institutions on the same standards, harmonising their profiles and thus reducing academic choice. How universities are compared inevitably influences their policies and, in turn, the future of the university landscape.

Assessing the quality of higher education and research does not consist in representing an objective reality that can be measured on a scale from 0 to 100. It is a complex issue, which must be dealt with scientifically, by competent people (in bibliometry, in higher education and research, in analysis and data processing techniques), in other words, by multidisciplinary teams who can accurately analyse the interpretation and utilisation of assessment results.

Assessing the quality of higher education and research can only make sense once a policy for higher education and research has been defined. It is our mission to define this policy.

References

Dehon C., A. Mc Cathie and V. Verardi (2009), "A Robust PCA on University Rankings", working paper.

Dewatripont M., F. Thys-Clément and L. Wilkin (2002), *European Universities: Change and Convergence,* Editions de l'Université de Bruxelles, Brussels.

Dewatripont M., F. Thys-Clément and L. Wilkin (2008), *Higher education in a globalized world: governance, competition and performance,* Editions de l'Université de Bruxelles, Brussels.

Hazelkorn E. (2007a), "Learning to Live with League Tables and Ranking: The Experience of Institutional Leaders", *International Ranking Expert Group*.

Hazelkorn E. (2007b), "L'impact du classement des établissements sur la prise de décision dans l'enseignement supérieur", *Politique et gestion de l'enseignement supérieur*, OCDE, vol. 19(2), pp. 95-122.

HICKS, D. (2004), "The four literatures of social sciences", In H.F. Moed, W. Glänzel and U. Schmoch (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*, Kluwer Academic Publishers, Dordrecht, pp. 473-496.

MOED H.F. (2002), "The impact factors debate: the ISI's uses and limits", *Nature*, 415, pp. 731-732.

MOED H.F. (2005), *Citation Analysis in Research Evaluation*, Springer, Dordrecht.

MOHRMAN K. (2007), "Educational Exchanges: What World-Class Universities Should Not Adopt from US Higher Education", WCU.

SADLAK J. and Liu Nian Cai (Ed.) (2007), *The world-class university and ranking: aiming beyond status*, Unesco-Cepes.

VAN RAAN, A. (2005), "Challenges in ranking of universities", powerpoints of paper presented to the first international conference on world class universities, Jaio Tong University, Shanghai, 16-18 June.

Websites

http://www.isihighlycited.com
http://ed.sjtu.edu.cn/ranking.htm
http://www.thes.co.uk
http://www.efmd.org

Evaluating research in Dutch universities: fifteen years of nationwide peer-review

Roel D. Bennink

Summary

This paper describes the system of external research assessment of the fourteen research-based universities in the Netherlands and how it evolved since the start in 1992. The system has no direct links to funding; it is, however, aimed at improvement and accountability. The main characteristics are presented in this paper and a number of evaluative questions are answered. How does peer review and public accountability contribute to the quality of research? What are the effects, advantages and drawbacks? In a rejoinder to this contribution, one of the peer reviewers participating in one of the Dutch research assessment exercises, reflects on the experience.

Keywords. Evaluation, Netherlands, peer review, research, universities.

1. Universities in the Netherlands

Let us start with a short presentation of the university landscape in the Netherlands. The Netherlands have fourteen research-based public universities (including the Open University of the Netherlands). Their combined budget amounts to about 5 billion Euro, they employ about 40,000 staff and have about 200,000 students.

There are three types of funding for research at universities. Their main source of funding is the "direct funding" from the Ministry of Education (1.4 billion Euro for research or about 60 % of the total research budget). This type of funding is not quality related; it is essentially stable, mainly based on the number of students. The second source of funding comes from the National Science Organisation (NWO) in grants for temporary projects. This type of funding is competitive, based on *ex-ante* assessment of proposals. The third source of research funding is also for temporary projects and comes from the industry, ministries and charity funds.

The second and third types of funding together amount to 950 million Euro (40 %). The universities are autonomous institutions; their accountability to the Ministry of Education is organised through annual reports and through legal requirements regarding quality assurance.

2. The tradition of external reviewing in the Netherlands

All publicly funded research in the Netherlands' universities must be submitted for external review every 6 years. The decision to start this system of external reviews was collectively taken in 1992 by the universities within the framework of the Association of Universities in the Netherlands (VSNU). In the first round (1993-2003), the reviews were organised nationwide per discipline. After a trial in 1993, a schedule of 28 reviews was drawn-up, so that about 5 reviews would be organised each year. The 28 "disciplines" were broad research domains, such as Physics, Chemistry, Socio-Cultural Sciences, Theology, Psychology, etc. International committees of peers performed independent assessments, according to a procedure laid down in a protocol. In 1997, the system was evaluated and a new protocol (VNSU-protocol 1998) was produced for the next round.

The current *Standard Evaluation Protocol 2003-2009 for Public Research Organisations* (SEP) was introduced in 2003, jointly agreed upon by the Association of Universities in the Netherlands (VSNU), the Royal Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO). Again, this protocol was based on an evaluation of the previous round, carried out jointly by VSNU, KNAW and NWO (*Kwaliteit verplicht…*, 2001).

Although the main aims and characteristics of the external reviews have not changed, it is obvious that the three rounds and the two evaluations have led to a number of changes, laid down in the protocols of 1998 and 2003. An important change in 2003 was that the mandatory nationwide reviews were abandoned and the universities were made individually responsible for organising the reviews. This meant that e.g. interdisciplinary reviews could be set up, and reviews organised by an individual university or a small group of universities. Several disciplines continue to organise nationwide reviews, however, because they regard the simultaneous, comparative element as valuable. No change was made in the requirement that all research must be submitted for external review every six years.

3. Characteristics of the current round of research reviews

The external reviews according to the *Standard Evaluation Protocol* (SEP) combine the following internal and external objectives:

— Improving the quality of research;
— Improving the research management and leadership;
— Accountability to government and society.

In view of these objectives, the main aspects to be evaluated are:

— Quality: the international recognition and innovative potential;
— Productivity: the scientific output in relation to the staff input;
— Relevance: the scientific and socio-economic impact;
— Viability: flexibility, management, leadership, future plans.

These four aspects are assessed by the committee and are scored on a five point scale that is comparable to the current scale of the Research Assessment Exercise (RAE) in the UK:

Table 1 Comparison of RAE-scales (UK) and SEP-scores (the Netherlands)

SEP-scores (The Netherlands)		*RAE-scores (United Kingdom)*	
5. Excellent	internationally leading; important and substantial impact	world-leading; a primary reference point of the field or subfield	4*
4. Very good	internationally competitive, national leader; significant contribution	internationally excellent; a major reference point that substantially advances knowledge and understanding of the field or sub-field	3*
3. Good	internationally visible, nationally competitive; valuable contribution	recognised internationally; a reference point that advances knowledge and understanding of the field or sub-field	2*
2. Satisfactory	nationally visible; adds to understanding	recognised nationally; a contribution to knowledge or understanding of the field or sub-field	1*
1. Unsatisfactory	flawed, not worthy of pursuing	below the standard of nationally recognised work	unclassified

The assessment method of the reviews is a combination of self-analysis and peer review. The protocol gives detailed instructions for the information that must be provided in the self-analysis, but the data definitions are in line with what is stored in the research information systems of the universities. This means that the reviews have a strong effect on the quality of the data systems, and vice versa. The quantitative data elements have a basic, multi-purpose character and keep track of the following elements:

— Research staff (tenured, non-tenured, PhD, support) per year;
— Funding (ministry; research councils; contracts) per year;
— Spending (personnel; other) per year;
— Results (publications).

The recurrent external reviews and the mid-term reviews that are also mandatory since 2003, have become important elements in the communication and policy development, on several levels of the universities. The reviews have become closely linked to research management, quality control and accountability to higher levels. This is enhanced by the uniform quality criteria specified in the protocol that all universities use, by the public nature of the reviews and (in the case of the larger reviews) by the simultaneous and comparative aspect.

Apart from the quantitative, factual data (the "metrics") that the self-assessments must contain, there are also a number of more qualitative and descriptive elements prescribed. The institutes and programmes must describe their mission, strategy and research processes (teamwork, supervision, quality control, etc.). They must provide evidence for their research reputation (reviews, awards, citations), and for their socio-economic impact (spin-offs, stakeholder surveys). Finally, they must provide an analysis of their strengths, weaknesses, opportunities and threats (a so-called SWOT-analysis).

Although through the years a shift has taken place towards management issues, the basis of the reviews still lies in the content of the research. The review panels consist of internationally recognised experts in the areas under review. For each research programme two panel members are selected as first and second reviewer. They receive copies of three key publications of "their" programmes and they take the lead in the panel discussions and in the interviews. They also write the assessment texts for the report, including critical remarks and recommendations. Ultimately, the review panel as a whole is responsible for all assessments and for the public report.

4. Self-assessments

The self-assessment documents perform a crucial function in the review process. They are a vehicle for self-reflection and they are the main source of information for the review panels. Gathering the necessary data and compiling the descriptive paragraphs takes a considerable effort from the institutes and research groups under review. Even though the quantitative data are generally stored in research information systems, properly presenting and analysing them is a time-consuming task. This process of compiling and discussing the information that will be presented to the panels, can be regarded as an important element in the "quality culture" of the research units, because it sharpens the shared values and expectations, it can be used to develop or change policy measures, and it can lead to the exchange of best practices[1]. The prospect of external feedback on the results of this process, adds a sense of purpose and urgency.

The different elements of the self-assessments that the protocol prescribes, are based on the notion that a systematic, cyclical monitoring of objectives, results and policies contributes to the quality of research (*Kwaliteit verplicht*…, 2001). Creativity, dynamics, responsibility, openness and professionalism were regarded as crucial for creating the conditions for high quality academic research. The reviews are intended to give feedback to the researchers and to the management on different levels in and between the universities, and to give substance to decisions about material, financial and human resources for research and about the direction of the research itself. This means that the review system must have the flexibility to accommodate many different situations, and yet be stable enough to provide the uniformity that is needed for comparison.

To achieve this, the Standard Evaluation Protocol was based on the so-called EFQM excellence model[2] and defines nine main areas of attention: Leadership (how are improvements stimulated and supported?), Strategy & Policy (long term objectives, coherence), Training & Selection, Resource management, Operational proc-

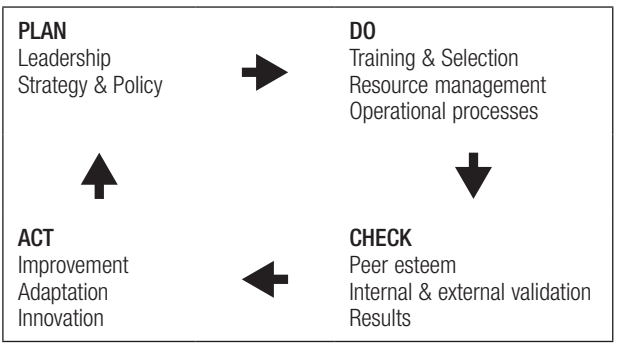
[1] For the notion of quality culture, see *Quality culture*… (2006).

[2] The EFQM model was developed in the private sector and can be regarded as an operationalization of Total Quality Management philosophies. The model consists of nine elements (leadership, policy and strategy, management of people, partnership and resources and processes, key performance results, and people, customer and society results). A basic premise of the model is that organizations with well-developed enablers will have excellent results. Organisations can use the model as a facilitator of change. See www.efqm.org.

esses, Peer esteem, Internal validation (what do we think of ourselves?), External validation (what do others think of us?) and Results of the research.

These nine areas of attention can be placed in the Plan-Do-Check-Act cycle (Figure 1).

Figure 1 Standard Evaluation Protocol as a "plan-do-check-act" cycle



The effort needed to produce a good self-assessment report obviously depends on what is already available. Groups or institutes that do not have a systematic quality assurance, will have much more trouble understanding the terminology and compiling the information than other units, even though the information is basically nothing more than what any research unit should have available at any time anyway, for example for annual reports, for grant applications, for policy purposes, etc.

The self-assessment reports produced by Dutch universities in the current round of reviews are mostly very professional and extensive documents; they often serve as internal and external reference documents and are sometimes also made public on the Internet[3].

5. Review committees

The review committees (or panels) must have the necessary competencies, disciplinary expertise and professional backgrounds to carry out the assessments. They must also be completely independent from the research institutes under review. Candidates are approached by or on behalf of the university board(s), on the basis of proposals from the units under review. In most cases the research expertise and international academic reputation of the candidates are the main criteria, but sometimes panel members are also selected on the grounds of their societal, political or managerial backgrounds. The size of the panels varies with the volume and breadth of the research programmes under review. Usually, the size is six to eight members. In the large interdisciplinary domain of environmental sciences smaller (sub)committees of three members have been used.

[3] See for example the self-assessment for the 2007 review of the Department of Industrial Design Engineering of Delft University of Technology, available on www.io.tudelft.nl.

In the research reviews organised by the independent agency QANU (Quality Assurance Netherlands Universities), the panel members sign a declaration stating that they will judge without influence from the institute, programme or other stakeholders, and without bias, personal preference or personal benefit. Any relationships with units under review must be reported and discussed in the panel.

QANU gives a short introduction to each panel on the use of the protocol and on the general background of the review system, but the quality of the reviews is largely determined by the fact that members of the international academic community have a shared notion of quality and the panel members are experienced in evaluating the work of their colleagues. In other words, the assessments are based on the collective wisdom of the panel members, and the protocol is the instrument that structures the documentation, the review process and the panel reports.

The work of the panels consists of preparation, interviews and reporting. The reviews always include interview sessions with the management and the programme directors, either on-site in the universities or on a central location in Holland. To prepare themselves for the interviews, the panel members read the documentation and the first and second reviewers for each programme make a preliminary assessment for the programmes assigned to them. These are discussed in the first meeting of the panel and they are the basis for the questioning during the interviews. The interviews with the programme directors usually take about 45 minutes and sometimes include short powerpoint presentations about the highlights of the programme. All participants regard these peer-to-peer sessions as indispensable elements in the review process. They add a personal touch and they contribute greatly to mutual trust. In terms of content they provide an opportunity to update and check the information provided in the self-assessments.

6. Reports

The assessments are laid down in public reports. For each research programme, the reports contain scores on the 5-point scale for Quality, Productivity, Relevance and Viability, plus an explanation of these scores in the assessment text. For each research institute, the reports must contain reflections on the leadership, strategy and policy, and assessments of the quality of the resources, facilities, academic reputation and societal relevance.

In case more than one institute is involved in the review, the reports contain a general reflection on the fields and subfields that they cover.

A draft version of the report is submitted to the units under review, for factual corrections and comments. The panels take these comments into account for the finalisation of the report. The final report is submitted to the boards of the participating universities, who are responsible for checking that the report is complete and consistent, and for formally accepting the report as an evaluation according to the national protocol. The university boards ask the institutes under review to react to the report; their reaction can be added to the report as an appendix.

Sensitive issues that are not suitable for the public domain, can be reported in a confidential management letter from the panel to the faculty of university board. Such issues can be of a personal nature (illness, conflict) or a strategic nature (interuniversity cooperation, large scale facilities, feedback on strategic plans).

Table 2 gives an overview of the number of "Excellent" scores for quality in the reviews that were held in the period 1998-2004[4].

Table 2 Absolute number of evaluated programmes and proportion of programmes marked with the "excellence" score

Year	*Discipline*	*Number of programmes*	*Excellent*	*%*	*Year*	*Exc./ nr. of progr.*	*%*
2002	Earth Sciences	26	9	34,6			
2002	Movement Science	6	2	33,3			
2000	Mechanical Engineering	35	10	28,6	2008		
2004	Mathematics	45	12	26,7			
2000	Philosophy	34	9	26,5	2006	8/27	30
1999	Agricultural Science	12	3	25			
2002	Chemistry	158	39	24,7			
1999	Biology	91	22	24,2			
2000	Electrical Engineering	39	8	20,5	2006	6/37	16
2002	Economics	60	12	20	2008		
2001	Socio-Cultural Sciences	34	6	17,6	2008		
1999	Veterinary Science	25	4	16			
2002	Business	20	3	15	2008		
2000	Environmental Sciences	14	2	14,3	2007	9/27	33
1999	Medicine	109	15	13,7			
2001	Civil Engineering	26	3	11,5	2005	5/16	31
1999	Psychology	59	6	10,2	2006	15/39	38
2004	Technology and Management	14	1	7,1			
2001	Pedagogics	35	2	5,7	2007	2/20	10
2004	Computer Science	41	2	4,9			
2002	Law	120	4	3,3			
1998	Arts	116	3	2,6			
2000	Theology	50	1	2			
2000	Maritime Engineering	4	0	0			
2001	Social Geography	18	0	0			
2002	Political Science	27	0	0			
		1218	*178*	*15*			

The table shows that ranking of disciplines on the basis of scores is not a useful exercise. There are such differences in the number of programmes, the nature of the domain, the approach of the panels and the degree of coverage of the review, that any conclusion based on these figures alone would be highly debatable.

[4] Sources: Commissie Dynamisering, 2006, available on www.minocw.nl/documenten/15506a.pdf. The three columns on the right indicate in which year the next review was held or will be held, and add some data from more recent QANU-reports, available on www.qanu.nl.

7. Consequences

The reviews have no direct financial impact; the Ministry of Education keeps at a distance and regards the reviews as a means of accountability. There is general agreement that linking the reviews to the direct funding would undermine their vital function[5]. The universities themselves use the reviews as an essential element in the steering mechanisms for research, on the level of the groups and institutes, on the level of the faculties, on the central university level and in the interaction between these levels. Quality as perceived by the academic community is regarded as a valuable criterion for policy decisions; the universities strengthen their research in high-quality coherently clustered programmes and the external assessments provide important input for the choices involved. Ranking of quality scores or productivity metrics are never the only basis for such policy decisions, because the profile or mission of a faculty will also play a role, as well as the courses that are taught in the bachelor and master programmes and graduate schools.

Because of the small financial margins at the disposal of the faculty management, excellent scores in the research reviews are no guarantee that the faculty will award extra funds, facilities or staff. High scores establish trust and goodwill; excellent groups will have easier access to external funds. The positive effect on their reputation will make them more attractive to researchers and students. Some universities do attach some financial rewards to high scores, for example in the form of a PhD-project. Low scores lead to tough questions and all kinds of policy measures, ranging from budget adjustments to leadership changes. The management at faculty and university level will try to strengthen the low-score groups by linking them to other groups and by supporting changes in the direction of the research and in the personnel.

The fact that the review reports are taken seriously by the university management, places a heavy responsibility on the panels. They are often confronted with elaborate comments on the draft report that they submit to the faculties. The programme directors sometimes seem to believe that any score below Excellent or Very good, will lead to severe consequences on the part of the faculty or the university. The university management, on the other hand, emphasises that not all research can be internationally leading and excellent. There are a number of reasons why it is inevitable that some research scores "only" Good or Satisfactory. Groups need around ten years to establish themselves firmly. Groups with a heavy teaching load can show a fully adequate performance without being internationally leading. Research with strong links to the national or regional professional practice (Law, Architecture, Business, Pedagogics) may have difficulties in reaching high-impact international journals. Finally, the necessary link between teaching and research can in some cases lead to groups that are almost sub-critical in size.

---

[5] This is an explicit conclusion of a survey that was carried out for a committee that was set-up by the Ministry of Education to advise on how to increase the dynamics of university research (Commissie Dynamisering). The survey was carried out by CHEPS (Center for Higher Education Policy Studies). The observations in this paragraph are largely based on that survey, which includes eight case studies and interviews with research directors and university managers. See Jongbloed and van der Meulen (2006).

The effects of the reviews can be summarised as follows:

— Visibility is increased;
— Management dialogues are enhanced;
— Management information improves;
— Publishing in high impact international journals is stimulated;
— Groups are merged, extended, redirected or stopped;
— High marks are an expression of the reputation of the group;
— Low marks lead to critical questions;
— Recommendations are taken seriously;
— Ministry is kept at a distance.

8. Lessons learned

Any system of external quality assessment will have many inherent tensions; the delicate balance between trust and mistrust, use and abuse, cost and yield, top-down and bottom-up, creates a high degree of complexity. Fifteen years of nationwide peer-review in the Netherlands have shown that a flexible system based on trust can work, but also that many checks and balances need to be built in. Constant monitoring of the system and periodic evaluations of its functioning are also necessary. The Dutch experience shows that there is a clear difference between the phase in which internal quality assurance systems need to be built up and the phase in which quality assessment has become a regular, continuous management instrument, supported by a shared quality awareness.

External peer review should never be the only tool for quality assurance. After all, just looking in the mirror or asking the opinion of others on how you look, does not suddenly make you presentable. A general agreement on basic quality indicators and performance data, plus adequate infrastructures for collecting, storing, analysing and exchanging those data, are a necessary foundation for internal and external assessments. Starting up external reviews can be the catalyst to establish or improve that foundation. The next step is then to take the gathered information at heart and take action in order to remediate dysfunctions.

The Dutch system is not unique or isolated from trends in quality assurance around the world. An important condition for the system is that it builds on the quality awareness in the international academic community.

The experience has shown that it was a good choice not to assess individual researchers or programme leaders. The information provided to the panels never includes CVs of individual researchers. The panels assess the performance and the potential of the research groups, not the individual members of the groups.

The experience has also shown that ranking (in the sense of ordering scores or metrics into a top-10 or top-100) is in itself not a useful tool for quality assurance. Rankings can perhaps be used as tools to identify broad differences in performance, but the multi-dimensionality of research quality and the large differences between fields make it necessary to employ more subtle tools for monitoring and steering.

A critical success factor seems to be that the Dutch system stays close to real organisational structures and management processes. The responsibility is placed at the level where it belongs. Feedback is given with reference to the stated mission

and objectives of the programmes, taking their particular context into account, but against the backdrop of current national and international trends. Financial decisions are not placed in the hands of the panels, but remain in the hands of the autonomous institutions.

Another lesson learned from the experience with the Standard Protocol (SEP) is that it is not easy to shift the emphasis of the reviews from content-oriented to management-oriented. Though this shift since 2003 was useful and generally successful, the content-oriented approach remains the main basis for the trust between the panels and the programmes, and the main basis for the assessments and recommendations of the panels. In a research review it is not possible to fully assess all management aspects of a Faculty, Institute or group. On the other hand, the management-oriented aspects in the review provide counterweight to the content-orientation that might otherwise become overly specialized.

In a rejoinder to this contribution, Seamus Hagerty provides the point of view of one of the external evaluators involved in the Dutch research assessment exercise. The external reviewers play a crucial role in the QANU research assessment, so it is worthwile to have a critical reflection on the process from that particular perspective. That is why we invited him to write an extended critical "footnote" to this contribution.

9. Conclusion

The Dutch experience shows that a research review system based on self-assessments, peer-review and public reports can become a valuable element in the quality assurance procedures of universities. Self-reflection and external feedback contribute to the management dialogue and to the proper use of available data. Policy decisions in terms of content, funding or facilities are taken by the responsible management levels and are never automatically based on high or low scores, or on rankings of any kind. A general consensus in the international academic community about basic criteria for the quality of research is at the heart of the system. The protocol that is used nationwide facilitates comparability but allows for enough flexibility to ensure that specific characteristics per subfield or unit can be taken into account. Though the workload is considerable, evaluations of the system have indicated that generally the efforts are considered worthwhile in the end.

References

Jongbloed B. and B. van der Meulen (2006), *De follow-up van onderzoeksvisitaties, Onderzoek in opdracht van de Commissie Dynamisering, Eindrapportage*, March, available on www.minocw.nl/documenten/15506b.pdf.

Kwaliteit verplicht, Naar een nieuw stelsel van kwaliteitszorg voor het wetenschappelijk onderzoek. Report of the Working group Quality Assurance Academic Research (2001), available on www.knaw.nl.

Quality culture in European Universities: a bottom-up approach, report on the three rounds of the quality culture project 2002-2006 (2006), European University Association (EUA) available on www.eua.be.

Standard Evaluation Protocol 2003-2009 for Public Research Organisations, published jointly by the Association of Universities in the Netherlands (VSNU), the Royal Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO), available on www.knaw.nl.

The Dutch research assessment exercise. An evaluator's point of view

Seamus HEGARTY

Summary

The Editors have invited Seamus Hegarty to briefly reflect on a number of issues he has encountered during his experience as an external evaluator in the Dutch research assessment exercise. This contribution is to be read as a rejoinder to the paper written by Roel Bennink, the QANU coordinator. The author focuses on the following issues: (a) is the workload acceptable for evaluators?; (b) does self-evaluation work?; (c) does an external panel have sufficient expertise? and (d) do standardized classification scales make sense?

KEYWORDS. Evaluation, Netherlands, peer review, research, universities.

The assessment of university research is fraught with difficulties as the massification of higher education and the demands for accountability in public expenditure challenge the traditional independence of the academic community. University systems around the world are tackling these problems in various ways and with varying degrees of success. The Dutch system offers a particular, and well defined, approach which has much to commend it, and these brief comments are offered from the perspective of an external panel member.

1. Is the workload for evaluators acceptable?

I served on the Pedagogics and Education Science panel in 2007. We were a total of five, with two members each from Germany and the United Kingdom and one from Belgium, serviced by a QANU officer. One of the five, who had served on a different panel for QANU on a previous occasion, acted as chair. All documentation and discussions were in English. The – voluminous – documentation reached us in good time before the meetings with programme directors.

The key strengths of the process were its efficiency, clarity and transparency. The self-assessment documents, which lie at the heart of the process, while substantial, follow a clear structure and assemble information which good research management would require in any case. The time and effort expended on the proc-

ess are accordingly modest in comparison, say, with the equivalent exercise in the United Kingdom. The demands on review panels, while considerable, are likewise contained: the self-assessment documents and three research articles must be read for each programme; there is a structured 45-minute meeting with the programme leaders; the judgments, in accordance with specific criteria, are drafted by the lead panel member and agreed in discussion (generally electronic) with the full panel; and feedback from the research programmes has to be considered.

2. Does self-evaluation work?

The key question is whether this review process with its crucial reliance on self-assessment and relatively modest use of external review results in judgments that are sufficiently robust to make it worthwhile. To reach a view on this, I propose to look briefly at three issues: self-assessment; panel expertise; and the five-point scale.

Our panel's view of the self-assessment reports was that they provided an informative and generally trustworthy statement of programmes' research context, activity and output. A detailed template has been laid down by QANU and this was followed closely in all cases, so that the reports coming to the panel followed a common structure. We considered these reports carefully and probed their content in discussion with programme representatives. Our view was that they served our purposes well in grounding judgments about the comparative standing of different programmes (and should also serve internal management and accountability).

Where we had reservations, these related to two areas: staff numbers; and publication citations. There was an occasional tendency to inflate programme size by including staff *and their output* whose time commitment to the programme was modest if not negligible. Thus, if a colleague with a 1 % time allocation contributed five research publications to a programme, credulity was strained. A second difficulty related to multiple authorship of publications. While this is common practice (and generally appropriate in view of the team nature of much research), there were a number of instances of publications where most of the authorship was external to the programme citing it. (Our panel recommended that future submissions should indicate clearly which authors were part of a programme and which were not.) While these observations indicate some dissatisfaction at the documentation received by the panel, we regarded them as relatively minor: the weaknesses were evident to us in the detail of the reports; and we were able to allow for them in making our judgments.

3. Can a panel have sufficient expertise?

What of the panel's expertise? We were only five people and, while we could claim some expertise in specific areas of educational and pedagogical research, we could not aspire to depth of scholarship across the entirety of this diverse field. This is where, inevitably, compromises have to be made. Short of assembling a very large team (which would, incidentally, greatly exacerbate the difficulty of securing a common interpretation of the scores on the five-point scale), panels *have* to find a way of making judgments they are prepared to stand over.

I believe our panel's work in this respect was robust. We rated research output in terms of methodological rigour and relevance to the research question, apparent contribution to the literature and place of publication (peer-reviewed journals were deemed to provide their own quality assurance). Each panel member rated programmes independently, with a degree of unanimity that was encouraging. Where we did not agree in our initial ratings, discussion led in all cases to an acceptable convergence of views.

A further source of validation of our judgments came from research programmes' reactions to our scores and narrative texts. Some recipients raised particular issues, which were responded to, but the majority accepted our reports and the judgments embodied in them.

4. Do the classification scales make sense?

The panel's judgments were structured in terms of the five-point scale outlined by Bennink above. Any such scale necessarily entails compromise and even simplification. A numeric scale cannot capture the full complexity of a research programme and implies too a precision to judgment making which is not warranted. These are familiar difficulties and need not be rehearsed further here. There are two particular points, however, that should be noted: the differential use of the scale by different panels; and the situation of practice-oriented areas such as education and pedagogical practice.

Bennink has referred to the first of these, arguing that disciplines should not be compared in terms of their scores. I want to underline the importance of this point but do not discuss it further here. The second point, regarding the particular situation of Pedagogics and Education Science, does need to be developed. While learning and teaching have many universal characteristics, they take place in specific situations which shape them in non-trivial ways. A 12-year-old learner in Amsterdam has a vastly different set of experiences from a 12-year-old in Albuquerque. Research which seeks to understand learning and teaching phenomena must take account of legislative, professional, societal and other factors which are integral to these phenomena. This makes for a particularity in research investigations which renders the application of the five-point scale difficult, especially in relation to the international dimension. Some educational research may have very high quality and great relevance to the local or even national context but not be internationally visible, much less internationally leading. When to this is added the difficulty of securing publication in international (mostly English-language) journals for research papers which are, quite properly, situated in local, Dutch-language contexts, the constraints imposed by the existing five-point scale become more apparent.

It is clear that assessment requires metrics of some kind. It is probable too that any numeric system will constrain judgment in certain respects. It is for consideration, however, whether a single scale, with the same descriptors, can do justice to the full range of scientific inquiry in a modern university. Knowledge generation in medicine and the sciences, say, is different from knowledge generation in education (and social work, law, etc.), and it may well be that effective quality assurance and evaluation demand a more differentiated approach.

5. Concluding thoughts

In summary, the Dutch system for assessing university research has much to commend it. It is efficient, transparent and not unduly laborious. This panel member's view is that it issues in robust judgments and, while there are improvements that could be made, the process is essentially sound.

The CHE approach

Sonja BERGHOFF and Gero FEDERKEIL

Summary

The CHE Ranking, started in 1998, has developed a particular methodology distinct from mainstream ranking as it refers to fields/programmes instead of whole institutions, is multi-dimensional and rejects the over-simplification of calculating a single composite indicator out of weighted indicators, and avoids exaggerating differences in performance inherent to league tables by ordering universities into three groups.

For the CHE University Ranking – mainly intended for prospective students who have to find a university and including indicators relative to teaching and learning, resources and facilities, and research activities – these methodological principles together with the interactive and individualized way of presenting the results on the web version give detailed insights into strengths and weaknesses of departments.

In parallel, the CHE Research Ranking gives a detailed insight into the research performance of German universities. This ranking is based upon indicators relative to the third-party funding, the publications, citations and patents, and the number of doctorates. In 16 specific subjects/disciplines, "strong-in-research" universities are identified and a summary of their research profile is published. These results are completed by information on university level concerning their subject specific research performance, analyses on the composition of third-party funding and on the correlation between different research indicators.

KEYWORDS. University ranking, research evaluation, multi-dimensional ranking.

1. Introduction

In the course of the last two decades, higher education rankings have emerged in many countries the world over. Despite their now long tradition (the first ranking by US News & World Report was published in 1983) rankings are still very controversial, in particular within higher education institutions: "Wherever rankings have appeared, they have been met with a mixture of public enthusiasm and institutional unease" (Usher & Savino, 2006: 3). Rankings were established to create transparency about the higher education system in a competitive world market – for prospec-

tive students, their parents, employers. Rankings are simultaneously the medium and the outcome of competition. They can be conceived as an imperative of the knowledge society. This means they reproduce the competitive structures they are trying to measure. As rankings are constructing – with high public visibility – such hierarchies of higher education institutions in terms of better and worse and as rankings might impact on the market situation of single institutions (e.g. applications, see Clarke, 2007), it is no wonder that they are followed by those institutions very attentively and in a sceptical way.

There is no single concept or model of ranking/league tables. Rankings vary in their aims and target groups as well as in terms of what they measure, how they measure it and how they implicitly define quality (see the comparative analysis of different ranking systems by Dill & Soo, 2005; Usher & Savino, 2006). And last but not least, as universities differ, rankings differ in their quality too. Nevertheless the majority of rankings share some basic methodological features:

1 Most rankings, both national and international, compare whole universities – either exclusively or some also introduce comparisons of broad discipline fields.
2 Most rankings aggregate their indicators into a single composite overall indicator of "the" quality of an institution. The weights given to the single indicators as well as the indicators differ quite a lot between rankings.
3 Results are displayed in a league table with individual rank positions from first to last.

The CHE ranking has a different approach, as explained below.

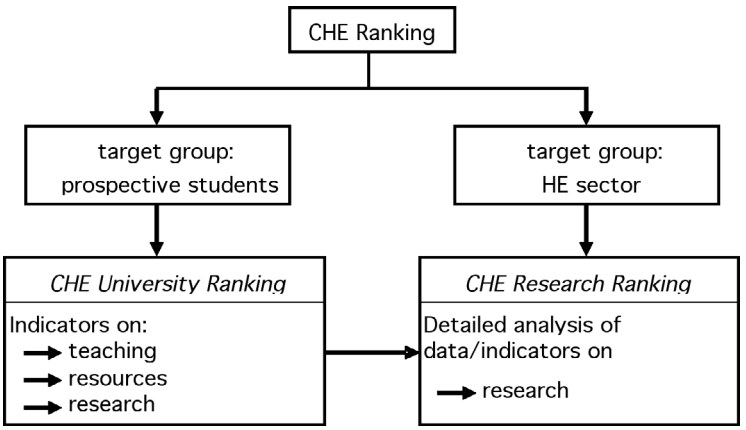
2. The CHE Ranking

The Center for Higher Education Development (CHE) was founded in May 1994 by the German Rectors' Conference and the Bertelsmann Foundation. The purpose of the Center is to initiate and to assist reform in the higher education institutions in Germany. The CHE defines itself as a "think tank" and consulting group for higher education. As a non-profit institution the CHE develops integrated concepts and, through pilot projects, explores possible options for future development. Transparency between German universities by means of ranking was one of the major founding tasks of the CHE.

The CHE started its ranking in 1998 after two years of intensive discussion with evaluation and methodological specialists as well as with students who gave insights on what information they expect from a ranking that is focussing on their need for information. Since 1999, the CHE Ranking is published in co-operation with a media partner in order to gain a wide public attention. But there is a clear division of responsibility: whereas CHE is responsible for the method, the selection of indicators, the collection of data and the calculation of results, the media partner is only responsible for the publication (print and online) and dissemination of results and has no influence on the methods. Since 2005 the CHE Ranking is published in co-operation with the German weekly newspaper *Die Zeit* which has a high reputation within academia; many issues of higher education are discussed in this paper.

The CHE Ranking portfolio includes two different publications with different objectives and different target groups (see Figure 1).

Figure 1 CHE Ranking portfolio



The *CHE University Ranking* focuses on information for prospective students who have to find a university. It includes indicators on teaching and learning, resources and facilities, on research activities as well as information that is important for this target group but is not related to the performance of universities (such as local rents, size of the universities, etc.). Research is included for two reasons: first, for a small group of prospective students, information about research activities and performance is relevant to their decision making about their future university right from the beginning (or, to put it in "Bologna terms", already when they are looking for a Bachelor programme), and, second, a ranking without information on research would probably not be accepted by universities and the higher education sector itself.

As research is not at the centre of a ranking devoted to prospective students, CHE decided to set up a particular *Research Ranking* that gives more detailed insights into research performance for an academic target group. In this ranking, the data on research are analysed and published in more detail. In addition to indicators on publications, citations, number of PhDs, research grants and patents, some bivariate and correlational analysis is included. A detailed description of the CHE Research Ranking is given in Section 5 while Section 4 focuses on the CHE University Ranking for students.

In the context of the Bologna process where there is a growing demand for international comparative information on higher education institutions and programmes, the CHE started to internationalize its University Ranking in 2004. In the early stages, universities from Austria and then (2005) Switzerland[1] were included in the ranking. In 2006/07 the CHE started a pilot project in co-operation with the Center for Higher Education Policy Studies (CHEPS) at Twente University, which

[1] The whole country (not only German-speaking Switzerland).

was funded by the European Commission, to test the possibility of including Dutch universities and *hogescholen* as well as those from the Flemish part of Belgium. There was a decision right at the beginning that the results would not be published in the ranking in this first round. In 2009, most Dutch universities will participate in the ranking.

3. Methodological principles

Both CHE rankings – the University Ranking and the Research Ranking – share the same philosophy and basic methodological principles that are distinct from the mainstream ranking outlined in the introduction. The CHE Ranking is characterized by three basic principles (for a detailed description of the methodology, see Berghoff *et al*., 2008b).

A. Level of ranking: programme/field instead of whole institutions

Most rankings compare whole institutions (e.g. US News & World Report, THES World Rankings, Jiao Tong Ranking). This model implies that institutional-level comparisons are adequate for comparative assessment of universities implying that the institutions as a whole are responsible for quality and good performance. Evidence from the CHE ranking shows that universities can be very heterogeneous with regard to the performance of their individual departments. A university might perform well and hence be ranked high in physics and, at the same time, perform poorly and be ranked low in history. Academics usually have a strong commitment to the academic community in their own field – reputation is mainly attributed by peers within a specific academic field. In a pilot study the CHE sought to establish a field-specific ranking of European top universities in mathematics and natural sciences (physics, chemistry and biology) making a pre-selection of the top institutions per field by bibliometric analysis as the basis for a broader ranking including additional indicators and perspectives. One of the study's most interesting findings is that only a very few universities were among the top universities in all four fields and the majority of institutions were pre-selected only in one or two of those disciplines. Hence an institutional ranking that compares whole universities inevitably levels out such differences in performance within universities. These differences result in many cases from explicit strategic decisions by universities concerning their priorities and the development of specific strong fields.

Another point against institutional rankings is directed at their use by prospective students. The ranking are intended to give information and orientation to this specific target group. This proposal is poorly suited to the European situation. General bachelor degrees are not important in influencing future academic and professional career paths, but the quality of an institution's specific subject matter is. Prospective students, therefore, are much more interested in information about subject/programmes within a university than its overall ranking. The information that a particular university is well ranked can be useless if the department to which the student would like to go is not as well ranked.

B. Multi-dimensional ranking instead of composite overall score

The number of indicators differs between rankings, but independently from that number, most rankings calculate an aggregated overall score by giving particular weights to the indicators. By selecting a particular set of indicators and assigning specific weights to each indicator, rankings impose a specific definition of quality. According to the US National Opinion Research Center, neither a theoretical nor an empirical basis is used in developing such weighting procedures. Also, the heterogeneity of decision preferences in the target group of students or even for other stakeholders can lead to the avoidance of a specific choice of weighting scheme. Some students are looking for a university with high research activities (as measured e.g. by research grants, publications, etc.) while others may look for a university with close contacts between students and teachers, good mentoring and short study duration. Calculating an overall score is thus too restrictive.

Furthermore, institutional-level scoring levels out differences between particular aspects of a programme or of a university's performance. This is most evident in rankings including indicators both on teaching and on research. A university with good research performance does not necessarily provide good teaching and learning experiences for their students and vice versa (although this is a belief held by some academics in Europe – the traditional Humboldtian ideal of the university). Multi-dimensional rankings can provide a differentiated insight into the strengths and weaknesses of a university. This is the only way to take into account the multi-perspectivity nature of quality. This view leads Usher & Savino (2007: 23) from their analysis of ranking systems to conclude that "one of the main reasons of institutional unease [with rankings] is the tendency of institutional ranking schemes to use weighted aggregates of indicators to arrive at a single, all-encompassing quality score".

C. Groups instead of league tables

In the tradition of the US News & World Report rankings, universities are usually arranged in league tables with individual rank positions. This approach suggests that each difference in the numeric value of an indicator marks a difference in quality/performance between the entities ranked. League table comparison inevitably involves the danger of misinterpreting small differences in the numeric value of an indicator in terms of differences in performance or in quality. For example, in the 2001 edition of the US News & World Report ranking of national universities, the difference between rank 13 and rank 22 was only 6 on a 100 point scale. In many cases, data are insufficiently precise to establish clear cut and unambiguous table positions in a reliable way. Or, to put it in statistical terms, such a procedure ignores the existence of standard errors in data.

Hence, for each indicator, the CHE ranking classifies universities into only three groups: a top, a middle and a bottom group. The procedure followed to determine these groups is explained in Section 4.C. There is no additional distinction made within groups; in all publications, universities are ordered alphabetically within groups – so there is no league table.

4. The CHE University Ranking

A Indicators

The choice of indicators is crucial to rankings. Rankings can be distinguished according to the data sources to which they refer and to the quality (relevance, validity) of indicators. Indicators should be relevant to the target group(s). In a preparatory phase of almost two years, the CHE tried to identify relevant indicators with the help of an advisory board (including evaluation experts and members of professional and university associations) and by group discussions with school leavers and students. Those discussions are repeated regularly in order to adjust indicators to changing demands for information within the target group.

Out of this process a "model for decision making" was derived containing nine *components* relevant to the decision process (see Table 1). Each component comprises several indicators – all in all some 35 (depending on subjects/fields). The components range from general information on towns (e.g. mean rents) and the university (size, year of foundation, type), student characteristics, central issues of courses & teaching, some aspects of employability, research and labour market to some overall judgements made by professors and students. Depending on the field, the ranking covers 20 to 25 indicators. A more detailed description of these indicators may be found on http://ranking.zeit.de/che9/CHE_en?module=Baustein.

Table 1 The nine components of the "model for decision making" in the CHE University Ranking

City, university	Students	Study outcome
Internationalisation	Teaching	Ressources
Research	Labour market, employability	Overall assessment (students, professors)

The CHE-ranking follows a multi-perspective approach. First, each component comprises indicators from different data sources. Taking "research" as an example, some indicators are constructed on the basis of data delivered by the faculties (e.g. research grants, number of PhDs), others are derived from bibliometric analyses on the basis of various data bases (e.g. *Science Citation Index* and *Social Science Citation Index*, but also some specific German data bases for specific fields). The CHE also uses indicators based on the professor reputational survey (e.g. research reputation).

Second, the set of indicators comprises objective empirical data as well as subjective judgements. In the component "teaching", for example, there are fact indicators such as student-staff ratios or average study duration (which varies tremendously between German universities, in some diploma-courses up to 3 years!) as well as judgements provided by professors and students, e.g. on course organisation, contact between professors and students, libraries, computer facilities, etc.

B. Data sources

With regard to its multi-dimensional approach, the CHE University Ranking is based on a multitude of data sources which give a multi-perspective view on higher education institutions. The ranking tries to combine facts as well as subjective judgements and evaluations on programmes and institutions.

First, at the core of the ranking, there is a survey at the faculty or department level, collecting data on staff, facilities, students, research and individual degree programmes. Second, a student survey gives detailed insights into the students perspective on their programmes and their universities. The survey includes 500 students per field and institution. Students give detailed feedback on various issues such as e.g. organisation of programmes, teaching and learning, facilities, contacts with teachers and other students. Furthermore, there is a complete survey among the professors of the fields included, in which they give information about the reputation of institutions in their field. The CHE is conducting bibliometric analyses to evaluate the activities of publications and to measure the frequency of citations; in the relevant fields the number of patents are analyzed too. Recently the CHE started to conduct surveys among graduates/alumni in order to get more information on issues of employability and the labour market.

C. Presentation of results

The third methodological principle of the CHE Ranking states to divide, for each indicator, the set of universities in only three groups: a top, a middle and a bottom group. The procedure used to compute these groups differs according to the nature of the indicator.

For a factual indicator, groups are computed by using the quartiles of the measured values. The top and bottom classes contain the universities for which the indicator takes a value greater than the third quartile or smaller than the first quartile, respectively; the middle class is composed by those institutions for which the indicator's value belongs to the interquartile interval.

If the indicator turns on a subjective evaluation and results from a survey of students or professors, the mean of the judgements[2] given by the respondents is determined for the whole set of universities as well as for each individual institution; moreover, in order to take into account not only the mean score but also the number of respondents and the heterogeneity of judgements within each individual university, a confidence interval is computed for its average judgement. Then, a university is placed in the top group or in the bottom group if the confidence interval for its average judgement is, respectively, completely above or below the observed global mean for all the universities; a university is classified in the middle group if the confidence interval for its average judgement contains the observed global mean.

Results of the CHE University Ranking are published in a threefold manner, designed to serve the different users. First, there is series of articles in the weekly edition of *Die Zeit* in which selected results are presented together with background

[2] The subjective judgements are measured on a six points Likert-scale where 1 means "very good" and 6 indicates "very bad".

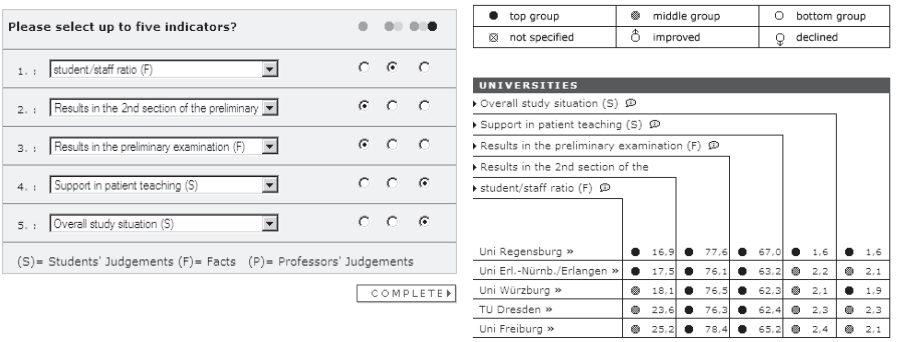
information and additional analysis. This targets a broader public as well as the academic sector. Second, there is a special magazine ("study guide"), which is made particularly for (prospective) students. Again, there is a presentation of basic, selected results of the ranking plus information on fields/subjects and additional information on universities and student life for "freshmen". Third, the complete results are presented in a web-version of the ranking (www.das-ranking.de) both in German and English.

There are several ways to enter the data. First, there is a basic overview table listing all institutions offering degree programmes in a field (in alphabetic order, as there is no overall score) and displaying the rank group for five selected indicators (which are marked by traffic light colours in CHE rankings where green stands for top group, yellow for middle group and red means bottom group). These lists are published in print as well as in a web version, which is the third medium for publishing the results. The internet offers a wide range of interactive ways for dealing with the results. In the web version the overview lists can also be sorted by indicators, clicking on the name of an institution in the list leads to more detailed information about that department and its programmes listing all indicators plus a range of descriptive information.

The most important feature of the web version, however, is an interactive ranking (called "my ranking") in which – according to the basic approach of a multidimension ranking that does not give general weights to indicators – the user can select up to five indicators and decide which groups (only top, only top and medium, all groups) will be displayed and hence gets an individual ranking according to his own preferences and priorities. As the lists normally differ substantially depending on the selection of indicators (e.g. more focussing on teaching and learning versus research) this instrument can identify specific profiles and strengths and weaknesses of the institutions.

The following example, drawn from medicine, shows a personalized ranking with a selection of indicators focussing on study outcomes (results in national exams), students' judgements on their programme (support by teachers in patient teaching and the overall study situation) and student-staff ratio. The selection was made in such a way that only those universities that are in the top group with regard to the results in the first (after 3 years) and the second examination (after 5 years) and those who are at least in the middle group with regard to student-staff ratio, are displayed. Only five universities out of 37 fulfil those criteria.

Figure 2 Personalized ranking



D. Impact of the CHE University Ranking

Rankings may affect students as well as universities themselves. The effects of the ranking on students, that the CHE has been able to measure in a separate study, are quite considerable. According to survey data, about one third of students use rankings for orientation, which is substantial in the German context, given the persistence of the myth that all universities are equal and the fact that for a long time rankings were not accepted in the scientific community. The CHE University Ranking assists in helping make people aware of differences that exist in the quality of teaching and research. The proportion of students using the ranking varies across different subjects: from about 50 % in engineering to only 19 % in literature. Generally it can be said that particularly achievement-oriented students make use of the ranking.

A good example of the impact of these rankings can be shown for psychology, which was first included in 2001 in the CHE Ranking. In the following year, the number of applications at the recommended universities increased notably while the overall aggregate numbers remained stable. The increase was approximately 19 % for universities that had been recommended for the students of "researcher"-type and about 13 % for those who just want to study rapidly and efficiently with adequate monitoring. The CHE investigation demonstrated that good ranking results had more effects on applications than bad results.

At the institutional level, it has been observed that universities and departments take the ranking as a starting-point for analysis of their strengths and weaknesses. In this context the CHE offers detailed analysis of the student survey for single departments that goes beyond the published indicators. After a first phase in which poorly ranked departments often expressed fundamental criticism of the ranking, the CHE now gets considerable positive feedback even by those departments who came off badly (or at least by some professors or vice-deans who are engaged in matters of teaching) telling that they want to make use of the results for an analysis of problems and for reforms.

E. Perspectives of the CHE University Ranking

In the context of the Bologna-process, student mobility within Europe is growing and will probably grow further within the coming years. Accordingly, information for students about programmes in an international perspective will become more important. With this perspective, the CHE is striving for a European ranking.

The internationalization strategy is determined by two goals. First, the ranking should achieve high acceptance within the higher education system and within individual universities of the respective countries. Second, the comparative ranking must – in its methodology and the choice of indicators – take into account specific characteristics of the higher education systems and academic culture of other countries, otherwise the comparison will not be able to produce valid information about those countries. In particular, we have to check carefully the availability of adequate databases for comparative bibliometric analysis in order to avoid biases disadvantaging a specific country.

By this approach, the CHE ranking differs from "world rankings" that put together and analyze commonly available data on different countries without regard to differences in the structure of higher education and academic cultures. In the long run, the aim is to build a European ranking of universities. The task then will be to define clusters or groups of universities which can be compared to each other. A classification of European universities would be a good tool for this endeavour.

In 2007, the CHE published an "Excellence Ranking" in mathematics, physics, chemistry, and biology. It follows the basic CHE approach (field specific, multidimensional, rank groups) and is a ranking for one particular type of university within Europe: top research universities. After a pre-selection of universities mainly based on bibliometric and on internationalisation/europeanisation indicators, results on research indicators as well as those derived from a Master- and PhD-student survey were shown for a group of top universities for each of the fields included (results can be found under: www.excellenceranking.org).

5. The CHE Research Ranking

A. Methodological principles

The CHE Research Ranking is based on the same data as the CHE University Ranking. As explained in the previous sections, the University Ranking presents a lot of information for forthcoming students, such as study duration or students' evaluation of different aspects of the study situation. The Research Ranking concentrates on presenting and giving detailed information on the research performances of the German universities.

Currently, the CHE Research Ranking includes 16 subjects from sciences, social sciences and the humanities. Its aim is to present, for each subject, a top group of universities which are strong in research. Besides the absolute values of various indicators, e.g. the number of publications, the amount of third stream funding or the number of doctorates, some "per capita" indices are also taken into account for the ranking. The results are aggregated at different levels: besides the lists for each

indicator, tables of the strongest universities are presented for each subject as well as an overall list containing the strongest disciplines for all universities.

The CHE Research Ranking adheres to the same methodological principles as the CHE University Ranking:

— no aggregation of indicators across a whole university, but subject-specific analysis of the data and presentation of the results; for each subject a different set of indicators is used and the data is derived from different sources depending on the subject;
— no weighted or non-weighted total value for the research performance of a department, but examination of different indicators in a multidimensional ranking;
— no individual rank positions, but for each indicator a group of the best performing universities is identified.

B. Disciplines examined in the CHE Research Ranking

The subjects examined at present in the CHE Research Ranking are listed below. The year of first publication is indicated in brackets: english studies (2007), biology (2006), business studies (2005), chemistry (2006), electrical engineering (2007), pedagogy/education science (2007), history (2007), mechanical engineering (2007), mathematics (2006), medicine (2006), pharmacy (2006), physics (2006), psychology (2007), sociology (2005), economics (2005), and dentistry (2006).

As already mentioned, the used data sources and the constructed set of indicators vary from one subject to another.

Note that some subjects are still totally missing in the Research Ranking, for example information sciences, German studies[3], political science and law. The reason is that an adequate publication analysis in these disciplines has not yet been established; it seems unreasonable to present a research ranking for these subjects without results of a publication analysis.

C. Data sources

Data for CHE Rankings are raised from different sources. The most important survey is the institutional survey. An online-questionnaire is used to collect data directly at the department level or from institutes. Departments have the opportunity to use the questionnaire via a password, and some questions may be sent to other sections of the university, e.g. the data on third-party funding may be filled in by the central unit in charge of the funding. The questionnaire is only open for about two months, after which it is closed for reasons of data control. In the last phase of this survey, it is reopened and departments/institutes as well as university authorities have the opportunity to react on comments made by the CHE directly in the questionnaire and to correct and complete data where necessary. This last step regarding completeness and reliability of the data is very important. The CHE Research Ranking makes use of the number of doctorates, the amount and composition of third-party funding and of the data concerning personnel.

[3] German literature and language.

Before the actual institutional survey a pre-survey asks for example for the list of researchers' names necessary for the publication analysis. These lists are used for queries in publication databases specifically chosen for each subject. Some of these analyses are performed by the CHE; others are outsourced and carried out by specialized agencies.

Further information used in the CHE Research Ranking is collected through surveys of universities and professors, both of which are explained directly in relation to the respective indicators in the following section.

D. Indicators

The CHE Research Ranking contains and shows details and subject-specific information on different research indicators relative to the amount of third-party funding, the number of publications and citations, the number of PhDs and – if appropriate – the number of patents.

Third-party funding

The *absolute* indicator on third-party funding displays the three-year-average money spent. Data is raised as part of the institutional query within the CHE University Ranking. This survey collects data directly from the departments and institutes of the universities concerned. Funding is divided into different subgroups depending on the source, e.g. the German Research Council or private foundations.

Data quality is assured by different methods:

— Detailed categories do not leave much room for wrong allocation of third-party funding, numbers given for the category "other" must be justified.
— The Research Ranking publishes data only for those departments or institutes with complete data for all three years to avoid cases where only data for "good" years are submitted.
— Data is tested for plausibility and outliers.
— Departmental data is contrasted with external sources, e.g. The German Research Council or the statistics of federal states, as much as possible to show the reliability of the numbers.
— In the process of data collection, the CHE sends all data delivered by universities and departments back to them before computing the indicators. Hence universities have the possibility to complete the data and correct errors.
— An advisory board for the respective subjects checks the plausibility of the results. In addition, the accumulated data is checked by the CHE to identify extreme cases and inconsistencies. Parts of the data collected may not be published because only reliable and valid data should be published in the CHE Rankings.

For computing a *per capita* (or *relative*) indicator, the sum of third-party funding is set in relation to the total number of researchers in a department.

Publications

For different disciplines different approaches are necessary depending on the publication habits in the specific subject field.

The number of publications is counted for an interval of three years. Not all publications of each department or institute are taken into consideration but only a certain subset of "relevant" publications which are selected by the databases used, by authors' names and of course by the time window used[4]. These subsets should represent the publication activity of each unit and are used to set up the ranking.

Sources used for the different disciplines are shown in Table 2. If the Web of Science is used, a citation analysis is also conducted and its results published. If very heterogeneous databases containing everything from thick books to very short articles are used (as, for example, the database "wisonet" for business/economics), a weighting scheme is applied taking into account the number of pages among other things. In very few cases, a set of core journals has been established to differentiate important publications and allow them more weight.

The databases show interesting details on publication behaviour in different subjects. For example, looking at history, one can see that more than 80 % of the publications listed are single-author publications, nearly 4 % of the publications count more than 500 pages, and less than 14 % are shorter than 10 pages. Education science shows a different picture: less than 60 % of the publications are written by single authors, more than 35 % of the listed publications count less than 10 pages and the number of books with more than 500 pages is very low. These facts have to be taken into account when choosing adequate weights for computing the publication indicators.

Table 2 Sources for the different disciplines

Subject	*Database*	*Types*		*Adjustment number of authors*	*Adjustment length*	*Core journals*	*Citations*
		Articles	*Monographs*				
Business/ Economics	wisonet	I	I	I	I		
	Web of Science	I					I
Electrical engineering	INSPEC, Web of Science	I	I				
English studies	AREAS	I	I	I	I		
History	Historische Bibliographie AHF	I	I	I	I		
Mathematics	MathSciNet	I					
Pedagogy	FIS Bildung	I	I	I	I	I	
Psychology	Web of Science, PSYINDEX	I					I
Sciences	Web of Science	I					I
Sociology	Solis	I	I	I	I		
	Web of Science	I					I

[4] For the sciences, for example, bibliometric analyses take only *international* journal articles into consideration; publications in regional journals are not counted. In this sense, the indicator is only based on a subset of publications – those which may be considered as the most relevant to represent the publication activity of a department/institute.

Queries in the publication data sources are based on the lists of names of professors and senior researchers. This technique needs a lot of work but, compared to the institutional approach, has some advantages which induced the CHE to use it. The institutional approach counts all publication produced by an institution in a certain field. In the Web of Science these fields are defined by sets of journals, which means that publications by a physicist in a biology journal will probably not be counted. On the other hand, if one wants to compute *per capita* indices for the institutional approach, one must estimate the number of possible authors which might bring in a new source of errors. The query by name takes into account all publications of the persons on the list if they are listed in the database. Furthermore, as the number of authors is known, computations of *per capita* (or *relative*) indices are sharp and numerator and denominator match. In some of the discipline-specific databases used, institution names are often missing; in this case a query by name is the only way out. Another advantage of the query by name is that the publications of newly-appointed professors or researchers may be counted for their new department. To do so means not only to look back on an institution's achievements during the last year but to try to predict its performance for coming years based on its personnel.

Patents

To represent application-oriented research for engineering and natural sciences the number of patents is counted for several subjects. Since 2006, in Germany, all inventions made by university researchers have been owned first by the university and not by the inventor. Any researcher who wants to have an invention patented by the German or the European patent office has to inform the university first and only if it refuses to get the invention patented can the researcher do so on his own. This regulation makes it possible to ask the university offices concerned with the transfer of knowledge and technology directly for the number of notified inventions.

This was done for biology, chemistry, medicine and physics as well as mechanical and electrical engineering, yielding all inventions reported to the university by researchers of the respective fields in the years 2002-2004 and 2003-2005, respectively. The number of researchers was collected in the institutional survey directly at the departments. This made it possible to show the *absolute* numbers of inventions per year alongside the *relative* number of inventions per ten researchers at the same time.

Doctorates

The number of doctorates is asked for at the department level in the institutional survey; numbers are collected for a time interval of three years. Published in the Research Ranking are the mean number of doctorates per year (*absolute* indicator) and the number of doctorates per professor (*relative* indicator).

Reputation

The reputation of departments or institutes, in respective disciplines, is included in the survey of professors. Professors are asked to name up to five departments they consider to be leading in research in their area throughout Germany. Departments

receiving recommendations by at least 25 % of the professors who answered are sorted into the top group concerning this indicator. This indicator is NOT used as a selection criterion for the group of "strong-in-research" departments but shown as information to see whether standing and performance correspond or whether they do not.

E. Presentation of results

Grouping

For each subject/discipline and each indicator, the departments or institutes included in the CHE Research Ranking are divided in three groups: a top, a middle and a bottom group. But the procedure for computing these groups is slightly different for *absolute* and *relative* (*per capita*) indicators.

For a *relative* indicator, the grouping follows the same principle as the one proposed in the CHE University Ranking for factual indicators. The departments are classified into three groups by using the quartiles of the distribution of the values observed for the indicator: the departments for which the indicator has a value smaller than the first quartile belong to the bottom group and the ones which have obtained a score greater than the third quartile are in the top group. The middle group contains the departments for which the value of the indicator falls in the interquartile interval.

If an *absolute* indicator is considered, cumulated distribution of its values is taken into account. The observed values are sorted decreasingly and their shares in the total sum are accumulated. Departments at the top of the list, which together cover at least 50 % of the total amount of values collected, form the top group for this indicator. Departments at the end of the list covering at most 10 % of the total form the bottom group.

"Strong-in-research" departments per discipline

The data on research are displayed in setting up the group of departments with "excellence" in research when considering both absolute and relative indicators. As explained above, a top group is determined for each indicator; departments/institutes which belong to the top group for at least 50 % of the indicators are classified as "strong-in-research".

These departments are presented in a table which reveals their top group placements on the different indicators and thus shows a kind of very short research profile of the respective departments rather than just a single number. For example, Table 3 shows the "strong-in-research" group in physics in 2006. The first two columns show the university's name and the number of top placements the physics department of this university received. In physics, there is no university that manages to achieve the maximum number of seven top group placements. In other disciplines, it may happen that some universities reach the optimal value but in most cases there are none.

The third column shows whether the respective department was classified as strong in the last cycle (++) or whether it is new in the group of strong departments (+). In physics, four universities are new and – this can be seen at the bottom of the table – five universities fell out of the group. Their performance in the present cycle is also shown in the table.

The next seven columns within the frame show the top group placements; a line means that the department achieved a place in the top group of the respective indicator. Other assignments are not shown here.

The last column reveals the results from the survey of professors. Lines mark those departments that have the highest standing in the opinion of their colleagues.

Table 3 Strong departments in physics

University	*Number of top group placements (maximum 7)*	*Comparison to last cycle (2003)*	*Absolute*			*Relative*				*Reputation*
			third-party funding	*publications*	*doctorates*	*third-party funding*	*publications*	*citations*	*doctorates*	
RWTH Aachen	4	++								
Uni Bochum	5	+								
Uni Bremen	4	+								
TU Dresden	4	+								
Uni Göttingen	4	++								
Uni Hamburg	5	++								
Uni Heidelberg	6	++								
Uni Karlsruhe	6	++								
Uni Mainz	4	++								
LMU München	6	++								
TU München	5	++								
Uni Stuttgart	4	+								
Uni Würzburg	5	++								
No longer in the group of strong departments										
FU Berlin	1	+								
HU Berlin	3	+								
TU Berlin	3	+								
Uni Bonn	2	+								
Uni Freiburg	1	+								

Results on university level

An overview of the results is given in a comprehensive table containing all universities in the research ranking and their respective subjects. The table below lists those universities which succeeded in placing at least 50 % of their subjects under review in the respective "strong-in-research" groups.

Table 4 Overview of the "strong-in-research" groups at university level

University	*Number of subjects in CHE Research Ranking*	*Thereof in top groups*	*Percentage*	*Subjects in CHE Research Ranking (bold: top group)*
TU München	8	7	87.5 %	**Biologie, BWL, Chemie, Elektro- und Informationstechnik, Mathematik, Maschinenbau/ Verfahrenstechnik,** Medizin, **Physik**
Uni Heidelberg	13	9	69.2 %	**Anglistik/Amerikanistik, Biologie, Chemie,** Erziehungswissenschaft, Geschichte, **Mathematik, Medizin, Pharmazie, Physik, Psychologie,** Soziologie/Sozialwissenschaft, VWL, **Zahnmedizin**
Uni Karlsruhe	6	4	66.7 %	Biologie, **Chemie, Elektro- und Informationstechnik,** Mathematik, **Maschinenbau/Verfahrenstechnik, Physik**
Uni Freiburg	13	8	61.5 %	**Anglistik/Amerikanistik, Biologie,** Chemie, Erziehungswissenschaft, **Geschichte,** Mathematik, **Medizin, Pharmazie,** Physik, Psychologie, **Soziologie/Sozialwissenschaft, VWL, Zahnmedizin**
Uni Stuttgart	10	6	60.0 %	Anglistik/Amerikanistik, **BWL, Chemie, Elektro- und Informationstechnik,** Erziehungswissenschaft, Geschichte, **Mathematik, Maschinenbau/Verfahrenstechnik, Physik,** Soziologie/Sozialwissenschaft
LMU München	14	8	57.1 %	**Anglistik/Amerikanistik, Biologie, BWL, Chemie, Erziehungswissenschaft, Geschichte,** Mathematik, **Medizin,** Pharmazie, **Physik,** Psychologie, Soziologie/Sozialwissenschaft, VWL, Zahnmedizin
Uni Göttingen	13	7	53.8 %	Anglistik/Amerikanistik, **Biologie,** BWL, **Chemie,** Erziehungswissenschaft, **Geschichte, Mathematik,** Medizin, **Physik,** Psychologie, **Soziologie/ Sozialwissenschaft,** VWL, **Zahnmedizin**
Uni Frankfurt a.M.	14	7	50.0 %	Anglistik/Amerikanistik, Biologie, BWL, Chemie, Erziehungswissenschaft, Geschichte, Mathematik, Medizin, Pharmazie, Physik, Psychologie, Soziologie/Sozialwissenschaft, VWL, Zahnmedizin

38 universities place at least one of their departments in the respective "strong-in-research" groups. 17 universities do not succeed in any of the subjects, though eight of them were examined in ten or more disciplines.

F. Some further analyses

The presentation of the profile of the "strong-in-research" departments per discipline and of the ranking results on university level is completed by some very informative analyses. Three examples of such supplementary analyses are briefly described below.

Correlation analyses

Contrasting different indicators in scatterplots gives detailed insight into their relation. Looking at medicine, for example, shows that publication output and third-party funding are strongly correlated. The picture is dominated by the Charité in Berlin followed by the Ludwig Maximilian University in Munich (see Figure 3a). But the strong correlation also holds for the universities with less output as can be seen by enlarging the lower left corner of the scatterplot shown in Figure 3a (see Figure 3b).

Figure 3a Scatterplot between publication output and third-party funding

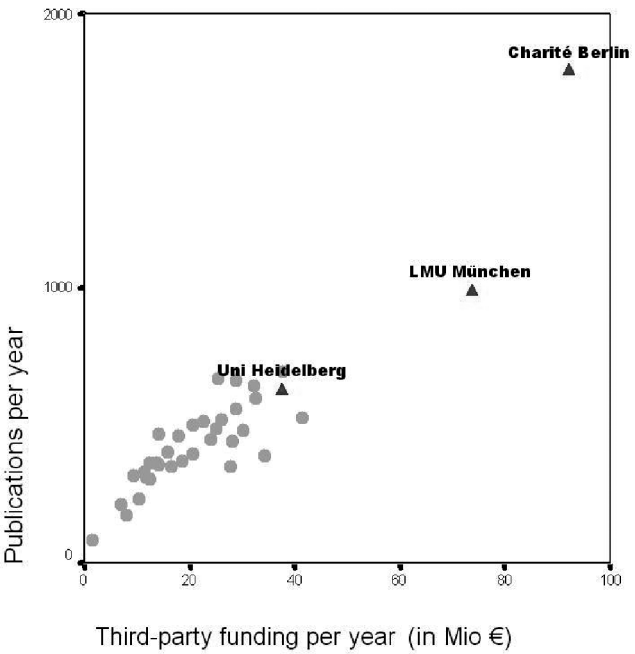
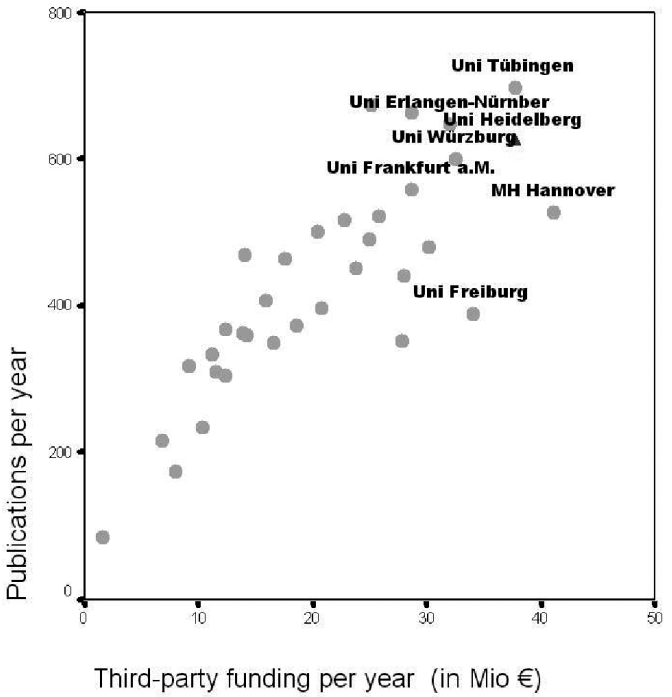


Figure 3b Scatterplot between publication output and third-party funding for universities with less publication output



Let us mention another example. The correlation analysis between the relative and absolute numbers of doctorates, publications or patents allows to illustrate the influence of size on research performance for each discipline.

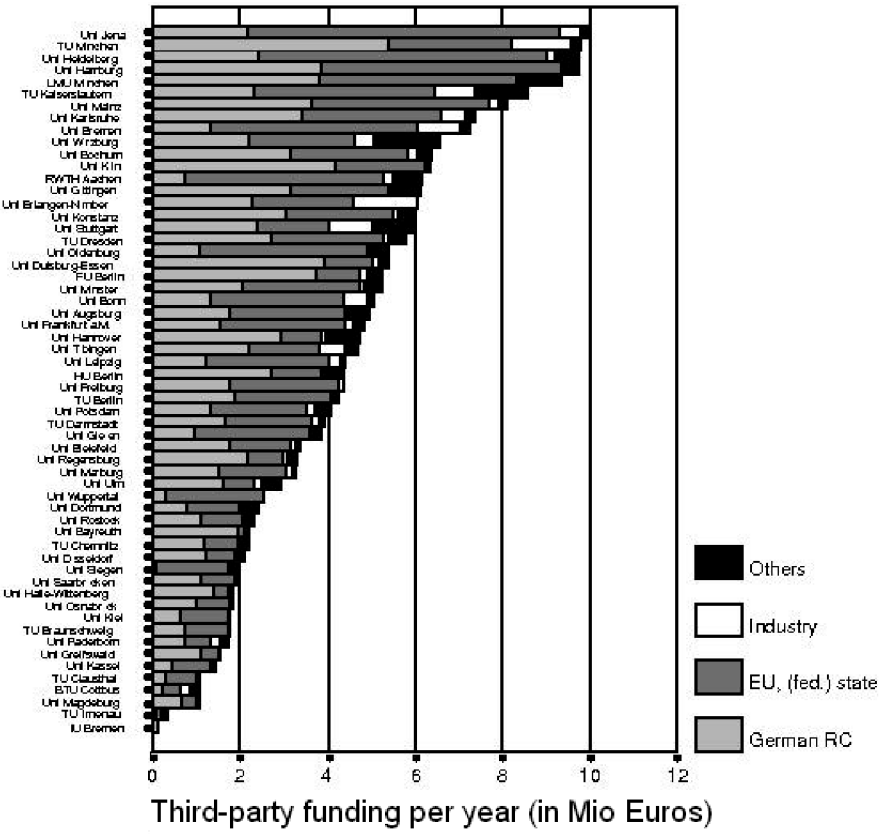
Composition of third-party funding

Regarding third-party funding, unsurprisingly, there exist large differences between the different disciplines. For example, the overall percentage of third-party funding by the German Research Council in mathematics is about 56 %, more than a half of the total third-party funding. On the other hand e.g. in medicine, the percentage of funding by the German RC accounts for about a quarter of the total amount. This clearly shows the interest to present, for each subject, the distribution of third-part funding according to the various potential sources: German RC, EU projects, government, federal state government, industry or foundations, other origin.

Moreover, in each specific subject field, the composition of third-party funding may differ a lot from one department to another. Looking at Figure 4 which shows the composition of third-party funding for the physics departments of German universities, it is obvious that by using only a single source for third-party funding e.g. the money coming from the German RC, the sequence of the departments would

change dramatically. The university in Duisburg-Essen or the Berlin Free University would rank much higher in a ranking based on German RC funding only.

Figure 4 Composition of third-party funding for physics departments



Other disciplines show a much more visible concentration of the amount of third-party funding in very few universities. For example in medicine, the Charité in Berlin and the Ludwig-Maximilian-University together share about 20 % of the total sum of third-party funding; in mathematics the first three together receive about 20 %; in physics, five universities are needed to reach this percentage.

These analyses and figures are often used by universities to analyze their performance and to make comparisons with other departments.

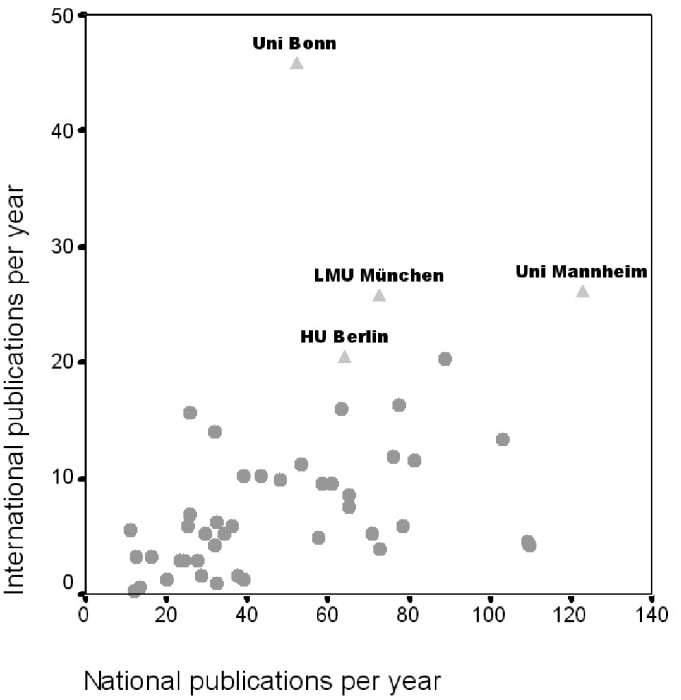
Publication analysis

An interesting discussion arose when in 1998 the first results of a publication analysis for economics was published. The results seemed to contrast with all analy-

ses that had been carried out by economists themselves during previous years. Everybody judged that because of that the CHE results must be wrong.

The different results were due to the different methods used. Rankings known to economists were mostly based on articles in international journals. In contrast to that, the CHE used a database that contained books and articles in edited volumes as well as a lot of national publications in the German language. The reason for choosing this approach was that one could not expect too many international journal articles to be written by members of the German economics departments so the numbers might not be sufficient for a national comparison of the publication output. So the indicator based on the original CHE method gives a picture of the output in general, whereas the indicator based on the first mentioned approach represents the international visibility of a department.

Figure 5 Economics departments in Germany



Since 2005, the CHE Research Ranking publishes two indicators for economics, one based on the more national-oriented database and a second based only on articles in international journals. Contrasting these two indicators shows the different profiles of economics departments in Germany (see Figure 5). The University of Bonn shows a distinct international profile concerning publication output, the University of Mannheim performs very well on both indicators, whereas other universities perform very well on the national one but are internationally not visible.

6. Conclusions

Among the different instruments of quality assessment in higher education, rankings probably receive the most public attention. Rankings are a growing phenomenon in higher education and are published in many countries throughout the world. Despite their controversial nature, they are here to stay as they correspond to a need for transparency about higher education in an increasingly competitive system. The primary aim of rankings is to create transparency about higher education from an external and comparative perspective. Institutional enhancement is at best a secondary aspect of rankings. Nevertheless, their results are taken seriously by the institutions ranked – in terms of marketing, with regard to strategies for climbing in league tables (up to a degree that could be classified as neither the intention nor the purpose of rankings) but also in a way that universities seek to cope with weaknesses identified by rankings. It is only in this sense that rankings can contribute to the quality assurance of institutions. They can be a starting point for institutions to analyze their strengths and weaknesses compared to their competitors'.

The CHE Ranking developed a particular methodology that was appraised very positively by several comparative studies on rankings (Usher & Savino, 2006; Marginson and van der Wende, 2007). This approach is distinct from mainstream ranking as it refers to fields/programmes instead of whole institutions, is multi-dimensional and rejects the over-simplification of calculating a single composite indicator out of weighted indicators and avoids exaggerating differences in performance inherent to league tables by placing universities into three groups.

For the CHE University Ranking – mainly intended for prospective students who have to find a university and including indicators relative to teaching and learning, resources and facilities, and research activities – these methodological principles together with the interactive and individualized way of presenting the results on the web version give detailed insights into strengths and weaknesses of departments, providing in this way a profile of the latter. This ranking serves both the need of prospective students helping them to find the best university *for them*, as well as the need of the faculties/departments and researchers themselves to compare with other institutions.

The CHE Research Ranking gives detailed insight into the research performance of German universities. Besides a ranking of the universities concerning their subject specific research performance, it presents profiles of strong departments/institutes for each subject. Furthermore, information is given on the composition of third-party funding or on the correlations between different research indicators. This makes the CHE Research Ranking a useful benchmarking tool for universities.

References

Berghoff S. & S. Hornbostel (2003), "Das CHE hinter den sieben Bergen", *PWP Perspektiven der Wirtschaftspolitik Eine Zeitschrift des Vereins für Socialpolitik*, pp. 191-195.

Berghoff S., G. Federkeil, P. Giebisch, C.-D. Hachmeister, M. Hennings & D. Müller-Böling (2008a), *CHE ForschungsRanking 2007*, AP102, Gütersloh

Berghoff S., G. Federkeil, P. Giebisch, C.-D. Hachmeister, M. Hennings, D. Müller-Böling & I. Roessler (2008b), "*Das CHE-Hochschulranking. Vorgehensweise und Indikatoren*", CHE-Arbeitspapier 106, Gütersloh, http ://www.che.de/downloads/Methoden_Hochschulranking_2008_AP106.pdf.

Berghoff S. *et al*. (2008c), "Identifying the Best: The CHE Ranking of Excellent European Graduate Programmes in Natural Sciences", CHE Working paper 99, Gütersloh, http://www.che.de/downloads/CHE_ExcellenceRanking_2007_AP99.pdf.

Clarke M. (2007), "The Impact of Higher Education Rankings on Student Access, Choice, and Opportunity", Institute for Higher Education Policy, pp. 35-48.

Dill D. & M. Soo (2005), "Academic Quality, League Tables, and Public Policy: A cross National Analysis of University Ranking Systems", *Higher Education*, 49, pp. 495-533.

Marginson S. & M. van der Wende (2007), "Globalisation and Higher Education", OECD Education working paper series, no. 8.

Sadlak J. & Nian Cai Liu (ed.) (2007), *The World-Class University and Ranking: Aiming Beyond Status*, UNESCO-CEPES, Bucharest.

Usher A. & M. Savino (2006), *A World of Difference. A Global Survey of University League Tables*, Toronto, http://www.educationalpolicy.org/pdf/World-of-Difference-200602162.pdf (on 15 September 2008).

On the "multi-dimensionality" of ranking and the role of bibliometrics in university assessment

Wolfgang GLÄNZEL and Koenraad DEBACKERE

Summary

The complexity of university activities does not allow the reduction of the multidimensional space of those activities and their outcomes into one dimension of linear ranking. The difficulty of quantification as well as the all too frequently experienced arbitrariness in defining composite indicators often result in an inadequate representation and an irreproducible product and hence are in clear conflict with the *Berlin Principles on Ranking* of Higher Education Institutions. Even focussing on one single, however important aspect, such as the assessment of research performance, remains a multifaceted endeavour. Using the example of bibliometrics, we point to caveats and pitfalls in the challenge of comparative research assessment of colleges and universities.

KEYWORDS. Bibliometrics, research evaluation, composite indicators, university ranking.

1. Introduction

Performance-based listing of research and education, and above all, the academic ranking of colleges and universities, has become one of the most favourite issues in the assessment of higher education institutions. At least since the publication of the first edition of the Shanghai Jiao Tong world university ranking in 2003 (ARWU, 2007) and the successive lists, such as the *Times Higher Education Supplement* – QS World University Rankings in 2005 (THES-QS, 2007), the comparative evaluation of the quality of Higher Education Institutions (HEIs) has been brought into the focus of public and policy interest. World rankings have been followed by national lists in several European countries, in Canada and the US. Although their methodology has been improved since and guidelines for quality management (see Berlin Principles on Ranking of Higher Education Institutions compiled by the International Ranking Expert Group [IREG, 2006]) have been elaborated, university ranking remains controversial. Methodological and general issues such as the question of how complex multidimensional criteria can be transformed into linearity have been

addressed, and are at present discussed with keen interest. In short, the complexity of university activities does not allow the reduction of the multidimensional space into one dimension of linear ranking. The difficulty of quantification and arbitrariness in defining composite indicators result in an inadequate representation and an irreproducible product leading to a clear conflict with the Berlin Principles on Ranking of HEIs. Even focussing on one single, however important aspect, such as the assessment of research performance, remains a multifaceted endeavour. Proceeding from our experience, we illustrate this with two examples. The first one describes the clustering of research institutions on the basis of their publication profiles for comparison of institutional research performance among *likes* and therefore to avoid the effect of "comparing apples with oranges". The second example visualises a "two-dimensional approach" to university ranking, proceeding from second-generation relational charts. Based on these examples, the study also points to caveats and pitfalls in the challenge of comparative research assessment of universities.

2. A concise discourse on ranking?

Before we tackle the question of to what extent reliable and reproducible ranking lists are at all possible, we attempt to clarify the notion of *ranking* by presenting the following comprehensible but nonetheless precise definition. In verbal terms, ranking is *positioning comparable objects on an ordinal scale based on a (non-strict) weak order relation among (statistical) functions of, or a combination of functions of measures or scores associated with those objects*.

These (mainly statistical) functions, which are usually based on variables for evaluative purposes, are called *indicators*. Different indicators X_k representing different aspects of quality, form the components of a *composite indicator Y,* the basis of the ranking; this composite indicator is usually a linear combination of the X_k's, that is,

$$Y = \Sigma\, \lambda_k \cdot X_k,$$

where λ_k ($k = 1, 2, \ldots, p$) are p pre-defined weightings and, without loss of generality, verify the equality $\sum \lambda_k = 1$ (this last relation implies that Y is actually a *weighted mean* of the individual indicators X_k). The use of composite indicators always reflects a certain arbitrariness and a level of simplification as we will show below. The most problematic issues in applying composite indicators are listed below.

— *Possible interdependence of components*

The underlying variables represent factors influencing performance. These factors are often not separable and, consequently, individual variables do not amount to one unique factor each. Variables are therefore often interdependent. For instance, the variables *funding*, *personnel*, *publication output*, *citation impact*, *peer reviews* are not independent. A change on one variable can therefore have unpredictable effects upon other variables defining the composite indicator.

— *Altering weightings can result in a different ranking*

The choice of weightings is in practice arbitrary. The selection is guided by the rankers' preferences rather than by methodological or empirical findings.

Rounding to "plausible" values, e.g. 10 % or 25 %, further emphasises this arbitrariness.

— *Results might be obscure and irreproducible*

The mixture of possibly incommensurable indicators, superposition of interdependent variables and arbitrary weighting can make the methodology obscure and the results obtained irreproducible.

— *Random errors of statistical functions are usually ignored*

Composite indicators are linear combinations of statistical functions which can themselves be subject to random errors. The standard errors of the statistics – such as means and shares – are influenced by the size set of objects measured by the variable in question and the underlying probability distributions. Different positions in the ranking list might therefore be interpreted as ties.

— *Multi-dimensional space is reduced into linearity*

This is one of the most crucial issues in ranking. From the mathematical viewpoint, a linear combination as applied through the composite indicator is a result of a projection into a subspace. Since projections are irreversible, valuable information is definitely lost by reducing multi-dimensional space into linearity.

Besides the aforementioned statistical and methodological problems, several data-related issues are relevant as well. In the first place, we mention the "cleanness", compatibility and hence the reliability of the data used. Data collection for large-scale ranking still remains a challenge if it is at all feasible. The time-variant nature of the underlying data sources is a further problem. Thus the incorrect institutional assignment of staff or research-output data taken from different sources might result in incompatibility issues. Combining alumni and staff winning Nobel Prizes and Fields Medals with recent publication- and citation-related data might serve just as an example for such a problematic methodological approach.

3. Selective vs. integrated ranking

In order to account for the complexity of university activities, two basic approaches are possible: selective ranking and integrated ranking. *Selective* ranking focuses on measuring and ranking according to one selected activity whereas the *integrated* or "holistic" ranking procedure attempts to capture the complex set of all or at least of the most important activities. The advantages of the first method are obvious. As compared with the holistic approach, information loss and incommensurability can be reduced and reliability can be increased. Of course, individual lists have to be prepared for each activity aspect. In the following section we give a concise description of examples for selective and integrated college and university ranking.

A. Evaluation of education

In 1993 a national education-related university ranking was published in Germany (*Der Spiegel-Spezial*, 1993). The ranking was survey-based. Questionnaires

had been sent to students and professors. A breakdown by fields was presented as well to give a more differentiated picture, to reveal "strengths and weaknesses", and to help students and academic staff make a selection. Because of differences and peculiarities of national educational and accreditation systems, such endeavours are practically restricted to the national level.

B. Research performance

With the Shanghai Ranking (ARWU, 2007), first published in 2003, the focus was shifted to research assessment. The composite indicator build by Shanghai Jiao Tong University is used to rank the world's major institutes of higher education on the basis of the following weighted key indicators, alumni winning Nobel Prizes and Fields Medals (10%), staff winning Nobel Prizes and Fields Medals (20%), highly-cited researchers according to highlycited.com (20%), articles published in *Nature* and *Science* (20%), publications indexed in the *Science Citation Index – Expanded* (SCIE) and the *Social Sciences Citation Index* (SSCI) of Thomson Scientific (20%) and the size of the institution (10%). This world-wide ranking was to a large extent facilitated by the availability of the multidisciplinary bibliographic database *Web of Science* and its derivatives.

C. "Holistic approach"

The broader approach chosen by THES-QS, which largely relies on *peer review score*, could not overcome the limitations of previous attempts and remained controversial as well. It actually marks a new direction in university ranking, particularly the trend towards integrated evaluation. The holistic approach, i.e., the comprehensive and integrated quantification of university performance and a world-wide ranking based on *all* HEI activities, including education, research and third mission, however, remains utopian at least for the present.

The question arises whether there is really any need for an integrated ranking. The evaluation of selected activities within the HEI missions (such as quality of education, research performance or the assessment of important third-stream activities) might provide more valuable information for the interested users in the relevant sectors and domains.

The *Centre for Higher Education Development* (CHE, 2007) has chosen a third route. Their approach is strictly subject-oriented but the evaluation extends to both research and education. The ranking is based on bibliometrics and questionnaires. Although CHE aims at internationalisation, its methodology remains subject to the above-mentioned limitations.

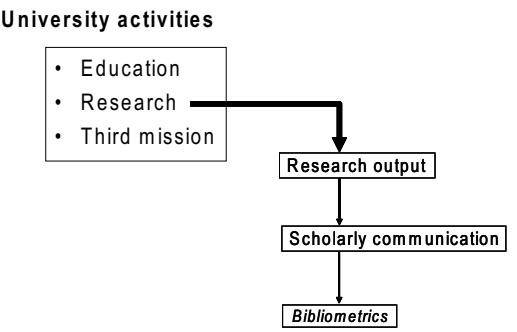
4. Bibliometrics and the "multi-dimensionality" of research activity

In this section, we take a critical look at the possible role of *bibliometrics* in (selective) university ranking. Although measuring only *one*, however, important part of research activities, bibliometrics proved an efficient tool in research assessment. Figure 1 sketches the position and function of bibliometrics in quantifying and measuring activities of higher education institutes.

As in the case of *all* HEIs rankings, first and foremost the following two issues have to be solved for the bibliometric approach: the quality of results stands and falls with the correctness of data collection, pre-processing data and the application of sound methodology. This includes correct institutional assignment and the selection of normalised standard indicators that guarantee the robustness and the reproducibility of results.

Another issue arises from the institute-specific specialisation or diversification as even multi-disciplinary research and education institutions usually have more specific research profiles. Thus the practice in institutional evaluation is benchmarking and comparison of institutional performance with reference institutions with similar research profiles. Computerised or semi-computerised classification of research institutions according to their publication profiles (e.g., Thijs and Glänzel, 2008, 2009) can assist both the selection of reference units and the realisation of comparative analysis. From the perspective of validity, comparison of institutions with completely different mission and research profiles should of course be avoided. Although the quantification of research output should in principle allow such treatment, putting business schools and medical schools on the same list would not make sense. On the other hand, large universities with originally different profiles like medical and technical universities do have overlapping research activities. Thus the methodology applied should nevertheless be suited for intra- and inter-class comparison where and whenever this makes sense. Finally, an efficient method to further compensate for the biases caused by subject-specific profile heterogeneity in the context of specialisation and diversification, is the consequent standardisation and normalisation of the bibliometric indicators to eliminate subject-specific biases.

Figure 1 Function of bibliometrics in quantifying and measuring activities of HEIs



In order to obtain a more realistic and differentiated picture of research at higher-educational institutions, the following *three scenarios* are suggested.

I. Clustering of similar objects
II. Breakdown by fields
III. Standardisation of indicators

The proposed issues are preconditions for the correct use and interpretation of bibliometrics-based indicators and should therefore be applied in combination.

A. Clustering of similar objects

Ranking HEIs with completely different profiles as, for instance, based on the comparison of medical schools with business schools, still remains an exercise of "comparing apples with oranges". In order to find an appropriate profile classification for universities, colleges and research institutes, we have clustered more than 2,000 institutions from fifteen European countries[1] according to their publication profiles in the period 2001-2003. The stopping rule introduced by Duda and Hart (1973) was applied to determine and to optimise the number of clusters. The optimum has been found at eight profile clusters. Table 1 presents the classification of European institutions according to Thijs and Glänzel (2008, 2009).

Table 1 The eight clusters resulting from the optimum solution

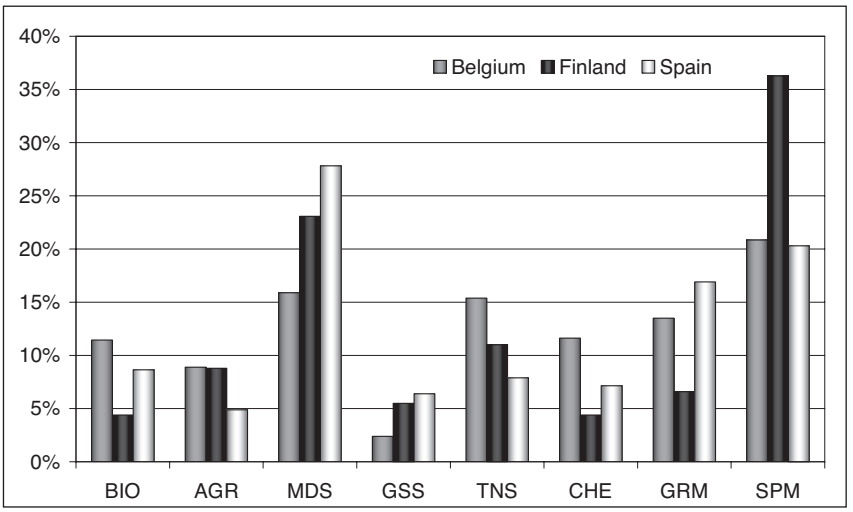
Cluster		*Code*
Cluster 1	*(Biology)*	BIO
Cluster 2	*(Agriculture)*	AGR
Cluster 3	*(Multidisciplinary)*	MDS
Cluster 4	*(Geo & Space Science)*	GSS
Cluster 5	*(Technical & Natural Sciences)*	TNS
Cluster 6	*(Chemistry)*	CHE
Cluster 7	*(General & Research Medicine)*	GRM
Cluster 8	*(Specialised Medicine)*	SPM

Source: Thijs & Glänzel (2008, 2009) based on WoS (Thomson Scientific).

The application of the university classification alone seems to be insufficient in practice. Using the example of Belgium, Finland and Spain, one can easily see that national characteristics of scientific research in higher education might strongly influence the constitution of clusters in different countries (see Figure 2). Ranking according to clusters would therefore be biased by national representation. Therefore we suggest the additional application of one of the following scenarios.

[1] This set comprises fifteen countries, namely Switzerland and the members of the European Union *before 2004* (EU15) except Greece.

Figure 2 Examples for different national cluster profiles



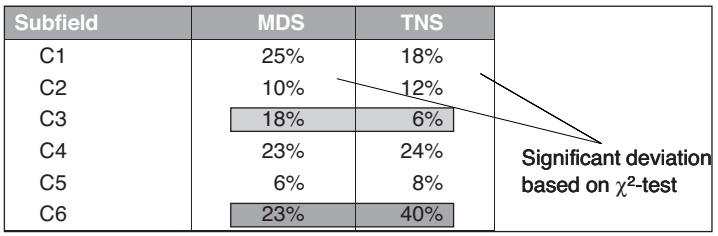
Source: Thijs & Glänzel (2008, 2009) based on WoS (Thomson Scientific).

B. Breakdown by fields

The research performance of a university might differ among its faculties, departments and thus in different fields. Institute-specific specialisation is often contrasted by, or even combined with, diversification. Research in the same field carried out by different institutions can still have different profiles as shown in Figure 3. The profile of chemistry research in multidisciplinary universities significantly deviates from that in technical universities, although the overlap is, of course, considerable. In particular, 18 % of the publications in chemistry research coming from multidisciplinary universities (MDS) belong to the organic and medicinal chemistry subfield (C3); the corresponding percentage for the technical universities (TNS) is equal to 6 %. On the other hand, the part of publications classified in the materials science subfield (C6) is much higher for the technical than for the multidisciplinary universities (40 % for the technical universities vs. 23 % for the multidisciplinary universities).

The consequences of these institutional peculiarities are obvious: overall gross publication and citation counts can be misleading when ranking institutions with multidisciplinary and specialised research profiles. This effect can be reduced by breaking down institutional research activity by fields and subfields. This breakdown might further help reveal institutional "strengths and weaknesses".

Figure 3 Example of the deviating field structure in different clusters[2]



Source: Thijs & Glänzel (2008, 2009) based on WoS (Thomson Scientific).

C. Standardisation of indicators

An effective method to further compensate for biases caused by subject-specific profile heterogeneity in the context of specialisation and diversification, is the consequent standardisation and normalisation of the bibliometric indicators used in comparative studies. In an earlier study (Glänzel *et al*., 2009), we described an appropriate set of adjusted standard indicators that meets the requirements of meso-level analyses. In particular, proceeding from the indicators used in Budapest and Leuven, we defined an adequate level of standardisation that makes it possible to use standard indicators for both intra- and inter-cluster analysis, for domain-specific as well as multidisciplinary studies and their adequate graphical presentation. The relative indicators developed in Budapest in the 1980s and preferably presented in relational charts (e.g. Schubert and Braun, 1986), were used as the starting point for the development of the instruments for cross-institutional comparisons. While in the relational charts the *Mean Observed Citation Rate* (MOCR)[3] was plotted against the journal-based *Mean Expected Citation Rate* (MECR)[4], the new version of relational charts uses subfield normalised observed and expected citation rates to avoid possible biases caused by subject-specific peculiarities or by different activity profiles as described above. Subject-normalisation is done by dividing the two indicators by the corresponding values of the subject-based *Field Expected Citation Rate* (FECR)[5]. A detailed description of these indicators can be found in Glänzel *et al.* (2009).

While the *y*-axis presents the factual "performance" measured through citations, the *x*-axis stands for the impact standard of the journals in which the institution publishes (see Figure 4). Both measures shed light on two important aspects of research

[2] The subfield abbreviations according to the Leuven subject-classification scheme (Glänzel and Schubert, 2003) are C1 – analytical, inorganic & nuclear chemistry, C2 – applied chemistry & chemical engineering, C3 – organic & medicinal chemistry, C4 – physical chemistry, C5 – polymer science, C6 – materials science.

[3] Mean observed citation rate (MOCR) is defined as the ratio of citation count to publication count.

[4] The journal-based expected citation rate of a single paper is defined as the average citation rate of all papers published in the same journal.

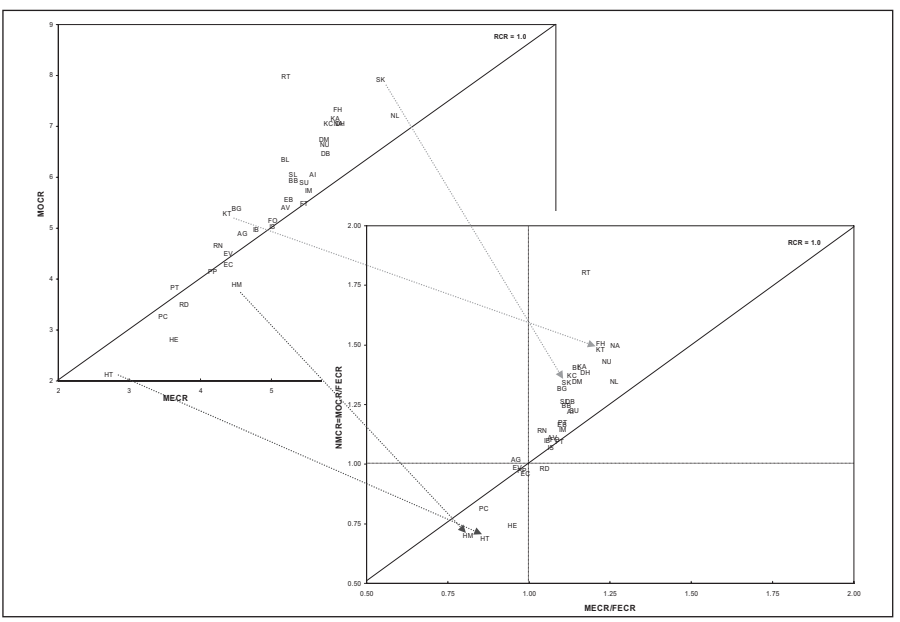
[5] Analogously to the MECR, the FECR of a single paper is defined as the average citation rate of all papers published in the same subject in the same year.

assessment, in particular "visibility-" and "quality-related" issues. In order to overcome profile-specific biases (which might occur even within the same profile cluster) a strict standardisation and normalisation of indicators should be applied. The two straight lines in the lower right-hand chart of Figure 4, whose equations are $x = 1$ (for the vertical line) and $y = 1$ (for the horizontal line), indicate two equilibrium situations, particularly, the conformity with the corresponding underlying reference standards (Braun and Glänzel, 1990). A further relative indicator, called *Relative Citation Rate* (RCR), is obtained as the ratio of the two indicators. The straight line with equation $y = x$ indicates the balance between observation and expectation.

The effect of normalisation is demonstrated using the following example. The observed citation impact of 39 European universities in thirteen selected countries (including the three largest HEIs each of the corresponding country) is plotted against its expectation, once without and another time with subject normalisation. The publication period is 2001-2003 and a three-year citation window for each publication year has been applied. The changing "ranks" of medical and technical universities in the two-dimensional presentation are quite impressive.

The scatterplot of (MECR, MOCR) values is presented first in a traditional relational chart (see upper left-hand corner of Figure 4). Both expectation and observation cover quite a large range of citation impact. The fact that SK is a medical university whereas KT is a technical university, certainly contributes to the huge

Figure 4 Traditional and subfield-normalised relational chart for 39 selected universities



Source: Glänzel *et al.* (2009) based on WoS (Thomson Scientific).

deviations of their respective citation impact indicators. We also mention that both impact indicators (MECR and MOCR) of SK considerably exceed those of the technical university KT. We observe a similar situation for HM and HT, however, at a much lower level. Also HT, as a technical university, appears in the low-end group of this diagram. We obtain a completely different situation if subfield-based normalisation is applied.

The plot of subfield-normalised observation against subfield-normalised expectation can be found in the lower right-hand corner of Figure 4. The positions of universities SK and KT have interchanged; the same applies to HM and HT. The effect of field-specific lower impact of technical universities and the usual high impact of medical universities has thus been eliminated.

Two lessons can be learned from this example. (1) Non-normalised, non-standardised counts, such as gross-publication or gross-citation counts, can be strongly affected by university-specific profiles. This effect is measurable even if shares or mean values are used. Only appropriate normalisation can compensate and (nearly) eliminate profile-specific biases. (2) Reducing dimensions means losing information, and might hence result in misinterpretations. The universities AG and NL have almost the same RCR value although both institutions hold different positions in the two charts (cf. Figure 4). While both expected and observed citation impact of AG are in line with the world standard, the indicator values of NL reveal that this university belongs to the high-end with respect to research performance.

5. Conclusions

The idea of ranking HEIs according to simple, seemingly objective and robust indicators is perhaps tempting. However, robustness is easily lost by building composite indicators with partially interdependent or even incompatible components and arbitrary weightings. Reality is more complex than can be described this way. Instead of any linear ranking of colleges and universities, a more detailed, complex analysis is necessary to capture and reflect several important aspects of performance among the manifold activities of a university.

Bibliometrics can contribute to the evaluation of at least *one* of these aspects. One lesson from bibliometrics is that standardisation and normalisation help eliminate biases and facilitate longitudinal ranking analysis as well. Another lesson from bibliometrics is that even normalisation of indicators cannot disguise the fact that comparing HEIs with *completely different profiles* nonetheless remains an exercise of "comparing apples with oranges".

References

ARWU (2007), *Academic Ranking of World Universities*, accessible via: http://ed.sjtu.edu.cn/ranking.htm.

Braun T., W. Glänzel (1990), "United Germany: The New Scientific Superpower?", *Scientometrics*, 19(5-6), pp. 513-521.

CHE (2007), *CHE University Ranking*, accessible via: http://www.che-ranking.de.

Duda R.O., P.E. Hart (1973), *Pattern Classification and Scene Analysis*, Wiley, New York.

Glänzel W., A. Schubert (2003), "A new classification scheme of science fields and subfields designed for scientometric evaluation purposes", *Scientometrics*, 56(3), pp. 357-367.

GLÄNZEL W., B. THIJS, A. SCHUBERT, K. DEBACKERE (2009), "Subfield-specific normalized relative indicators and a new generation of relational charts: methodological foundations illustrated on the assessment of institutional research performance", *Scientometrics*, 78(1), pp. 165-188.

IREG (2006), *Berlin Principles on Ranking of Higher Education Institutions*, accessible via: http://www.che.de/downloads/Berlin_Principles_IREG_534.pdf.

SCHUBERT A., T. BRAUN (1986), "Relative indicators and relational charts for comparative-assessment of publication output and citation impact", *Scientometrics*, 9(5-6), pp. 281-291.

Spiegel-Spezial (1993), *Welche Uni ist die beste?*, 3/1993, pp. 3-168.

THES-QS (2007), *Times Higher Education – QS World University Rankings 2007*, accessible via: http://www.topuniversities.com/worlduniversityrankings/results/2007/overall_rankings/top_400_universities/.

THIJS B., W. GLÄNZEL (2008), "A structural analysis of publication profiles for the classification of European research institutes", *Scientometrics*, 74(2), pp. 223-236.

THIJS B., W. GLÄNZEL (2009), "A structural analysis of benchmarks on different bibliometrical indicators for European research institutes based on their research profile", *Scientometrics*, 79(2), in press. DOI: 10.1007/s11192-009-0425-z.

A statistical approach to rankings: some figures and explanations for European universities

Michel LUBRANO[1]

Summary

The Shanghai ranking is based on indicators that are too sparse for European universities. Publications provide much more statistical information which is used in this paper to rank European economics departments. Publications are considered as random variables, so that a standard deviation can be associated to the total score of a department. And we can test if two departments are statistically different. Finally, a multilevel model is adjusted to the data that provides another way of ranking departments. It opens the way to explaining why some departments are more productive than others. We provide stylised facts which contrast small northern European countries and big southern European countries.

KEYWORDS. Ranking, research, economics departments, statistical significance, multilevel method.

1. Introduction

In 2000, the European Economic Association launched a research programme on the ranking of economics departments in Europe that finally involved four different teams. Lubrano *et al.* (2003) was the main contribution of the Louvain-Marseille team to this project. Since 2003, many events contributed to change what we said at that time, including the last conference held at the Université Libre de Bruxelles in December 2007. This paper aims at summarising the main results of Lubrano *et al.* (2003) while introducing some new material.

[1] This paper is a revised version of the paper given at the conference Ranking and Research Assessment in Higher Education held in Brussels, 12-13 December 2007. The author thanks all the participants for their comments and especially Luc Bauwens. The editorial comments of Dirk Jacobs were crucial in preparing the present version of the paper. A special mention should be made to Abhishek Chandan and to Mathieu Goudard for their computational assistance. Remaining errors are solely mine. Financial support of the ANR research project NT05-3-41515-STAHN-Hubert: Economie de la Connaissance is gratefully acknowledged.

During the last five years, the international competition between universities and countries increased a lot. The Jiao Tong university of Shanghai produced its famous ranking for internal purposes at the origin. It was targeted at the Chinese government with the aim of showing that the Jiao Tong university compared favourably to other universities in the world and thus deserved the money that it received. Very rapidly, foreign countries took interest in this ranking, presumably for two reasons. First, it was done by a country which could be supposed to be independent and objective. Everybody assumed that the best universities were located in the Western world. So China had no interest in manipulating this ranking. Secondly, China is sending abroad a huge number of students. It is fairly straightforward to suppose that this ranking was also used to dispatch worldwide this enormous potential of Chinese students in an efficient way.

We shall focus our attention on European universities and more specifically on their economics departments. As a guideline, we have followed the paradigm of a graduate student looking for a PhD programme in Europe. This student has to apply for a grant. He must choose first a country where to apply, and then inside that country a particular economics department. This is a decision problem under uncertainty, because he does not know all the characteristics of the country and because if he chooses a particular country, he might be admitted in a university which would not be his first choice. To solve this decision problem, we assume that the student maximises his utility function and that he has access to observations of a random variable attached to the various opportunities among which he has to choose. This problem was formalised in Lubrano and Protopopescu (2004), building on the classical results of the decision theory literature. The result is that the student chooses the country that dominates the others for the given observed random variable in a sense that we shall define latter on. We then generalise this approach to rank universities and departments.

The paper is organised as follows. In section 2, we select the random variable that can best help to solve the student's decision problem. In section 3, we detail various concepts attached to the definition of a university and an affiliation. Section 4 gives a short summary of a mathematical theory for ranking. In section 5, we apply this theory for ranking European economic departments and testing their differences. In section 6, we provide some stylised facts that could explain the obtained ranking. In section 7, we investigate the capabilities of multilevel models to provide a statistical basis to obtain an alternative ranking and that could take into account exogenous variables related to economic policy. Section 8 concludes.

2. Which indicator for ranking universities

Universities are institutions producing knowledge. They are using various inputs and have a large variety of outputs. They can be ranked on the quality of these outputs while their efficiency can be measured by the quantity of needed inputs in order to achieve their outputs.

The first visible output concerns the number of students who get a degree, undergraduate or PhD. More important is the future of these students, what they will become, their wages and/or their future scientific achievements.

The second visible output concerns research. There are all sorts of variables which can be used to measure and qualify this output. The most immediate one is the number of publications, possibly weighted by their quality. A more refined criterion is based on the citations of a paper in order to appraise its impact on the scientific community. We have then criteria which are much more elitist because they concern less and less people such as being the editor of a journal, being a member of the restricted list of highly cited researchers, holding a scientific prize or finally receiving the Nobel Prize.

The last visible output concerns the impact of the university on the outside economic world. It concerns joint ventures with the industry, the creation of scientific parks and the holding of patents. This last type of output is rarely taken into account for rankings.

A. Analysing the Shanghai ranking

The University Jiao Tong of Shanghai uses a mix of different criteria based on the quality of teaching measured only by the number of Nobel Prizes and Fields Medals among former students, the quality of the institution measured as the number of Nobel Prizes and Fields Medals and the number of highly cited researchers among the staff of the university and finally two publication indicators (one based solely on the number of papers in *Nature* and in *Science*, the other based on the number of papers in the *Science Citation Index* and in the *Arts and Humanities Citation Index*; the *Social Science Citation Index* is ignored). This ranking is published every year.

Let us try to appraise the relevance of these criteria by considering the list of the top 100 universities in the world published for 2004, 2005, 2006 and 2007. If these rankings are done in a efficient way, they should not vary too much from year to year. We have computed the Spearman rank correlation between the ranking of 2007 and those of 2006, 2005, and 2004. These correlations should slightly decrease as time elapses, but should remain relatively high, just because universities are slowly evolving institutions. We made these computations for the top 100 universities in the world, and then for the sub-sample of top 36 European universities.

Table 1 Stability and un-stability of the Shanghai ranking

Period	*Number of obs.*	*2007-2006*	*2007-2005*	*2007-2004*
World	106	0.993	0.985	0.968
Europe	36	0.904	0.814	0.646

The last three columns indicate the Spearman rank correlation coefficient between the ranking obtained for two different years. Six universities do not appear in all 4 rankings. We ranked them 101-106, using the alphabetical order.

Table 1 shows that the world ranking is fairly stable, because the Spearman rank coefficient of correlation decreases only slightly. When we turn to Europe, the correlation between the ranks is roughly the same at the beginning, but it diverges rapidly. This decrease in correlation means that the European ranking is not stable across time. We cannot convincingly suppose that European universities were so

much affected by the Shanghai ranking that they dramatically improved their governance in four years. It is more realistic to assume that their ranking is not based on a sufficiently large information set and that their ranking is blurred by statistical noise. If we want to rank European economics departments, we should use criteria which provide more information per department. The Shanghai ranking does not use enough observations per university so as to get convergent information, except for the very major universities.

B. The van Damme formula

Among the four sources of data used in the Shanghai ranking, many have zero observations for medium range universities: the number of Nobel Prizes or even the number of highly cited researchers. For instance, France has only four highly cited researchers in business/economics. The number of publications provides a much denser information set. The *Social Science Citation Index* collects the papers of 160 journals, while the CD-ROM of the *Journal of Economic Literature* concerns more than 650 journals. This data base covers economics in a broad sense as it includes also econometrics, finance and management. A large place is devoted to national journals, so that the language bias should not be too important. We have chosen it as our main source of information.

The van Damme (1996) formula is an aggregation rule which determines the publishing score of an author. It is based on counting the number of publications of an author for a given year and giving them an appropriate weight.

Definition 1 *A researcher i is attributed a score* $q_{i,j,t}$ *for his publication* p_j *that appeared in year t. This score is defined by*

$$q_{i,j,t} = \frac{b(p_j)}{a(p_j)}\, v(p_j), \tag{1}$$

where $b(p)$ is a number related to the length of the publication, $a(p)$ is a number related to the number n_j of authors of the publication, $v(p)$ is a number related to the quality of the publication. The total scores $s_{i,t}$ of a researcher i during year t is equal to the sum of the scores of the $n_{i,t}$ publications to which he contributed during year t:

$$s_{i,t} = \sum_{j=1}^{n_{i,t}} q_{i,j,t}. \tag{2}$$

The use of the van Damme formula implies that many choices have to be made. We must choose actual numbers and definitions for $a(p_j)$, $b(p_j)$ and $v(p_j)$.

— $a(p_j)$ is usually equal to the number of co-authors. But we might want to favour co-authorship. In this case, $a(p_j)$ is taken equal to the square root of the number of co-authors.

— $b(p_j)$ is usually equal to the number of pages of the article. But this might introduce unwanted volatility in the measurements, so Lubrano *et al.* (2003) have chosen to set it equal to 1.

— $v(p_j)$ is an index meant to formalise that a paper published in a confidential journal is worth less than a paper published in a top international journal. We have chosen to take 10 for a small group of top international journals and 1 for low quality national journals. There is of course a grading for intermediate journals. Note that the global scaling is arbitrary. This index requires the existence of a journal ranking, that we discuss below.

There are basically two ways of ranking journals. A first method is based on citations and makes use of impact factors. The ranking published in Combes and Linnemer (2001) is based on expert opinions. In Lubrano *et al.* (2003), we have taken the Combes and Linnemer (2001) ranking and we have confronted it to citation data to produce an updated ranking. See our paper for more details.

The score of an author is a random variable for which we have observations. The randomness can be attributed to several facts that are essential to understand. There is a variable time between submission and publication. The probability of acceptance of a paper is influenced by the choice of the names of the referees. The choice of the journal to which the paper is submitted is not necessarily optimal. The contribution of each author to an article is not necessarily the same. The length of a paper is not always strictly related to its quality and impacts. Some of the notes published in *Econometrica* receive much more citations than regular articles.

The production of an author might vary a lot across the year either because of cyclical productivity or because the author is at the beginning or the end of his professional life cycle. Consequently, it is wise to smooth the production of an author by cumulating it over a large span of time. Lubrano *et al.* (2003) have chosen a span of 10 years covering 1991 to 2001.

C. An informal evaluation of countries

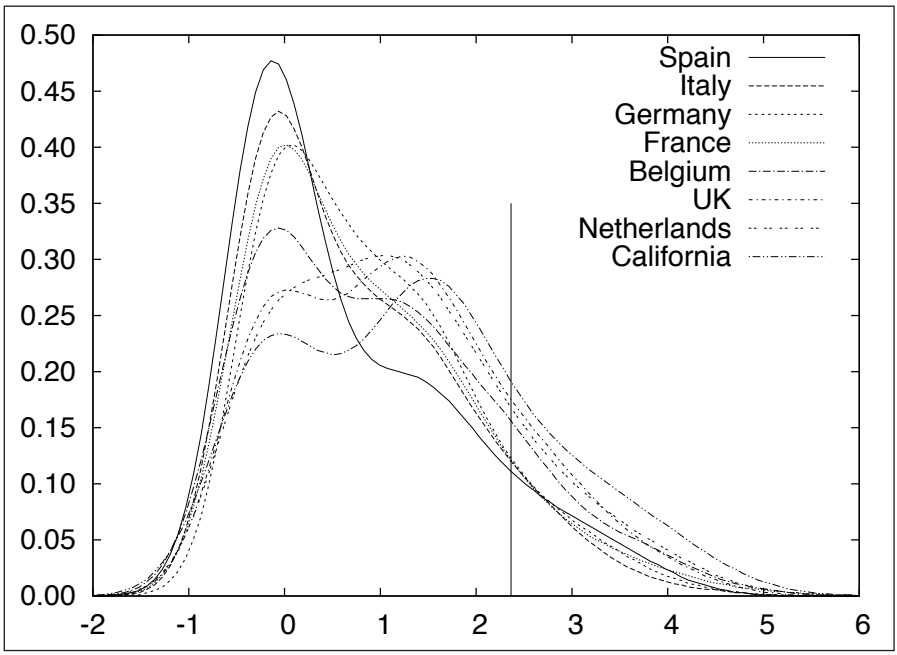
The basic observation is a researcher for which we provide a score computed using formula (2). We know his country of origin, but for the moment we ignore his institutional affiliation. Let us compare the distribution of these scores. For statistical reasons, it is convenient to take the logarithm of the individual scores.

Figure 1 displays a non parametric estimation of the densities of the log scores for seven European countries[2]. We have added California as a point of reference. Out of the 12 top US universities of the Shanghai ranking, 5 are located in California while the total number of economics departments is 52 in this State. It thus can be taken as a proxy for the whole USA. We can distinguish two groups of European countries:

— A first group of countries has a large mode around zero, which means that in these countries most authors obtained only one publication over ten years. These are Spain, Italy, Germany and France. They are large countries from the south or centre of Europe.

[2] A non parametric estimate of a density is a kind of smoothed histogram which represents the frequency of the observations ordered in small classes.

Figure 1 Density of log individual publishing scores grouped by countries



Country tags are given by decreasing height of the first mode of the densities.

— A second group of countries has a marked secondary mode just below 2. This is equivalent to seven publications in ten years. These are Belgium, the UK and the Netherlands. For the latter, the second mode is dominant as well as for California. These European countries are representative of the north of Europe. We could have added Denmark, Sweden and Switzerland to this group.

The distinction between these two groups could have its roots far back in history. It can be related to the opposition between two models of universities: the Napoleonic model for southern European countries and the Anglo-Saxon model for northern European countries. The third model, the Humboldt model which used to be dominant in Germany is no longer exactly in use, at least not in its original form.

3. Universities and affiliations

Universities can be ranked according to the aggregate score of their members. But what is the definition of a university? It is not as simple as it might appear at first sight.

Definition 2 *An academic research institution is defined at time t as a collection of individuals having a research and a teaching activity. These individuals have a common physical location. They acknowledge their current affiliation in their scientific publications. They constitute the collective human capital of the institution.*

This definition considers current and not past affiliations. It makes no credit to the history of the institution. Once an author leaves his institution, he leaves it with all his publication stock. And, when a new member arrives, his new institution is credited with all his past scientific achievements, discarding the fact that he has written them elsewhere. In fact, this definition aims at measuring the current human capital of an institution. A PhD student is not interested in past Nobel Prizes, but is interested in having a PhD supervisor who is a nowadays highly cited researcher. Finally, this definition insists on common location. This means that the Tinbergen Institute in the Netherlands, the CNRS in France, the CEPR in the UK, the Max Planck institutes in Germany... are not institutions and thus cannot be ranked.

An alternative definition can be used as well which does justice to another view. The web site of a university always mentions its past Nobel Prizes, even if they are dead. This means that they still contribute to the reputation of the university and that they are the sign that a new Nobel Prize can appear in the future. A visitor usually indicates the temporary affiliation of the hosting institution on the papers he has written during his visit. When writing his report to ask for subsidies, a dean uses in fact the following definition of his institution:

Definition 3 *An academic research institution is a "moral person" having the intellectual ownership of all the present and past research hosted in its walls and financed on its funds.*

Following these two definitions, we can propose two contrasting measures of the score of an institution. Let index i covers all the n economists of a country and let index j correspond to the m year span. Let us define $\Theta_{k,t}$ as the set of members affiliated to institution k at time t. The human capital definition corresponds to

$$sd_k = \sum_{i=1}^{n} (i \in \Theta_{k,t}) \sum_{j=0}^{m} s_{i,t-j}. \tag{3}$$

while the copyright definition means

$$\widetilde{sd_k} = \sum_{i=1}^{n} \sum_{j=0}^{m} (i \in \Theta_{k,t-j}) s_{i,t-j}. \tag{4}$$

Remarks:

— Bauwens (1999) implicitly uses the legalist definition in his yearly ranking of Belgian economists and Belgian academic institutions. He took $m = 4$, but considers two periods 1992-1996 and 1993-1997. His institution rankings do not vary much, but his ranking of individuals is unstable through time.

— To our knowledge, Cribari-Neto *et al.* (1999) are the only authors to present rankings obtained according to the two definitions. But they do not interpret the economic or legal meaning of these two rankings.

— The yearly ranking produced by CentER at Tilburg University is based only on current year publications. For this ranking $m = 1$ so that the two definitions become identical.
— Information about the affiliation at the time of publication is directly given by the JEL database. It appears to be much more difficult to get the list of the members of an institution at time t. There is no simple way to reconstruct it from the data contained in the JEL database.
— Ranking the institutions of a country is equivalent to achieve the complete ordering of the set Ω_t representing all the authors of that country. Institutions have to form a partition of that set. For a given t, $\Theta_{k,t}$, $k = 1,q$ has to operate a partition of Ω_t. So for instance Oxford University and Nuffield College cannot appear in the same ranking.

4. A mathematical theory for ranking

How can we use the aggregate score of an institution to rank it? Clearly, there is a size effect so that a large institution has a higher probability of getting a higher rank than a small one. In its multi-criteria ranking the University Jiao Tong of Shanghai takes into account a size effect, but with a small weight. Here again the paradigm of a PhD student looking for a department where to apply proves to be fruitful. Let us define a minimum level of academic achievement using the van Damme formula as $z = 10 / \sqrt{2} = 7.07$. This minimal score can be reached with either 10 papers in low journals or one paper in a top journal (according to our predefined scale), all written with one co-author. A PhD student would look for a place where he can find:

1. a great number of academics that are above the minimum level z of achievement,
2. a large aggregate score, resulting from individual scores greater than z,
3. a large number of high level academics.

In Lubrano and Protopopescu (2004), we have shown, that the student decision problem led to choose the country or the institution that stochastically dominates the others at the order one, two or three, depending on his degree of risk aversion. Let us detail what we mean by stochastic dominance at the order one. We have defined a random variable which corresponds to the total score of individual authors. Let us now consider all the authors of country A, all the authors of country B and the distribution of their respective scores. We call $F_A(x)$ the cumulative distribution of the scores in country A and $F_B(x)$ the cumulative distribution of the scores in country B. In Lubrano and Protopopescu (2004), we say that country A dominates country B at the order one if whatever the value of $x > y$, we have $F_A(x) > F_B(x)$. Higher order of dominance are obtained by using integral transformations of $F(x)$.

Stochastic dominance is a notion which is difficult to manipulate. However for a given z, we can define a class of indexes (which mimics the poverty indices introduced by Foster, Greer and Thorbecke, 1984) and which can be taken as proxies for stochastic dominance. More precisely, the score of any collection of authors, such as a country, a university or a department can be measured using:

$$TS_\alpha(z) = \sum_{i=1}^{N_A} (x_i - z)^\alpha (x_i - z), \tag{5}$$

where α is a predetermined parameter that can be 0, 1 or 2. Department *A* will be said to dominate department *B* according to the index $P^{\alpha}(z)$ if $P_A^{\alpha}(z) \geq P_B^{\alpha}(z)$ for a given level z of academic achievement. If we let z vary over its starting value in IR^+, we get stochastic dominance at the order α + 1. So these indexes can be seen as particular cases for the more general notion of stochastic dominance. Let us detail some of the characteristics of theses indices, depending on the value of α.

— If α = 0, we are ranking departments according to the number of active academics having a production greater than z. This is a head count measure which is invariant to the degree of activity of productive academics, provided they produce above the minimum level z. This is a good indicator for the effective size of a department.
— If α = 1, the ranking takes into account the cumulated production of authors having a production greater than z. For $z = 0$, this is the usual measure employed in the main rankings.
— If α = 2, a larger weight is given to the most productive academics. This is in a way a more elitist criterion.

Under which conditions can we obtain the same ranking, whatever the criteria we use? There is a strong mathematical result that we shall briefly state. When there is stochastic dominance at the order one, the cumulative distributions do not cross. It is then quite easy to show that stochastic dominance at the order one implies stochastic dominance at higher orders. Consequently and only in that case, the above indexes will give the same ordering, whatever the value of α.

If scientific production is a random variable, then the measure used for ranking departments and universities is also a random variable. Consequently, rankings cannot be taken at face value. We have to provide a standard deviation for the measures $TS_{\alpha}(z)$. Are two departments which are ranked differently, really that different? A statistical test for equality of the scores can be easily devised. Let for a given α TS_A and TS_B be the total scores of two departments *A* and *B*. Let us call s_A^2 and s_B^2 their respective variances. If the central limit theorem applies, these are two Gaussian random variables and their equality is tested by the test statistic

$$t = \frac{TS_A - TS_B}{\sqrt{s_A^2 + s_B^2}} N(0,1), \tag{6}$$

which is the test statistic of the equality of the means. The 5 percent critical value is 1.96 and the 10 percent one is 1.66 for a bilateral test. A law of large numbers can be invoked for the consistency of the estimators of the variances. This test is very similar to that of Kakwani (1993) for poverty indices because our proposed indices are quite similar to the poverty indices of Foster, Greer and Thorbecke (1984).

5. Ranking European economics departments

Let us now present a first ranking using the total score measure, $TS_{\alpha}(z)$ with α = 1 and $z = 7.07$. Table 2 was computed with the data of Lubrano *et al.* (2003), with two modifications concerning France. Paris School of Economics is a new

creation formed with regrouping three research centers in Paris: DELTA, CERAS and CEPREMAP. Paris Grandes Ecoles is a planned grouping of also three institutions: HEC, Polytechnique and CREST-INSEE. More accurate information should be gained by updating the database to more recent years. This is planned for future work.

A. Ranking using alternative criteria

Table 2 European rankings using three different measures

Institution	$TS_1(z)$	*std. dev.*	*Rk*	$TS_0(z)$	*Rk*	$\sqrt{TS_2(z)}$	*Rk*
LSE	2637.33	(256.97)	1	150	1	334.81	4
U Tilburg	2433.20	(331.86)	2	108	3	336.32	3
U Oxford	2074.31	(181.48)	3	119	2	262.28	8
Paris Sc Eco	2003.02	(291.80)	4	64	12	383.24	1
U Cambridge	1919.50	(200.46)	5	101	4	276.24	7
U Erasmus	1692.40	(115.06)	6	92	5	278.20	6
U Louvain	1611.94	(257.60)	7	73	8	281.56	5
U Amsterdam	1435.42	(183.19)	8	68	11	240.18	11
U Warwick	1378.41	(124.10)	9	70	10	205.33	19
U Toulouse	1331.90	(306.18)	10	43	24	366.06	2
Paris Gr. Ecoles	1312.94	(157.15)	11	76	7	216.99	13
U Paris I	1229.64	(117.92)	12	79	6	196.33	21
U College London	1224.10	(204.36)	13	62	13	256.03	9
U Nottingham	1169.51	(163.70)	14	43	25	240.59	10
U York	1102.71	(144.75)	15	53	17	208.50	17
Stockholm Sc Eco	1066.09	(145.35)	16	56	16	206.74	18
U Maastricht	1064.94	(226.13)	17	60	14	196.68	20
U Essex	988.45	(143.02)	18	37	32	214.86	16
U Stockholm	935.42	(111.78)	19	50	19	220.54	12
U Autonoma Barc	932.92	(193.32)	20	46	22	179.62	23
U Bonn	900.15	(226.65)	21	57	15	147.96	28
London Bus Sc	883.44	(100.50)	22	53	18	156.70	25
Free U of Amsterdam	852.47	(146.16)	23	47	21	190.53	22
U Manchester	844.95	(60.38)	24	72	9	115.87	30
U Libre de Brx	844.28	(156.34)	25	33	33	215.83	14
U Copenhagen	824.28	(151.62)	26	41	27	176.19	24
KU Leuven	800.26	(140.50)	27	41	28	152.48	27
U Groningen	780.62	(98.24)	28	44	23	154.84	26
U Aix-Marseille	752.47	(155.06)	29	29	35	214.99	15
U Pompeu Fabra	744.36	(103.20)	30	40	30	147.69	29

Column 2 is used to rank departments according to the total production. The corresponding rank is given in column 4 while column 3 gives the standard deviation of the total score. Column 5 represents the number of active members and produces the ranking given in column 6. Column 7 represents the square root of $TS_2(z)$ and produces the ranking given in column 7.

Table 2 presents three different rankings for the thirty main economics departments in Europe. The first ranking is the most usual one, because it is based on the

total score of a department. However, in this total we have included only the production of members above a personal minimum score z. That minimum level can have been gained outside their present institution. The ranking obtained in column 4 (and also given in column 1) according to this criterion is in accordance with common intuition[3].

The notion of a minimum level of activity starts to enter the criteria used by governmental evaluation agencies. For instance, the French AERES (Agence d'Evaluation de la Recherche et de l'Enseignement Supérieur) defines a criterion for identifying publishing scientists: a publishing scientist in the social sciences is a person publishing at least four articles in refereed journals over a period of four years; that number can be modulated by the quality of the journals and the number of co-authors. This definition is very similar to our z and is in a way even stricter. The French AERES uses the total number of publishing scientists to characterise a research center. In columns 5 and 6 in Table 2, we have used the headcount measure TS_0 to obtain an alternate ranking of the departments. This criteria has a tendency to favour big departments. For instance Paris I in France, Manchester University in the UK, or even Bonn University in Germany are favoured by this criterion.

The last criterion $\sqrt{TS_2(z)}$, which puts a stronger weight on very productive authors, has a strictly reverse effect. It favours very much smaller departments which are composed of very productive researchers because it takes the sum of their squared production. This is the case for PSE and even more TSE in France. It is an elitist criterion.

How should we interpret these differences in ranking? For some institutions, the rankings are roughly the same, whatever the method. For other institutions, there are huge differences. In fact, it is useful to go back to the notion of stochastic dominance that we have introduced above. We have chosen a fixed level z and said that A dominated B if $TS^{A}_{\alpha}(z) > TS^{B}_{\alpha}(z)$. We have stochastic dominance at the order α if this relation is valid whatever the value of z. If we have stochastic dominance at the order one ($\alpha = 0$), then department A will dominate department B at any higher orders. In this case, the choice of the criterion does not matter. Stochastic dominance at the order one is obtained when cumulative distributions do not cross. If distributions do cross, then we must use higher orders of stochastic dominance to compare A and B. When our indices produce different rankings, this simply means that we do not have stochastic dominance at the order one and consequently that there does not exist an unambiguous ranking. Let us now try to determine for which departments there are strong differences.

To be more precise, we define three categories by observing the gain or loss in ranking between TS_2 and TS_0. A difference of 5 is considered as non significant. Let us identify and characterise the winners and losers.

Some large and well established departments are not very much affected by the type of criteria used for their ranking. From the top of the list to its bottom, fall in this category the LSE, Tilburg, Cambridge, Erasmus, Louvain, Amsterdam, University

[3] Tilburg is a university specialised in the social sciences. Its economics department had these last twenty years a very active recruiting policy which explains its position in this ranking.

College London, York, Stockholm School of Economics, Barcelona, The Free University of Amsterdam, Copenhagen, Katholieke Universiteit Leuven, Groningen and Pompeu Fabra. We can say that these departments are fairly homogenous and stable. Their position cannot be deeply modified by the arrival or departure of a member. This is an indication of stochastic ordering at the order one.

Some departments owe their position in this ranking mainly to their size. They have a significant number of active researchers, but these researchers are not outstanding. They suffer a lot if we adopt a more elitist criteria such as TS_2. In this category fall surprisingly Oxford and Warwick in the UK, but also Paris Grandes Ecoles and Paris I in France, and Maastricht, Manchester, Bonn, London Business School, lower down in the ranking. These departments should make a significant effort in the quality of their recruitment. We have to say again that this result is surprising for Oxford and Warwick.

The final group of institutions made a real effort on the quality of their recruitment. The best example is Paris School of Economics which is in the top five, but manages to be first in Europe if we follow TS_2. Even more striking is Toulouse which is only in the top ten, but comes second with TS_2. But with this last case, we could say that it lacks a scientific basis formed by a bulk of intermediate researchers. The department seems too small to be solid in the long term. With TS_0, Toulouse would be 24th. The departure of some of its most brilliant members could be very damaging. In the same list, but lower down in the ranking we have Nottingham, Essex, Stockholm University, Université Libre de Bruxelles and Aix-Marseille. In these universities, the quality is based on a small group of top people.

B. Testing for differences

Let us go back to the first ranking, the one based on the index TS_1. We give in column 3 of Table 2 standard deviations that can be used to compute the test of the equality of the means defined in formula (6). With this statistics, we can test whether two departments with different rankings are in fact not statistically different. Under a normality assumption, the 5 % critical value of that equality test is 1.96. We shall say that two departments are statistically different if the associated test statistics is greater than that critical value. With 30 departments, we can compute 435 different statistics and build a 30 × 30 table. This would not be very illuminating. We are thus not going to give this full table, but will try to produce some salient features that can be drawn from examining it. We proceed as follows. We have ordered the departments according to their TS_1. We start from the department which is ranked first and test if its immediate follower is statistically different. If it is not, we define a first group with these two departments, saying that they are equivalent. We go down the list till we reach the first department which is statistically different from the top department of the list. This department will constitute the starting point of a second group and so on. This is a very heuristic procedure which is just a convenient mean of summarising a large table. In each group, all departments are equivalent. Between two consecutive groups, the first departments of each group are statistically different, but in general, the last department of the first group is not statistically different from the first department of the second group.

At the top of the list, we find four leading departments which can be considered as equivalent: LSE, Tilburg, Oxford and PSE. This means that LSE cannot be distinguished from the three other ones, but is different from the top department of the second group, Cambridge. Starting now from Cambridge, we find a second group of four departments: Cambridge, Erasmus, Louvain and University of Amsterdam. This means again that the leader Cambridge cannot be distinguished from its three followers. But this does not imply that any member of this group is different from the first group. For instance, Oxford and Cambridge are not statistically different. The third group is larger with 9 universities: Warwick, Toulouse, Paris Grandes Ecoles, Paris I, University College London, Nottingham, York, Stockholm School of Economics and Maastricht. Finally, if we start now from the bottom of the ranking, we must go up to rank 15 in order to find a department which is statistically different from the last one of the ranking. There is thus a large group of equivalent departments, between Stockholm School of Economics and University Pompeu Fabra. That means that from rank 16 till rank 30, no department is really different from the other when they are ranked according to TS_1. On average, these departments have 47 active members with a minimum of 29 and a maximum of 72. This is to be contrasted with the top 15 group which has on average a size of 80 active members with a maximum of 150 and a minimum of 43. Departments in the top group are much bigger departments.

Let us now try to detail the position of the five French departments that appear in this ranking of thirty European departments. These departments have very different characteristics. One is huge: Paris I; one is very small: Aix-Marseille. Three are departments inside a public university: Toulouse, Paris I and Aix-Marseille, while the remaining two (PSE and PGE) are members of the selective system of *Grandes Ecoles*. Toulouse has such a large standard deviation that it cannot be statistically different from any of the other four departments. This is in a way a special case. On the contrary, PSE is equivalent to Toulouse, but dominates the four other departments. PGE and Paris I are equivalent despite the fact that they are organised in a totally different way. However, they both dominate Aix-Marseille.

6. Some elements for explaining the rankings

Up to now, we have given details on a methodology for ranking economics departments. We have shown that the global Shanghai ranking was in fact based on a too narrow information set. Turning to economics departments, we had access to a much larger data base. However, even with that data base, there remains a large uncertainty on the produced rankings. Nevertheless our rankings and the Shanghai ranking have a common point: European departments do not compare favourably to the US departments (see Lubrano *et al.* (2003) for more details).

In this section, we shall find some heuristic explanations to the poor rating of European institutions. Some of the data we shall exploit concern only economics departments, some other data concern entire universities, because usually these figures are either not available or not significant at the department level.

The difference in quality between universities is the result of a long historical process, of differences in scientific habits and of different institutional contexts. A ranking is interesting if it can be matched with stylised facts that could provide at

least an intuition for devising economic policy measures that would improve the position of Europe in the world knowledge economy and within Europe the position of countries that have a weak system of research.

A. A heterogeneous output

The rankings we gave in the previous section are based on publications in economics journals. To explain differences in ranking, the first intuition consists in analyzing which types of publications are practiced in the different countries, both from a quantitative and a qualitative point of view, and that at the country level.

A first simple calculation consists in dividing the number of papers published in a given country by its total population. This gives an idea on research intensity or productivity per country. The second type of analysis looks at the proportion of national journals in the total production of a country and compares it with the percentage of the same production that appears in top international journals. These figures are given in Table 4 and Table 6 of Lubrano *et al.* (2003).

In Europe, the most research intensive countries are small northern European countries plus the UK. They include Belgium, Denmark, Finland, the Netherlands, Norway and Sweden. In that group of countries, the number of published papers in economics per million inhabitants is greater than 75 and goes up to 115 for the UK. California has a similar ratio of publication as the UK. On the contrary, large countries from southern Europe have a low research intensity: France, Germany, Italy and Spain have published on average less than 46 papers per million inhabitants. These figures are in accordance with the estimated densities given in 1. It would be hard to pretend that these differences appear just because of a selection bias. The JEL data base that we used provides a large coverage for national journals, including France and the other southern European countries. This would not be the case, had we used for instance other data bases such as the *Social Science Citation Index*.

This contrast between the two groups of countries is amplified by the fact that these large countries publish between 65 % and 85 % of their production in national low ranked journals. This proportion is on average 30 % for the small northern European countries and goes down to 8 % for the Netherlands. The group of small northern countries publishes more than 30 % of their total production in top international journals[4], while for the group of four large countries, this percentage is below 15 %. The UK is a particular case. It uses national journals for 40 % of its total production, but these journals are in general of a good quality. However, it uses top quality journals for only 20 % while this percentage soars up to 60 % for California. As underlined in Drèze and Estevan (2007), Europe has to put a huge effort on the four large countries if it wants to reach the Lisbon objectives[5]. This is also coherent with the

[4] Top international journals are defined in Lubrano *et al.* (2003). as journals which get a grade between 6 and 10. In our list of 650 journals, there are 70 such journals. Thus this list is not too elitist.

[5] The Lisbon objectives aim to make the European Union "the most dynamic and competitive knowledge-based economy in the world", through education, research and development (quoted from Drèze and Estevan, 2007). To reach this goal, one objective is to reach a share of 3 % of GDP for research and development spending.

econometric findings of Bauwens *et al*. (2007) where the dummy variable English language in publications is found to be an indicator of research performance.

B. Institutional data

Table 3 Miscellaneous data for some European universities (I)

University	*Creation date*	*Students*	*Academic staff*	*Total staff*	*Nobel Prizes*	*Annual budget*
Amsterdam	1632	27,000	2,100	5,000		487
Erasmus	1913	20,000	1,925	3,700	1	461
Tilburg	1927	11,500	1,300			117
Louvain	1425	21,000	1,334	5,000	1	360
ULB	1834	20,000	1,300		4	226
Oxford	1167	20,000	1,465	8,419	48	900
Cambridge	1209	18,000	1,600	8,600	83	900
Warwick	1965	30,000	1,800	5,000		465
LSE	1895	7,800	1,460		14	66
Toulouse I	1229	17,000	568	1,049		64
Paris I	1150	40,000	1,611	2,464		54

The annual budget is expressed in millions of euros. It includes wages of the academic staff. For French universities, the official figures do not include academic wages which are paid directly by the State. For the LSE, the annual budget does not include academic wages, which are paid directly by the University of London.

In most European countries, research is produced inside universities[6]. The way these institutions are organised and financed could explain some of the differences in output. So it is interesting to produce some figures and to contrast a panel of European universities. We have regrouped these figures in Tables 3 and 4. We have chosen eleven universities which are representative of four different countries: the UK, the Netherlands, Belgium and France. We have regrouped universities by countries. The last category regroups universities in the social sciences for which budget figures do not cover wages. These figures were collected on the Web, either on Wikipedia or directly on the Web sites of the universities.

The first striking fact is that universities are most of the time very old institutions created during the Middle Ages. The University of Amsterdam was created in the seventeenth century because the Netherlands got their independence from Spain only in 1581. A second generation of universities was created at the end of nineteenth century and during the twentieth century.

Financial figures are very difficult to interpret because they do not cover the same implications. For that reason, we separated LSE, Toulouse I and Paris I because academic wages are not included in their budgets. Apart from these three cases, the

[6] There are also national research agencies. We have not taken them into account. A specialised ranking can be found on the web www.webometrics.info. However, it must also be noted that members of these agencies can be working inside universities like for instance most of the CNRS researchers in economics.

indicated budgets cover wages and daily functioning. Endowments are not included and can be huge as for LSE, Oxford and Cambridge. The average budget is around 450 million of euros per year for a standard European university, except of course if that university is specialised in the social sciences. The immediate striking fact is that the budget of Oxford and Cambridge is twice that figure. This should perhaps be confronted to the 131 Nobel Prizes that these two universities collect among the 151 European Nobel Prizes reported here. Top scientific research is an expensive activity.

The usual size of a university is around 20,000 students. LSE is a special case because it is not a university, but simply a department of the University of London. French universities are within that range, except for Paris I which has twice this average size. Italian universities can be much larger. Università di Roma La Sapienza has 147,000 students.

Table 4 Miscellaneous data for some European universities (II)

University	*Tuition fees*	*Admission undergrad.*	*Budget per student*	*Academics per student*	*Total staff per student*
Amsterdam	1,538	free	18,037	0.08	0.19
Erasmus	1,538	free	23,050	0.10	0.19
Tilburg	1,538	free	10,174	0.11	
Louvain	788	free	17,143	0.06	0.24
ULB	788	free	11,300	0.07	
Oxford	4,860	AAA	45,000	0.07	0.43
Cambridge	4,800	AAA	50,000	0.09	0.48
Warwick	4,500	mild	15,500	0.06	0.17
LSE	4,500	strong	8,462	0.19	
Toulouse I	544	free	3,764	0.03	0.06
Paris I	544	free	1,350	0.04	0.06

AAA refers to grades obtained at GCSE for UK students. All students applying to a UK university must go through UCAS, which is a centralised service for Universities and Colleges Admissions, www.ucas.com. Tuition fees are given for undergraduate programs and European students.

Admission fees can vary a lot across countries, but are uniform inside countries. The given figures correspond to undergraduate studies. There are virtually no fees in Germany. French fees are very low. Fees in the UK are pretty high compared to the other European countries. There is no selection at the entrance of undergraduate studies in most countries, except in the UK and in Germany. Oxford and Cambridge require AAA grades to the GCSE. But there is no running examination like in the French *Grandes Ecoles*. There used to be one in Cambridge, but it was suppressed. There is always some kind of selection at the entrance of PhD studies in every country. As a conclusion, we have two groups of universities. British universities form the first group. They impose high fees and a selection at the entrance. The degree of selection depends on the ranking of the university; but it is always possible to find a university corresponding to one's qualification. The second group is formed by the continental universities where there is no selection at the entrance and where fees are moderate or low.

Let us finally examine financial figures and staff. The average budget per student is around 15,000 euros either in Warwick or in continental Europe. In Oxford and Cambridge it is thrice that figure. The last three lines of Table 4 show the dramatic financial situation of the French universities. On average in France, the cost of a student is 6,850 euros, while the cost of a student in *Classes Préparatoires aux Grandes Ecoles* is 13,200 euros. We find the same type of disparity when we look at the academic staff per student. The average number is 0.10 including Oxford and Cambridge, when it goes down to 0.04 in France. For total staff, the contrast is more pronounced. Oxford and Cambridge are twice above the European level while France is three times below the European level.

C. Confronting output and institutional settings

It is rather difficult to relate the above figures describing the organisation of European Universities to their ranking. Aghion *et al.* (2007) (reprinted in this volume) have tried to relate country performance in the Shanghai ranking to some of these key variables, using regression analysis. They also have variables describing university governance. Bauwens *et al.* (2007) use a Poisson model which explains the number of highly cited researchers in a country by a set of variables such as English proficiency and RD funding. Both papers work at the country level. However, the final regression in Aghion *et al.* (2007) concerns universities. How could we confront our heterogenous output data to some of the elements of organisation that we have put forward. Let us first try an informal reading grid before presenting in the next section what can be done with multilevel models.

We already alluded to the three existing organisational models for universities. The Napoleonic model is based on large public universities which deliver national diplomas, having roughly no fees and no selection at the entrance. This model is valid in three large southern European countries (France, Italy and partially in Spain). The analysis we provided for the output of these countries could be taken as a clue for the lack of success for this model. The various experiences in France show that this model is hard to reform.

Germany was for a long time ruled by the Humboldt model in which small groups of students were having periodic seminars with professors and where universities are autonomous (contrary to the Napoleonic model). This model could account nowadays for the functioning of the EHESS in France, but certainly not for the present organisation of the German university which teaches a large number of students. It is slightly selective and has no admission fees. In term of research, this model did not prove to be efficient, at least for economics.

The last identified model is the Anglo-Saxon model which concerns the UK and the USA. It is characterised by autonomous and competitive universities, important admission fees and a strong selection at the entrance. It would be difficult to pretend that this model is fully efficient simply because of the large differences in output that we identified between the UK and California. But it is certainly more efficient than the Napoleonic model.

The good results of the small nordic countries (Belgian, the Netherlands in our tables, but also Denmark, Norway and Sweden) are not explained by any of the three

above models. In these countries, universities might have a large degree of autonomy (like in the Anglo-Saxon model), but there is roughly no selection at the entrance and fees are moderate (unlike the Anglo-Saxon model). They receive large subsidies from their government. These countries could provide good suggestions to reform the Napoleonic model of France, Italy and Spain. This intuition has to be formalised in a detailed statistical model.

7. A statistical model for explaining academic performance

The bibliometric data that serve to measure the performance of a department have a rather complex structure which has to be taken into account in order to model them correctly. Multilevel models are the appropriate tool for that purpose. We cannot give here a complete description of all the results we already obtained, because our research is still ongoing. We will just describe the first results and show how multilevel models can be used for ranking and where all the major institutional variables that we have detailed can enter.

A. Multilevel data

In order to find explanations to academic performance, we must first precisely describe the different levels according to which these data can be organised:

1. Individual researchers and professors are publishing articles over a given period. We have data on those publications which are the main ingredient to measure research output.
2. We have departments and universities where the authors are affiliated. They constitute their first environment. They are the object of rankings.
3. Finally we have countries with specific policy indicators such as total spending in higher education as a percentage of GDP.

It is quite illuminating to distinguish between these three different levels, because they are the source of different types of variables that can influence the performance of a department. We shall try to list these variables.

— Individual variables are personal characteristics that can influence the scientific production at a given period. There are basic variables such as age and gender. Age is interesting if we want to point out a life cycle effect as in Rauber and Ursprung (2006). Other variables concern the place where the PhD was delivered (abroad or not), past production, and recognition by peers (being or not being elected fellow of the econometric society). Concerning individual production, the percentage of papers written with foreign co-authors, the number of foreign co-authors, the percentage of papers published in national journals are indicators that can be drawn directly from bibliographic data bases.

— Institutional variables at the university level cover input and environmental variables which can be listed as follows: The budget per student, the number of students, the presence of a selective system at the entrance, the amount of fees, the importance of non academic staff, the way of governance (autonomy).

— Variables at the country level might be thought of having a large impact in term of economic policy. Some countries have a large number of economics departments and other have only very few. In some countries most authors publish in national journals in their national language. For some other countries, it is just the reverse. The use of English as a scientific vehicle was shown to have a tremendous impact in Bauwens *et al.* (2007). Finally the percentage of GDP devoted to research is thought to be a major indicator. There is the Lisbon objective of 3 % of GDP which is far from being reached by most of the European countries.

The correlation structure between these variables depends heavily on the level at which they are observed. Multilevel models take full account of this correlation structure.

B. Multilevel models with exogenous variables

Multilevel models have for long been used to rank secondary schools and hospitals as explained in Goldstein and Spiegelhalter (1996), but never for ranking universities. The basic model can be presented as follows. Let y_{ij} be the score of author i belonging to institution j:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_{0j} + e_{ij} \qquad e_{ij} \approx N(0,\sigma_e^2) \\ \beta_{0j} &= \qquad\quad u_j \qquad u_j \approx N(0,\sigma_u^2) \end{aligned} \tag{7}$$

This is the most simple two level model we can think of. The first equation describes an author production as the sum of three factors: An overall mean, β_0, generally called the grand mean, a specific effect which depends on the institution of affiliation, β_{0j} and an individual random term of zero mean and variance σ_e^2. Parameter β_{0j} represents the deviation of the mean score of each institution from the grand mean. The second equation says that this institution effect is a random effect which has a zero mean and a variance equal to σ_u^2. There are thus three parameters in the model: β_0, σ_e^2 and σ_u^2. This version of the model is called a variance component model as it allows to separate the total variance of individual scores between a pure individual effect and an institutional effect.

This model is a necessary starting point for further analysis. Once it is estimated[7], the game consists in trying to reduce the two variances by adding exogenous variables. Let us call x_{ij} the duration of activity of an author for instance. This individual variable can be added in the following way:

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_1 x_{ij} + e_{ij} \qquad e_{ij} \approx N(0,\sigma_e^2) \\ \beta_{0j} &= \qquad \beta_0 + u_j \qquad u_j \approx N(0,\sigma_u^2). \end{aligned} \tag{8}$$

[7] Details for estimating those models can be found for instance in Goldstein (2003) or Raudenbush and Bryk (2002). The procedure MIXED of SAS was used to estimate the multilevel model of the next subsection. SPSS or MLWIN are also feasible options.

Just for convenience, we have moved the grand mean β_0 to the second equation. We have imposed that the influence of individual experience x_{ij} is the same across institutions. But of course a more general model can be proposed where β_1 varies randomly across institutions. This would imply the following model:

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_1 \quad x_{ij} + e_{ij} \\ \beta_{0j} &= \qquad\quad \beta_0 + u_{0j} \\ \beta_{1j} &= \qquad\quad \beta_1 + u_{1j} \end{aligned} \tag{9}$$

The second level is now described by two equations. We might like to introduce at this second level specific exogenous variables which could explain the characteristics of the institutions, such as for instance the number of students or total funding. We give here only this brief sketch just in order to underline the many possibilities of multilevel models and also to justify that we have not enough room here to provide a full analysis which is, by the way, still under investigation.

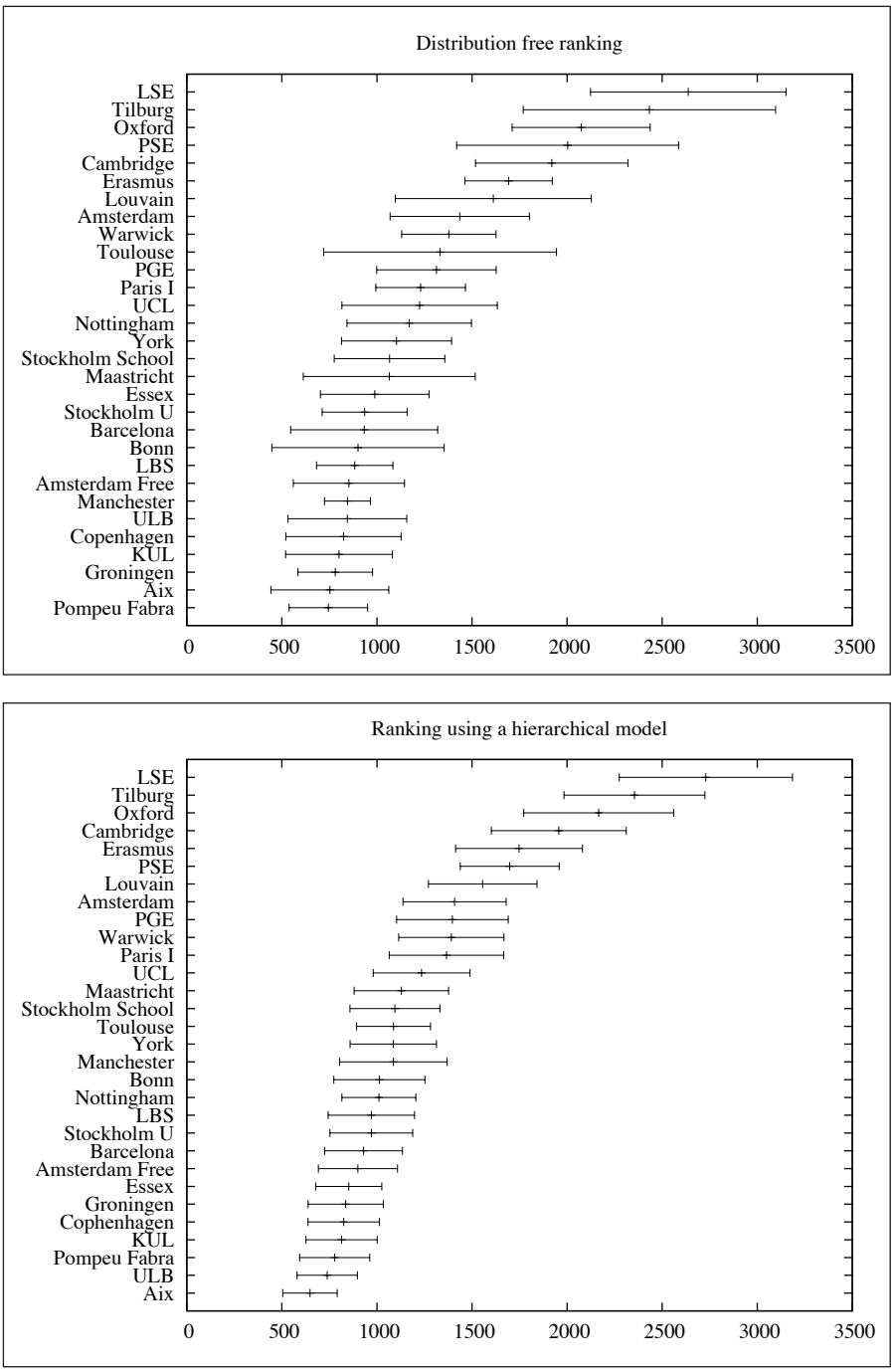
C. Ranking with multilevel models

Let us go back to the simple variance component model (7), where no exogenous variables are introduced. The u_j are the estimated random effects, which are obtained as a by-product of estimation. These random terms measure the deviance of each institution or department with respect to the grand mean. When added to β_0, they represent the mean score of each institution, providing thus a possible basis for rankings. This is the methodology used for instance in Goldstein and Spiegelhalter (1996) for ranking secondary schools. A standard deviation can be computed from σ_u^2, so that the final ranking is given with a confidence interval.

A mean score, such as a percentage of success to a national exam, is a meaningful value to rank secondary schools. In the rankings of universities that we detailed above, a size effect is taken into account. The more brilliant academics there are, the better it is for a university. An average score is not meaningful for ranking universities. Consequently $u_j \times u_j$ is the useful quantity to consider, where n_j is the number of authors in institution j. We have estimated the simple variance component model on our sample, assuming that e_{ij} and u_j were both normally distributed. It is interesting to confront the new ranking obtained to the one we initially obtained with a method that can be qualified as distribution free. In Figure 2, we present two error graphs (similar to those presented in Goldstein and Spiegelhalter 1996) so as to compare the two methods.

On the vertical axis, we have the institution names. On the horizontal axis, we have the predicted total score of the department together with a 95% confidence interval. The two methods give quite comparable rankings. The Spearman rank correlation between the two rankings is equal to 0.95. However, confidence intervals are narrower with the multilevel model. This apparent gain in efficiency is obtained at the expense of giving a lower rank to institutions that have a larger standard deviation, such as PSE and Toulouse and a better rank to those which have a smaller

Figure 2 Comparing two statistical methods for ranking institutions



standard deviation like Manchester. As we said above, variance component models are just a first step of analysis and exogenous variables have to be introduced.

8. Conclusion

One of the main points that we have tried to make in this paper is that scientific production is a random variable and that it has to be treated as such. This has a first consequence on rankings. Random variables have standard deviations. Whatever the criterion used for ranking institutions, a standard deviation has to be taken into account and this means that rankings are not deterministic. Two or more institutions can be statistically undistinguishable. The second point is that a standard deviation can be diminished by increasing the number of observations. In order to rank departments properly, we have to use criteria for which there are many observations. If the criterion used is too elitist, for example counting the number of Nobel Prizes, we simply are not going to have enough observations. The Shanghai ranking is subject to this type of criticism.

Ranking might be useful for distributing funds for instance. But, we definitely need an explanation about the average bad ranking of European institutions. We need a model relating scientific production to certain key variables. Some of these variables are individual variables, some are institutional variables and finally there are policy variables at the country level, such as those implied by the Lisbon agreement. We have given indications on how multilevel models could provide such a framework. Preliminary results, contained in Chandan, Goudard and Lubrano (2008), indicate that both personal publication habits and national variables such as total spending per student or the number of economics departments are of prime importance. More work is under way to expand these first preliminary results.

References

Aghion P., M. Dewatripont, C. Hoxby, A. Mas-Colell and A. Sapir (2007), "Why Reform Europe's Universities?", Bruegel Policy Brief 2007/04, Bruegel, Brussels.

Bauwens L. (1999), "Economic Research in Belgian Universities", Mimeo, CORE, Université catholique de Louvain.

Bauwens L., G. Mion and J. Thisse (2007), "The resistible decline of European science", CORE-DP 2007-92, CORE, Université catholique de Louvain.

Chandan A., M. Goudard and M. Lubrano (2008), "Using hierarchical linear models for explaining the ranking of universities", Manuscript, GREQAM, Marseille.

Combes P. and L. Linnemer (2001), "La publication d'articles de recherche en économie en France", *Annales d'Economie et de Statistique*, 62, pp. 5-47.

Cribari-Neto F., M. Jensen and A. Novo (1999), "Research in Econometric Theory: Quantitative and Qualitative Productivity Rankings", *Econometric Theory*, 15(5), pp. 719-752.

Drèze J. and F. Estevan (2007), "Research and higher education in economics: can we deliver the Lisbon objectives?", *Journal of the European Economic Association*, 5(2-3), pp. 271-304.

Foster J., J. Greer and E. Thorbecke (1984), "A class of decomposable poverty measures", *Econometrica*, 52(3), pp. 761-766.

Goldstein H. (2003), *Multilevel Statistical Models*, Oxford University Press, Oxford, 3rd ed.

Goldstein H. and D. Spiegelhalter (1996), "League Tables and their Limitations: Statistical Issues in Comparisons of Institutional Performance", *Journal of the Royal Statistical Society*, Series A, 159(3), pp. 385-443.

Kakwani, N. (1993): "Statistical inference in the measurement of poverty," *Review of Economics and Statistics*, 75(4), pp. 632-639.

LUBRANO M., L. BAUWENS, A. KIRMAN and C. PROTOPOPESCU (2003), "Ranking European economics departments: a statistical approach", *Journal of the European Economic Association*, 1(6), pp. 1367-1401.

LUBRANO, M. and C. PROTOPOPESCU (2004), "Density inference for ranking European research systems in the field of economics", *Journal of Econometrics*, 123(2), pp. 345-369.

RAUBER M. and H. URSPRUNG (2006), "Life cycle and cohort productivity in economic research: the continental European experience as exemplified by the case of Germany", Mimeo, Department of Economics, University of Konstanz.

RAUDENBUSH S. W. and A. S. BRYK (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods*, Sage Publications Ltd, London, 2nd ed.

VAN DAMME E. (1996), "Measuring Quality of academic journals and scientific productivity of researchers", Unpublished mimeo, CenTer, Tilburg University.

Why reform Europe's universities?

Philippe Aghion, Mathias Dewatripont, Caroline Hoxby,
Andreu Mas-Colell and André Sapir [1]

Summary

Recently published international rankings indicate that the performance gap between European and American universities is large and, in particular, that the best European universities lag far behind the best American universities. The country performance index we construct using the Shanghai ranking confirms that despite the good performance of some countries, Europe as a whole trails the US by a wide margin. The reason for this situation, which contributes to Europe's lagging growth performance, is two-fold. First, Europe invests too little in higher education. Total public and private spending on higher education in EU25 accounts for barely 1.3 % of GDP, against 3.3 % in the US. This translates into average spending of less than €10,000 per student in EU25 versus more than €35,000 in the US. Second, European universities suffer from poor governance, insufficient autonomy and often perverse incentives. We show that both factors contribute to the EU's poor performance and that reform should take place on both fronts, because autonomy also increases the efficiency of spending.

1. Introduction

European growth has been disappointing for the last 30 years but policymakers have only recently started to realize that Europe's growth performance is intimately linked with the research performance of its universities.

Europe invests too little in higher education. It is by now widely known that the European Union (EU) spends less than two percent of its GDP on R&D, compared to more than 2.5 percent in the United States (US). But the gap between Europe and the US is even wider for universities than for R&D spending. In 2001, total (public and private) spending on higher education in EU25 accounted for barely 1.3 % of GDP, against 3.3 % in the US. In other words, Europe spends every year two percent

[1] This article was published as a *Bruegel Policy Brief*, issue 2007/04, September 2007, © Bruegel – www.bruegel.org. We are very grateful to Aida Caldera, Indhira Santos and Alexis Walckiers for their excellent research assistance, and to colleagues across European universities in helping with the university survey used in this policy brief.

of GDP less than the US. In terms of expenditure per student, the contrast is starker still, with an annual spending of €8,700 in EU25 versus €36,500 in the US.

But the unsatisfactory research performance of Europe's universities also results from inadequate institutions. European universities suffer from poor governance, insufficient autonomy and often perverse incentives.

Europe started to recognize some years ago that its university system faces a problem. A first step was the Bologna Declaration that initiated the creation of a "European Higher Education Area". Recently, a growing number of individual EU member states have introduced reforms of their university systems.

However only the recent publication of global rankings, such as the Shanghai Jiao Tong University Academic Ranking of World Universities (the "Shanghai ranking") has made most policymakers aware of the magnitude of the problem and sparked a public debate on university reform. These rankings tend to reinforce the evidence that the US is well ahead of Europe in terms of cutting-edge university research.

The purpose of this Policy Brief is to examine what reforms are needed in order to enable European universities to produce world-class research and thus make the optimum contribution to growth[2].

In the first section of this Brief, we draw conclusions from the Shanghai ranking both about European university research performance in relation to that of US institutions and about differences in performance between European countries. We then report on our own survey of European universities listed in the Shanghai ranking, which we use to establish what determines university research performance. We also use comprehensive US data to analyse the interplay between autonomy and funding in boosting university research performance. Finally, we make concrete proposals about how to improve the conditions for research at European universities with the objective of boosting their contribution to growth.

2. Country performance

The debate on the funding and governance of European universities has been stirred greatly by the publication, since 2003, of the so-called Shanghai index which measures university research performance. Constructed by a group of Chinese scholars, the Shanghai index is a weighted average of six different indicators (see Box 1). While the weights are admittedly somewhat arbitrary, the main advantage of the index is its reliance on publicly available information.

Table 1 presents a detailed account of relative country performance, looking successively at the Top 50, Top 100, Top 200 and Top 500 universities in the Shanghai ranking. To better see how to read this table, consider first the column "Top 50". The best university in the Top 50 is given a score of 50, the next best university is given grade 49, and so on down to a score of 1 for the least performing university within the Top 50. For each country (or region), we then compute the sum of Top 50 Shanghai rankings that belong to this country, and divide the sum by the country's population. Finally, all the country scores are divided by the US score, so that each entry in the

[2] This Policy Brief does not deal with all the various roles and functions of universities, solely their research function. An upcoming Bruegel Blueprint will provide a fuller analysis of how universities perform against a broader set of objectives. Furthermore, this Policy Brief does not discuss the potential of EU-level policy to add value. This will also be dealt with in the upcoming Blueprint.

Box 1 The Shanghai index

This index aggregates six different indicators of research performance:

— The number of alumni from the university winning Nobel Prizes in physics, chemistry, medicine, and economics and Fields Medals in mathematics
— The number of university faculty winning Nobel Prizes in physics, chemistry, medicine, and economics and Fields Medals in mathematics
— The number of articles (co-)authored by a university faculty published in *Nature* and *Science*
— The number of articles (co-)authored by a university faculty published in Science Citation Index-expanded and Social Science Citation Index
— The number of highly cited researchers from the university in 21 broad subject categories
— The academic performance with respect to the size of the university.

Note that the Shanghai index tends to undervalue countries where a great deal of academic scientific research takes place outside universities (the Max Planck institutes in Germany) or in centres whose researchers are affiliated with several universities (the CNRS laboratories in France). This partly explains the poor performance of France and Germany in Table 1.

column "Top 50" can be interpreted as a fraction of the US per capita performance for the Top 50 universities. This gives our Country Performance Index for the Top 50 universities. The same logic applies, respectively, to the "Top 100", "Top 200" and "Top 500" columns, where the best university receives a score of, respectively 100, 200 and 500, and the last one always receives a score of 1. There are, obviously, fewer zero entries in a column as one moves from the Top 50 to the Top 500 as it is easier for a country to have universities among the latter than the former.

Table 1 reveals several interesting findings:

— First, the United States completely dominates all European countries in the Top 50 universities. Only Switzerland and the United Kingdom rival the US on a per capita basis. By contrast, the EU15 and EU25, with a greater population than the US, score much lower.
— Second, the top 4 US states (Massachusetts, California, New York and Pennsylvania) score better than any European state in the Top 50 and Top 100.
— Third, country performance becomes more equalized as one enlarges the number of universities considered. In particular the gap between the EU15 or the EU25 and the US narrows down as one moves from the Top 50 to the Top 500. In part this is due to the way the scores are constructed, but it mostly reflects a reality: American universities dominate European universities in the top tier (the Top 50 and Top 100), but Europe has many good universities in the second (the next 100) and the third (the next 300) tiers.
— Fourth, there are important differences among European countries: Switzerland, the UK and Sweden do particularly well, even in the Top 100, where they out-perform (Switzerland and Sweden) or almost match (the UK) the United States on a per capita basis. The rest of Scandinavia (Denmark and Finland), Belgium and the Netherlands also do pretty well in the Top 200 and Top 500. By contrast, Southern and Eastern Europe lag far behind. France and Germany do relatively poorly, except in the third tier, the universities ranked between 301 and 500.

Table 1 Country performance in the Shanghai ranking (measured as percentages of the US per capita performance)

Country	*Population (in million)*	*Top 50*	*Top 100*	*Top 200*	*Top 500*
Austria	8.2	0.0	0.0	0.4	52.6
Belgium	10.4	0.0	0.0	61.3	122.4
Czech republic	10.2	0.0	0.0	0.0	13.1
Denmark	5.4	0.0	74.6	113.5	160.5
Finland	5.2	0.0	45.5	75.4	80.5
France	60.2	3.0	15.2	28.6	45.1
Germany	82.5	0.0	17.00	36.5	67.0
Greece	11.1	0.0	0.0	0.0	12.2
Hungary	10.1	0.0	0.0	0.0	13.3
Ireland	4.0	0.0	0.0	0.0	50.0
Italy	57.6	0.0	0.0	11.1	33.9
Netherlands	16.3	20.2	50.7	75.9	131.3
Poland	38.2	0.0	0.0	0.0	3.5
Spain	42.7	0.0	0.0	0.1	14.2
Sweden	9.0	6.7	116.5	178.8	216.9
UK	59.8	72.0	86.1	98.0	123.9
EU15	383.3	12.7	26.0	41.0	67.3
EU25	486.6	10.0	20.5	32.4	53.9
Norway	4.6	0.0	65.8	90.6	107.0
Switzerland	7.4	97.1	165.5	228.1	229.6
Australia	20.1	0.0	31.4	65.8	100.7
Canada	31.9	39.3	54.2	62.9	103.6
Japan	127.7	14.3	17.2	24.3	26.7
USA	293.7	100.0	100.0	100.0	100.00
California	36.1	234.2	198.5	163.2	103.2
Massachusetts	6.4	448.7	307.8	301.7	263.0
New York	19.3	195.7	167.4	138.7	147.7
Pennsylvania	12.4	110.7	176.9	161.0	115.2
Texas	22.9	32.7	60.9	82.8	102.5

3. What explains research performance in Europe?

An obvious starting point for economists is to look at money. Table 2 presents aggregate data on the levels of private and public expenditure on higher education across countries. The main findings are that:

— Richer countries spend relatively more on higher education than poorer countries.

Table 2 Public and private expenditure on higher education in 2001

In thousands of Euros per student				*As a % of GDP*		
Country	*Public*	*Private*	*Total*	*Public*	*Private*	*Total*
Austria	11.0	0.5	11.5	1.4	0.1	1.5
Belgium	10.6	1.6	12.2	1.4	0.2	1.6
Czech R.	2.3	0.4	2.7	0.8	0.1	0.9
Denmark	25.6	0.4	26.0	2.7	0.0	2.7
Finland	10.3	0.3	10.6	2.1	0.1	2.2
France	7.5	1.2	8.7	1.0	0.2	1.2
Germany	11.5	0.9	12.4	1.1	0.1	1.2
Greece	3.3	0.0	3.3	1.2	0.0	1.2
Hungary	2.6	0.6	3.2	1.1	0.3	1.4
Ireland	9.7	1.6	11.3	1.2	0.2	1.4
Italy	5.6	1.4	7.0	0.8	0.2	1.0
Netherlands	13.0	2.7	15.7	1.3	0.3	1.6
Poland	1.7	-*	-*	1.1	-*	-*
Spain	4.0	1.2	5.2	1.0	0.3	1.3
Sweden	18.9	1.8	20.7	2.1	0.2	2.3
UK	8.4	3.1	11.5	0.8	0.3	1.1
EU25	7.3	1.4	8.7	1.1	0.2	1.3
US	16.6	19.9	36.5	1.5	1.8	3.3
Japan	6.5	7.3	13.8	0.5	0.6	1.1

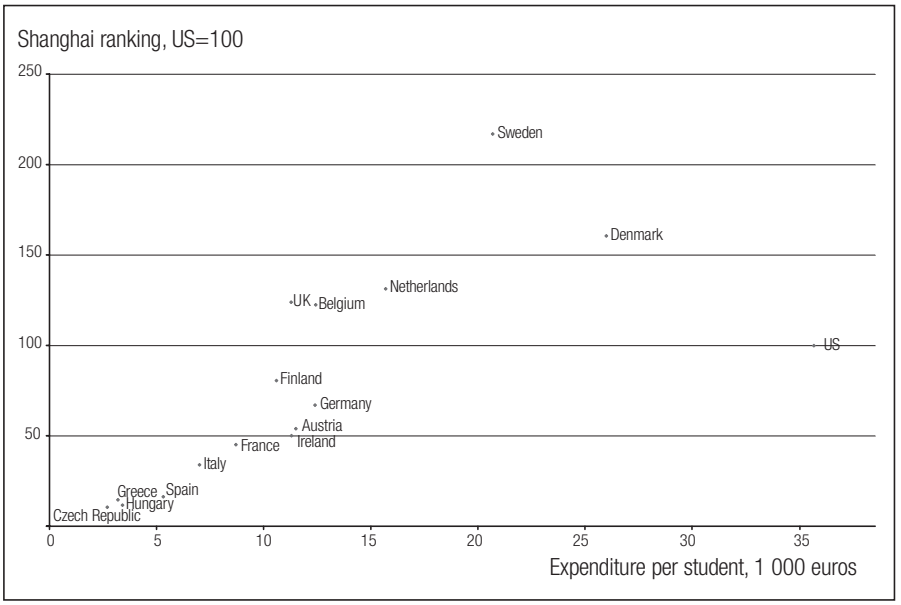
Source: European Commission, DG Research; *: not available. Note: not PPP converted.

— The US spends a lot more on higher education than any European country, especially thanks to private funding. But public spending alone is relatively higher than in the EU.
— Scandinavia also spends a lot, with most of the money coming from public sources.
— The UK spends surprisingly little (more on this later).

Figure 1 shows that there is a strong positive correlation between expenditure per student (from Table 2) and country performance (measured by the Top 500 performance values in Table 1).

However, these aggregate data do not indicate how the money is split between higher education institutions, in particular between research-oriented and teaching-oriented universities. In the remainder of this section we therefore present the results of a survey questionnaire which elicits information on individual budgets and on the governance of top research performers.

Figure 1 Relationship between expenditure per student and country performance



Source: Country performance index: Table 1; Expenditure per student: Table 2.

A. A survey of European universities

A survey questionnaire was sent to the European universities in the 2006 Top 500 Shanghai ranking[3]. We received 71 responses, an overall response rate of 36 %, which can be considered very satisfactory. We decided to focus on the ten countries for which the response rate was at least 25 % and the number of respondents at least two[4]. This left us with a total sample of 66 universities, with an average response rate of 41 % for the ten countries considered. We were able to check that, for each country, respondent universities have an average Shanghai 500 rank pretty close to that of the whole population of universities from that country, so that we could be satisfied of the representativity of our sample[5].

[3] The 2006 Shanghai ranking includes roughly 200 European universities belonging to the EU25 and Switzerland.

[4] The ten countries are: Belgium (4 responses out of 7 universities in the Shanghai 500 ranking), Denmark (2 out of 5), Germany (11 out of 40), Ireland (2 out of 3), Italy (9 out of 23), Netherlands (4 out of 12), Spain (6 out of 9), Sweden (5 out of 11), Switzerland (6 out of 8) and the UK (17 out of 43). We left out France, because only 4 out of 21 universities responded and moreover, university budgetary data are not comparable with those of other countries.

[5] In fact, respondents had a somewhat higher rank for all countries except for Spain.

Table 3 Characteristics of the universities in the sample (averages)

	Age (in years)	*Number of students (in thousands)*	*Budget per student (in thousand Euros)**	*Public status (1 if public, 0 if private)*	*Budget autonomy (1 if yes, 0 if no)*	*Building ownership (1 if yes, 0 if no)*	*Hiring autonomy (1 if yes, 0 if no)*	*Wage-setting autonomy (1 if yes, 0 if no)*	*% of Faculty with in-house PhD degree*
Belgium	284	21.7	11.3	0.5	0.4	1.0	1.0	0.0	63
Denmark	59	18.2	11.4	1.0	1.0	0.3	0.5	0.5	40
Germany	289	26.2	9.6	0.9	0.0	0.5	0.8	0.0	8
Ireland	259	16.3	12.7	0.5	0.5	1.0	1.0	0.0	49
Italy	444	44.9	10.1	1.0	0.9	1.0	0.4	0.0	24
Netherlands	217	21.4	20.5	0.8	0.8	1.0	0.8	0.2	33
Spain	342	44.8	7.0	1.0	0.5	1.0	0.5	0.0	69
Sweden	266	27.1	16.2	0.8	0.8	0.2	1.0	1.0	58
Switzerland	326	12.8	26.2	0.8	0.1	0.4	0.8	0.0	24
UK	242	14.6	24.5	0.5	0.9	0.9	1.0	0.8	8
Total	290	24.9	16.1	0.75	0.55	0.76	0.8	0.31	29

*: PPP adjusted.

Table 3 provides country averages on a variety of dimensions[6]. It confirms the high degree of heterogeneity between countries for the universities in the Top 500:

— Southern European (Italy and Spain) countries have very large (more than 40 thousand students on average) but not well-funded universities.
— Sweden and the Netherlands have universities of average size (20-25 thousand students), and better funded.
— The UK and Switzerland have small (10-15 thousand students) and very well funded universities. Comparing with the aggregate information on expenditure in Figure 1, one observes that the UK significantly favours top research performers since the universities in our sample (which belong to the group of top universities) have a budget per student about twice as large as the average for all universities in the country.

There is also a great deal of heterogeneity – albeit with some general trends – as far as university governance is concerned:

— State intervention is clearly pervasive, even when universities are not public.
— Wage-setting autonomy is rare, with Sweden and the UK being the foremost exceptions.
— Building ownership by the university is commonplace (except in Scandinavia and Switzerland).
— Hiring autonomy is prevalent, except in Southern Europe.

[6] We obtain very similar results when looking at medians rather than averages.

— Endogamy (measured as the percentage of faculty trained in-house at the PhD level) seems to be negatively correlated with country size: it is high in small countries (Belgium, Denmark, Ireland and Sweden, but not in Switzerland which is highly open to hiring scholars with PhDs from other institutions), and small in large countries (Germany, Italy and the UK, but not in Spain). This finding clearly reflects the absence of significant academic mobility between European countries.

A striking fact is thus the high variance in university governance across European countries, even among those which are performing well in terms of research. For example, among the three European countries with the best performance index, endogamy is high in Sweden but low in Switzerland and the UK, and universities are mostly public in Denmark, Sweden and Switzerland whereas they are mostly private in the Netherlands and the UK.

One dimension where there is little variance across European countries is the age of universities. Top European universities are old institutions: the average age of the 66 universities in our sample is nearly 300 years. It ranges from 220 years in the Netherlands to 450 years in Italy. The only outlier is Denmark where the average age is only 60 years. This suggests that European universities have a lot of accumulated knowledge, but may also be complicated to reform.

B. Preliminary evidence

Our survey allows us to examine how budget per student and various measures of university governance correlate with research performance measured by the Shanghai ranking. Table 4 shows that the research performance of a university is:

— positively correlated with the size of its budget per student: the higher the budget per student the better the performance;
— negatively correlated with its degree of public ownership: private universities perform better than public institutions;
— positively correlated with its budget autonomy: not being required to have its budget approved by governmental authorities is associated with better performance;
— not correlated with its building ownership: more autonomy with respect to buildings is not associated with better performance;
— positively correlated with its hiring and wage-setting autonomy: universities that decide on faculty hiring and set faculty wages do better;
— negatively correlated with its degree of endogamy in faculty hiring: universities which tend to hire their own graduates as faculty do less well.

Taken together these results suggest that the research performance of a university is positively affected by all our measures of university autonomy (except for building ownership), and also by funding. However, they not tell us: (i) which of these autonomy indicators dominates and how interrelated they are; (ii) whether funding and autonomy improve performance separately from one another, or whether there are positive interactions between the two. We now try to answer these questions with appropriate statistical instruments.

C. Funding and autonomy

We use regression analysis, a statistical technique for the investigation of relationships between variables, to assess the effect of budget and governance on research performance measured by Shanghai rankings.

We are interested in the effect of budget and university governance on university research performance. However we need to begin by taking into account two other factors that also affect Shanghai rankings, our measure of university research performance. The first is the size of the university. As Box 1 clearly indicates, other things equal, larger institutions are likely to have a better Shanghai ranking because they have more researchers. We do not have data on the number of researchers in our survey so we proxy the size of the university by the number of students. The second factor is the age of the university. Box 1 also indicates that, other things equal, older institutions may have a better Shanghai ranking because they have more alumni.

As expected, the regression analysis indicates that the research performance of universities is positively associated with their size and their age. More importantly, it also confirms the existence of a positive linkage between budget per student and research performance. These effects are statistically significant.

Once these three important factors (size, age and money) are taken into account, it turns out that one of the six governance indicators reported in Table 4, namely budget autonomy, has a statistically significant effect on research performance. The others have no statistical impact on performance.

Table 4 Correlation between budget and university governance, and research performance*

Characteristics	*Correlation coefficient*
Budget per student	+ 0.61
University governance:	
Public status (1 = public; 0 = no)	– 0.35
Budget autonomy (1 = yes; 0 = no)	+ 0.16
Building autonomy (1 = yes; 0 = no)	– 0.01
Hiring autonomy (1 = yes; 0 = no)	+ 0.20
Wage-setting autonomy (1 = yes; 0 = no)	+ 0.27
Percent of faculty with internal PhD degree	– 0.08

* Measured by the (logarithm of the) Shanghai ranking.

Table 5 Effect of budget and autonomy on research performance*

Variable	*Effect on research performance*
Size of the university (number of students)	+
Age of the university	+
Budget per student	+
Budget autonomy	+
Interaction between budget and autonomy	+

* Measured by the (logarithm of the) Shanghai ranking.

But our main result is not simply that more money or more autonomy is good for research performance. It is that more money has much more impact when it is combined with budget autonomy. To be more precise: we find that having budget autonomy doubles the effect of additional money on university research performance.

Hence, increasing budget per student helps research performance, and having budget autonomy doubles this beneficial effect.

This message based on the research performance of European universities is reinforced by the analysis of American universities presented in the next section.

4. Lessons from US evidence

The United States provide a wealth of information that can be used to go one step further in the analysis of research performance. Specifically, for the US we have access to a rich data set across US states and across time on education spending and patenting. For each state, we have at our disposal yearly information on university funding and governance and on patenting. We are able, therefore, to examine the effect of university funding and governance directly on innovation activity, rather than solely on university research performance.

Box 2 University funding, autonomy and innovation: Data and methodology

Data

For research expenditure, we use the detailed data in Aghion *et al.* (2007)[1] on how much each state spent on each type of education in all years from 1947 to 2004. We know in particular from these time series how much each US state spent on a given cohort of individuals (e.g. born in year X) in each year. Thus we know how much was spent on average on each individual at every stage of his or her studies (from primary school to post-graduate college).

For governance, we consider two alternative measures of university autonomy at the state level: (i) the percentage of universities that are private, keeping in mind that private universities are, on average, more autonomous than public universities; (ii) an aggregate autonomy index for public universities, which is constructed on the basis of several component factors. This index takes the maximum value when the public universities in the state: (a) set their own faculty salaries; (b) set their own tuition fees; (c) have lump sum budgeting (as opposed to line item budgeting); (d) can shift funds among major expenditure categories; (e) retain and control tuition revenue and/or grants; (f) have no ceiling on external faculty positions (and therefore need not hire faculty internally); (g) have no ceiling on external non-faculty positions (administrators or technicians); (h) have freedom from pre-audits of their expenditure; (i) can carry over year-end balances (rather than returning them to the state). It turns out that, like in the case of European universities, the most statistically important component factor of this aggregate index is budget autonomy.

Statistical test

We examine the effect on patenting in a US state, of increasing research education funding by $1,000 per year and per person over a sustained period, respectively in states with highly autonomous universities and in states with less autonomous universities.

Figure 2 illustrates a key result from our test: States with highly autonomous universities enjoy an accumulated impact of the research education funding on innovation which is roughly twice as high as that enjoyed by states with less autonomous universities.

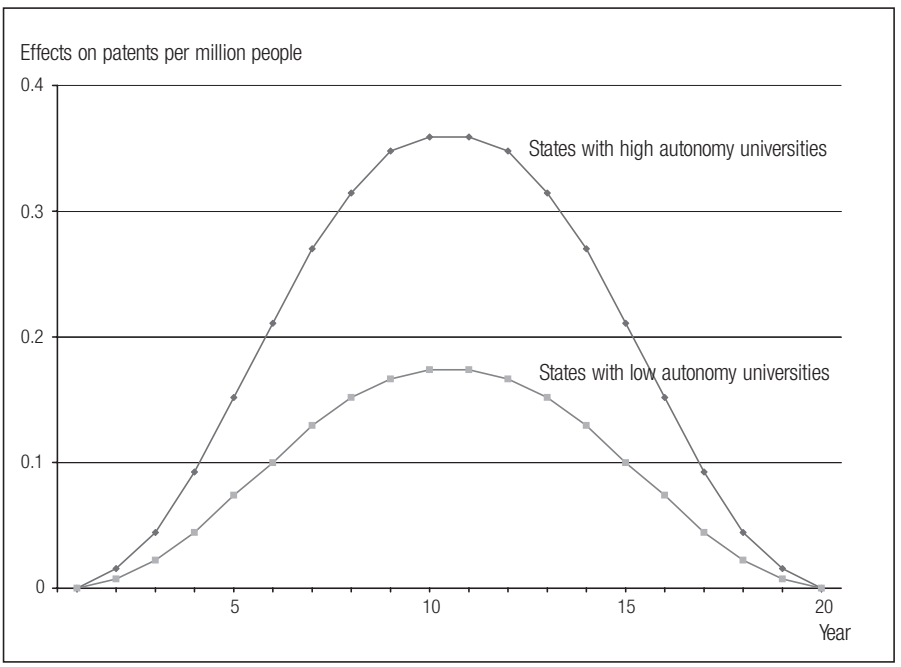
---

[1] P. Aghion, L. Boustan, C. Hoxby and J. Vandenbussche (2007), "Exploiting States' Mistakes to Evaluate the Impact of Higher Education on Growth", mimeo, Harvard.

Interestingly, there is considerable variation in university governance across states. States vary not only in the relative importance of private versus public universities, but also in the degree of autonomy granted by state authorities to public universities. Sometimes, even neighbouring states display sharp differences in governance. For instance, public universities in Illinois enjoy rather low autonomy on average, while their neighbours in Ohio enjoy instead high autonomy. These differences are persistent over time and often go back to the idiosyncratic origin of American universities, which in turn reflect differences in the preferences of university founders (e.g. Benjamin Franklin founded the private University of Pennsylvania, whereas Thomas Jefferson was the founder of the public University of Virginia).

Our strategy is to take US states' differences in university autonomy as given and then ask the following question: Does a given investment in higher education produce more patenting in a US state if universities in that state are more autonomous? The details of the statistical test are reported in Box 2. The answer to our question is a resounding yes: As illustrated in Figure 2, the effect of additional spending on patenting is roughly twice as high for states with more university autonomy. Autonomy therefore greatly enhances the efficiency of spending. This result confirms and nicely complements the one from Section 3.

Figure 2 Effects on patents of an increase in higher education expenditure, states with high autonomy vs. low autonomy universities



Source: Authors' own computations.
Note: The increase in expenditure is assumed to last from year 1 to 6. The effect on patenting accordingly starts in year 2, peaks in years 10 and 11, and ends in year 20.

5. Conclusions

In this brief we have investigated the relationship between university governance and funding on the one hand and various measures of performance on the other hand. In the first section we have tried to link our Country Performance Index based on the Shanghai ranking of universities to different aspects of university governance drawn from a survey questionnaire. In the second section of the brief we have assessed how university autonomy affects the patenting impact of university research funding.

Several interesting findings come out of our investigation.

First, the performance gap between Europe and America is large, in particular for the best-performing universities.

Second, as we broaden the investigation from the Top 50 to the Top 500 universities in the Shanghai ranking, the relative performance of European countries improves compared to the US. This, in turn, suggests strongly that quality variance is lower among European universities than among their American counterparts. It also suggests that what Europe lacks most is top-class universities.

Third, there is more than one model of university system that appears to work. For example, both Switzerland and Sweden are doing well with most universities being public, while the UK also performs well with a higher share of private universities, but also higher tuition fees and a higher degree of student selection. The UK, however, differs significantly from Switzerland and Sweden in one respect. All three perform very well in the top tier (Top 50 and Top 100), but the UK performs relatively less well in the remaining of the Top 500. This is due to the fact that the UK heavily concentrates its less than average higher education budget (in terms of GDP) on top institutions.

Indeed, a fourth lesson is that money helps performance.

Fifth, autonomy is good for research performance.

Sixth, autonomy and funding are complementary inputs to performance: more autonomy increases the extent to which additional research funding improves performance measures at the university and at the national/state/regional levels.

Policy lessons

What should be done to improve the performance of European universities?

1. European countries should invest more in their university systems. On average EU25 members spend 1.3 % of their GDP on higher education, against 3.3 % in the US. European countries should increase funding for higher education by at least 1 percentage point over the next ten years. It remains an open question how the burden of this increase is to be shared between public budgets and private funding, including tuition fees.
2. For this effort to pay off, European universities should become more autonomous, in particular with regard to budgets, and also in hiring, remuneration, programme and student selection, particularly at Master's level. What matters for good performance is both money and good governance. The two are complementary: increasing university budget has more impact with good governance and improving governance has more impact with higher budgets. We are aware,

however, that greater autonomy can be perverse and that it must be accompanied by greater performance evaluation.

Of course this Brief has focused mainly on the research function of universities and has left aside politically-sensitive issues of tuition fees and student selection, which are perhaps more directly related to the teaching function, although they also impact on research. Yet, we are confident that a reform stressing increased budget per student and greater autonomy (together with greater evaluation) will be performance enhancing, either alone or as part of a more radical overhaul of the university system, involving tuition fees and student selection. So far, our partial evidence, which will be further examined in our Blueprint, leads us to believe that there is more than one university system that works and, therefore that there are diverse paths to university reform.

Biographical notes

About the Co-Editors

Catherine DEHON holds a Ph.D. in Sciences (Statistics) from the Université Libre de Bruxelles (ULB), 2001. She is associate professor (chargée de cours) of Statistics and Econometrics at the Faculty SOCO, ULB. She is member of the board (Conseil d'Administration) of the Université Libre de Bruxelles. Her current fields of research include robust regression, robust multivariate analysis and recently the robustification of econometric methods. She is also interested in application of statistic and econometric methods on economic data, especially on economics of education. She has published articles in distinguished international reviews as *Statistics and Probability Letters*, *The Canadian Journal of Statistics*, *Statistical Papers*, *La Revue de Statistique Appliquée*, *Annals of the Institute of Statistical Mathematics*, *Oxford Bulletin of Economics and Statistics* and *CESifo Economic Issues.*

Dirk JACOBS has studied sociology at Ghent University (Belgium, 1993) and holds a doctorate (Ph.D.) in Social Sciences (Utrecht University, the Netherlands, 1998). He is associate professor in Sociology (chargé de cours en sociologie) at the Université Libre de Bruxelles (ULB) and visiting professor at the Facultés Universitaires Saint-Louis (FUSL) in Brussels. He has been a research fellow at the Research Foundation – Flanders and associate professor at the Katholieke Universiteit Brussel, the Katholieke Universiteit Leuven and the Vrije Universiteit Brussel. Main research interests of Dirk Jacobs include (ethnic) minorities, ethnocentrism, political sociology and quantitative and qualitative methodology. He has published widely on these and other topics in national and international scientific journals (including *International Migration Review*, *Journal of Ethnic and Migration Studies*, *British Medical Journal*, *Journal for International Migration and Integration*, *International Journal on Multicultural Societies).*

Catherine VERMANDELE holds a Ph.D. in Sciences (Statistics) from the Université Libre de Bruxelles (ULB), 2000. After four years as (part time) associate professor (chargée de cours) of Statistics at the Faculty of Psychological and Education Sciences, ULB, she is now full time associate professor of Statistics at the Faculty SOCO, ULB. Her main research interests are nonparametric statistics, sampling theory and teaching of statistics. Since 2000, she is also involved in different research projects relative to the student population in higher education. She has published on these various topics in international and national scientific reviews. She has published on these various topics in international and national scientific reviews, and is co-author, with Catherine Dehon and Jean-Jacques Droesbeke, of the 5th edition of *Eléments de statistique* (2008).

About the Authors

Philippe AGHION holds a Ph.D. in Economics from Harvard University, 1987. He is Robert Waggoner Professor of Economics at Harvard University, after holding academic positions at MIT, CNRS-DELTA, Nuffield College, and University College London. He was also Deputy Chief Economist at the European Bank for Reconstruction and Development. He was a Programme Director for Industrial Organization at the Centre for Economic Policy Research and a Member of the Council of the European Economic Association. He is a Fellow of the Econometric Society, the 2001 Laureate of the Jahnsson Medal and the recipient in 2006 of the Médaille d'Argent du CNRS. In 2005, he received a Honorary Degree from the Stockholm School of Economics. He has been invited to deliver the Clarendon Lectures (Oxford), the Zeuthen Lectures (Copenhagen), the Kuznets Lectures (Yale), the Gorman Lectures (University College London) and the Munich Lectures. He has written extensively on a range of topics in economics, in particular the theory of contracts and organizations and the theory of endogenous growth.

Roel D. BENNINK is coordinator research assessments and project leader for research and education assessments at the independent agency QANU (Quality Assurance Netherlands Universities, www.qanu.nl). He has supported more than thirty visiting committees as secretary and he also operates internationally as a consultant in the field of quality assurance. Before QANU became a separate agency in 2004, he was project leader and policy advisor at the Association of Universities in the Netherlands (VSNU) in quality assurance, funding policy and information policy. He holds academic degrees in English Language and Literature, Law and Sociology.

Sonja BERGHOFF holds a Ph.D. in Statistics (Universität Dortmund), where she also was a postdoctoral research fellow at the Department of Statistics. Since 2000 she is affiliated to the CHE (Centrum für Hochschulentwicklung – Centre for Higher Education Development), an independent German reform think tank for higher education and has developed what is considered to be the most sophisticated university ranking in Germany. Her activities at the CHE especially focus on the research indicators taken into account in the ranking. She is responsible for bibliometric analyses for all disciplines and has co-operated in several statistic assessment projects based on data from the CHE Ranking or other sources.

Koenraad DEBACKERE holds M.Sc. and Ph.D. degrees in Electrical Engineering and Management. He studied at the University of Gent and MIT, Cambridge, US. He is a full professor in Technology and Innovation Management at the University of Leuven. He has been a visiting professor at Nijmegen Business School and is a faculty member at the Vlerick Leuven Gent Management School. He is the head of the research division INCENTIM (International Centre for Research on Entrepreneurship and Innovation Management) at the University of Leuven. Koenraad Debackere has received several international awards and nominations for his research activities in the area of technology and innovation management. He obtained Best Research Paper Awards from the American Academy of Management and the Decisions Sciences Institute. He has authored over 70 articles and book chapters in this field and has been involved in projects for the European Commission, the Belgian and Dutch government and multinationals.

Mathias DEWATRIPONT holds a Ph.D. in Economics from Harvard University, 1986. He is a part-time Visiting Professor at the Massachusetts Institute of Technology and a Full Professor at ULB, where he was Co-Director of ECARES from its creation in 1991 until 2001. His general research area is the theory of incentives and contracts, with applications to the internal organization of firms, industrial organization and corporate finance, and the economics of higher education. He was Managing Editor of the *Review of Economic Studies* (1990-94),

and Council Member (1993-98) and President (2005) of the European Economic Association. He is a Fellow and Council Member of the Econometric Society, and was one of the three Programme co-chairs of its 2000 World Congress in Seattle. He is the laureate of the 1998 Francqui Prize and of the 2003 Jahnsson Medal. He has been Research Director of the Centre for Economic Policy Research since 1998. He is a member of DG Competition's EAGCP (Economic Advisory Group on Competition Policy) and European Commission President Jose Manuel Barroso's Economic Policy Analysis Group. In 2005, he became a Founding Member of the Scientific Council of the European Research Council.

Gero FEDERKEIL has a university degree in Sociology from the Universität Bielefeld. He has been research assistant at the Institut für Bevölkerungsforschung und Sozialpolitik (Population Research and Social Policy Department) of Universität Bielefeld and assistant at the German Science and Humanities Council. Since 2000, he is affiliated to the CHE (Centrum für Hochschulentwicklung – Centre for Higher Education Development) where his main activities concern ranking, performance indicators, evaluation and quality assurance. Within the ranking, he works in particular on indicators related to the relevance of study programmes in relation to the labour market as well as on coordination of the international expansion of the CHE Ranking.

Wolfgang GLÄNZEL holds a doctorate in mathematics from Eotvos Lorand University Budapest and a Ph.D. in Social Sciences from Leiden University. He is Professor of Managerial Economics, Strategy and Innovation at the Katholieke Universiteit Leuven. Main research interests of Wolfgang Glänzel are bibliometrics (Quantitative analysis and mathematical models of the information processes in scientific research) and Mathematics (Theory of probability distributions). He is also a Senior Research Fellow at the Steunpunt O&O Indicatoren (SOOI) at KULeuven since 2002. Wolfgang Glänzel is also Senior Scientist at the Institute for Science Policy Research, Hungarian Academy of Sciences in Budapest and Co-Editor of the international journal *Scientometrics*. Wolfgang Glänzel is author/co-author of several books and more than 150 papers in international journals and conference proceedings. In 1999 he received the international Derek deSolla Price Award for outstanding contributions to the quantitative studies of science.

Seamus HEGARTY is the former director of the National Foundation for Educational Research (NFER). NFER is the largest research organisation in the United Kingdom, comprising about 270 staff and running 70-80 research projects at any given time. He is now Chair of the International Association for the Evaluation of Educational Achievement. Born in Ireland, he took his first degree in Dublin and his doctorate in London. He has researched and written widely on special education, with a particular focus on inclusive education. He is founder editor of the *European Journal of Special Needs Education*. He is on the editorial board of five other journals. He has acted as adviser on special needs issues to UNESCO and numerous other national and international bodies. He served as principal consultant to UNESCO in preparing the Salamanca Declaration. He has completed an independent evaluation of the European Agency for Development in Special Needs Education. Seamus Hegarty was involved in the QANU-review of the research in Pedagogics and Education Science (6 faculties, 20 programmes, 2007).

Caroline HOXBY holds a Ph.D. in Economics from the Massachusetts Institute of Technology, 1994, after a Master's Degree from Oxford University which she attended on a Rhodes Scholarship. She is the Scott and Donya Bommer Professor of Economics at Stanford University and the Director of the Economics of Education Programme for the National Bureau of Economic Research. She is also a Senior Fellow at the Hoover Institution and serves as a member of the Board of Directors of the National Board for Education Sciences. She has received a

National Tax Association Award and is the recipient of the 2006 Thomas J. Fordham Prize for Distinguished Scholarship. She has written extensively on issues of educational choice, competition between schools, school finance, or the effect of unionization and of class size on educational outcomes.

Michel LUBRANO has studied Economics at Aix-Marseille University (France, 1976) and holds a doctorate (Ph.D.) in Econometrics (Toulouse, France, 1986). He is research fellow of CNRS and affiliated to GREQAM ("Groupement de Recherche en Economie Quantitative d'Aix Marseille"). GREQAM is a research center in Economics located both in Aix-en-Provence and Marseille. Main research interests of Michel Lubrano are Bayesian econometrics, Bayesian computations, econometrics of inequality, financial econometrics and bibliometry.

Andreu MAS-COLELL is a Professor of Economics (Catedrático) at the Universitat Pompeu Fabra, Barcelona, Spain. Formerly he was Professor of Economics at Harvard University (1981-96) and Professor of Economics and Mathematics at the University of California, Berkeley (1972-80). His research contributions go from abstract general equilibrium theory and the structure of financial markets to pricing policy for public firms or the economics of higher education. He has been a Sloan Fellow and Guggenheim Fellow. He holds Honorary Doctorates from the universities of Alacant, Toulouse and HEC (Paris). He has received the Rey Juan Carlos I Prize in Economics and the Pascual Madoz (National Research Prize). He has served as main Editor of the *Journal of Mathematical Economics* (1985-88), and of *Econometrica* (1988-92). He is a Fellow of the Econometric Society and was its President in 1993. In 1997 he was elected Foreign Associate to the US National Academy of Sciences and Foreign Honorary Member of the American Economic Association. In the year 2006 he served as President of the European Economic Association. From 2000 to 2003 he was Minister for Universities and Research of the Government of Catalonia. He has been designated General Secretary of the European Research Council from 2009 to 2011.

André SAPIR holds a Ph.D. in Economics from The Johns Hopkins University, 1977. He is professor at ULB, where he holds a chair in international economics and European integration. He is also a Senior Fellow of the Brussels European and Global Economic Laboratory (BRUEGEL) and a Research Fellow of the Centre for Economic Policy Research (CEPR). In addition, he is a member of European Commission President Jose Manuel Barroso's Economic Policy Analysis Group. He was an Economic Adviser to European Commission President Romano Prodi (2001-2004) and the Chairman of the High-Level Study Group appointed by him that produced the 2003 report "An Agenda for a Growing Europe", widely known as the "Sapir Report", published by Oxford University Press in March 2004. He is a founding Editorial Board Member of the *World Trade Review*, published by Cambridge University Press and the World Trade Organisation.

Philippe VINCKE, Rector of the Université Libre de Bruxelles, holds a Ph.D. in Sciences (Mathematics) from the Université Libre de Bruxelles (ULB, 1976). He is a leading scholar on the topic of multicriteria decision-making. His principal researches concern preference modelling, aggregation and decision support methods. He has published articles in distinguished international reviews as *Econometrica*, *Operational Research*, *European Journal of Operational Research*, *Journal of Multi-Criteria Decision Analysis*, *Informations Systems and Operational Research*, *Environmental Impact Assessment Review*, *Discrete Applied Mathematics*. In 2000, Philippe Vincke received the Georg Cantor Award given by the International Society on Multiple Criteria Decision Making. He was also the President of the European Association of the National Operational Research Societies and the Vice president of the International Federation of Operational Research Societies.

Table des matières

EDITIONS DE L'UNIVERSITE DE BRUXELLES

Fondées en 1972, les Editions de l'Université de Bruxelles sont un département de l'Université libre de Bruxelles (Belgique). Elles publient des ouvrages de recherche et des manuels universitaires d'auteurs issus de l'Union européenne.

Principales collections et directeurs de collection

- Commentaire J. Mégret (fondé par Jacques Mégret et dirigé jusqu'en 2005, par Michel Waelbroeck, Jean-Victor Louis, Daniel Vignes, Jean-Louis Dewost, Georges Vandersanden ; à partir de 2006, Comité de rédaction : Marianne Dony (directeur), Emmanuelle Bribosia (secrétaire de rédaction), Claude Blumann, Jacques Bourgeois, Laurence Idot, Jean-Paul Jacqué, Henry Labayle, Denys Simon)
- Aménagement du territoire et environnement (Christian Vandermotten)
- Economie (Henri Capron)
- Education (Françoise Thys-Clément)
- Etudes européennes (Marianne Dony)
- Histoire (Eliane Gubin)
- Philosophie et lettres (Manuel Couvreur)
- Philosophie et société (Jean-Marc Ferry et Nathalie Zaccaï-Reyners)
- Science politique (Jean-Michel De Waele)
- Sociologie (Mateo Alaluf et Pierre Desmarez)
- Spiritualités et pensées libres (Hervé Hasquin)
- Statistique et mathématiques appliquées (Jean-Jacques Droesbeke)
- UBlire (collection de poche)

Elles éditent trois séries thématiques, les *Problèmes d'histoire des religions* (direction : Alain Dierkens), les *Etudes sur le XVIII^e siècle* (direction : Bruno Bernard et Manuel Couvreur) et *Sextant* (direction : Eliane Gubin et Valérie Piette).

Des ouvrages des Editions de l'Université de Bruxelles figurent sur le site de la Digithèque de l'ULB. Ils sont aussi accessibles via le site des Editions.

Founded in 1972, Editions de l'Université de Bruxelles is a department of the Université libre de Bruxelles (Belgium). It publishes textbooks, university level and research oriented books in law, political science, economics, sociology, history, philosophy, …

Editions de l'Université de Bruxelles, avenue Paul Héger 26 - CPI 163, 1000 Bruxelles, Belgique
EDITIONS@admin.ulb.ac.be
http://www.editions-universite-bruxelles.be
Fax +32 (0) 2 650 37 94
Direction, droits étrangers : Michèle Mat.
Diffusion/distribution : Interforum Benelux (Belgique, Pays-Bas et grand-duché de Luxembourg) ; SODIS/ToThèmes (France) ; Servidis (Suisse) ; Somabec (Canada) ; Centre d'exportation du livre français (CELF) (autres pays).