DON F. WESTERHEIJDEN, GERO FEDERKEIL, LEON CREMONINI,
FRANS KAISER, AND MAARJA BEERKENS-SOO

RANKING GOES INTERNATIONAL

Piloting the CHE ranking of study programmes in Flanders and the Netherlands[1]

INTRODUCTION

The basic idea underlying the pilot project 'CHE Ranking of European Universities' is that the evolution of a common European Higher Education Area in the context of the Bologna process and a common European Research Area in the Lisbon strategy will lead to growing European mobility of students and higher education staff. Hence, comparable information about European higher education institutions will become more important for students as well as for academics in order for them to make well-informed choices in selecting where to go in the large European space, with perhaps 4,000 higher education institutions in more than 40 countries. Up to now such information is largely lacking. One possible instrument of providing such information is a 'student information system'—a better name than the more usual 'ranking'. Existing student information rankings are often national, thus not very useful for European mobility purposes, while international (so called 'world') rankings do not cover the European higher education area systematically and are in many respects biased in favour of English-speaking countries. In addition, the world rankings do not deliver useful information for students who are looking for an appropriate higher education institution as they show severe methodological weaknesses (e.g., most of them do not distinguish between subject areas, whereas students are looking for an institution within their particular subject/field). Moreover, they focus on institutions' research performance and do not include much information on teaching quality. We will go into the methodology in-depth in the next chapter.

The type of information that can underpin students' selection of study programmes can be provided by the CHE Ranking approach that has already started to become somewhat of a European ranking by including Austrian and Swiss universities in addition to German higher education institutions. The CHE Ranking approach has gained high acceptance in Germany and has been acknowledged by several comparative studies on ranking methodology. By successfully including Austrian and Swiss universities, a practical proof has already been provided that the method is valid for international ranking.

The aim of the project is to pilot the CHE Ranking further, beyond the German-language area, to higher education institutions in the Netherlands and in the

M. Rostan and M. Vaira (eds.), Questioning Excellence in Higher Education: Policies, Experiences and Challenges in National and Comparative Perspective, 175–193.

Flemish community of Belgium. In a first step, a common ranking of German, Austrian, Swiss, Dutch and Flemish higher education institutions is established for three subject areas. As was done for Austria and Switzerland, the extension was made in close co-operation with competent partners in the respective countries that have a profound knowledge of the national higher education systems and academic cultures. As the extension project to the Netherlands and Flanders is part of the existing, wider CHE Ranking of German, Austrian and Swiss higher education institutions, it profits from the information already available and provided through the existing ranking, which is funded independently of this project. The comparison with higher education institutions in these three countries is an additional benefit of the project. In the Netherlands, we can build on recent pilot experiences with ranking similar to the CHE approach supported by the Ministry of Education (www.studiekeuze123.nl, English version: www.studychoice123.nl).

For each subject area, data includes information on both teaching and research performance (although the present report focuses on the teaching side), as well as on facilities and services for students.

This first step was a one-year project, funded by the European Commission's Socrates programme, to test the methodology across (narrow) cultural borders, with a view to possibilities of further extension to other European countries at a later stage. In this first step, we focused mainly on first-cycle programmes, partly because second-cycle programmes had been introduced too recently to be evaluated in-depth in some of the countries participating in the CHE ranking. Further extension in the direction of second and third-cycle programmes was another 'next step'.

CONSIDERATIONS IN RANKING UNIVERSITIES

Increasing public interest in university rankings is reflected also in the amount of academic literature that has been devoted to the issue in recent years. University rankings have been examined from methodological, technical and conceptual aspects. While the literature points to many serious problems in university rankings, there seems to be a general consensus that "rankings are here to stay" (Merisotis, 2002) and that energy should go into improving rankings rather than fighting them (e.g. Marginson, 2007; Dill & Soo, 2005; van Dyke 2005). Moreover, well designed rankings can provide students with valuable information and encourage accountability in universities. Recent research consistently draws attention on potential dangers of poorly constructed rankings and suggests some principles that would make a ranking sound and beneficial.

Different rankings vary considerably in purpose and scope, in their definition of quality, the choice of indicators, and methodological designs (Usher & Savino, 2006). All these aspects are important not only for the quality of a ranking system, but also for its effect on the higher education system more generally. This chapter summarises key lessons from current international experiences in the field and points to major pitfalls that a well-designed ranking should try to avoid.

Critical Issues in the Design of University Rankings

Aggregate vs. multi-dimensional rankings. The biggest conceptual divide in university rankings is between aggregated and multi-dimensional rankings. All ranking systems collect information on various indicators such as research performance, student-staff ratios, university resources, etc. In most cases, select quality indicators are combined to produce an overall institutional ranking. The approach used is commonly known as 'weight-and-sum', which involves assigning a weight to each indicator according to its perceived importance and then using the weights to crunch the numbers into one easy-to-digest score (Clarke, 2002). Based on the aggregated scores a straight numerical ranking is produced, in which universities are put in straight rank order from best to worst. While this type of ranking is very popular because of its simplicity it has been found highly problematic from a methodological and conceptual standpoint.

An aggregated ranking assumes that there is a hierarchy of universities that is accurate for every potential user and for every purpose. Most rankings define potential students as their main target audience. An aggregated score presumes that all students have identical decision criteria and some universities are universally better for all types of students. This assumption, however, is not correct. Empirical evidence shows that students are not identical in terms of what they consider when choosing a university (see: Dill & Soo, 2005; Cremonini et al., 2007). Certain students may value research orientation while others consider more seriously the size of the institution, good mentoring, or international orientation. Aggregating the scores of all these dimensions hides significant performance differences between universities and fails to recommend the best university for the specific student.

Moreover, because aggregate rankings are based on the weight-and-sum approach, the choice of variables and the weights assigned are major problems. The choice of variables and weights reflects a set of assumptions about what promotes quality (in teaching, learning, or research). It is a matter of judgment that is not necessarily valid, comprehensive, relevant, or comparable (Dill & Soo, 2005; Clarke, 2002; Bowden, 2000; Eccles, 2002). For instance, measures of institutional environment do not appear to be straightforwardly linked to student outcomes, yet most ranking systems build formulae that implicitly assume a link (Pascarella & Terenzini, 1991).

An alternative to the aggregated ranking is to provide multiple scores for each university. The result would be a university 'report card' which provides information on different aspects of universities, but resists the temptation of ranking universities on one unique scale. This approach recognises that there is no unique hierarchy of universities, but the hierarchy depends on the individual criteria and relative importance of these criteria.

The difference between aggregated and multi-dimensional rankings reflects not only a different view on how students decide among universities, but reflects more broadly the difference in what exactly is being ranked. Any aggregated ranking is necessarily biased because it is based on a particular view on what constitutes

quality in the education (Marginson, 2007). Aggregated rankings are primarily trying to capture the prestige of a university and are designed in such a way as to keep the 'Harvards' and 'Oxfords' of the world at the top of the list. This type of rankings is heavily based on one of two criteria: excellence in research or prestige as measured by prestige surveys. Both approaches are quite problematic.

Prestige rankings are heavily based on research measures because universities' prestige is primarily generated by research excellence. The problem of research focused rankings is their relevance for students. Empirical evidence is quite sceptical about the link between research quality and teaching quality. On the contrary, research intensive universities tend to be less devoted to teaching and therefore may provide a less supportive learning environment for their students (see Dill & Soo, 2005). A university ranking that is based heavily on research quality is likely to provide students with information that could lead to less than optimal choices.

Prestige surveys contact academics, university administrators and employers and ask their opinion about the quality of the universities. The problems of such survey are manifold. It is unlikely that even academics are aware of the quality of research, furthermore teaching, in each university (Brooks, 2005). Famous universities are likely to produce misleading 'halo-effects'. A world famous university is more likely to get higher ratings not because the respondents are familiar with its performance, but because respondents assume high performance due to its reputation. A well known example is the ranking of Law Schools in the United States. According to this ranking, Princeton University was amongst the top 10 law schools in the country even though the university in fact does not have a law school (Frank & Cook, 1995). Such surveys also pose a problem of circularity because they ask deans, presidents etc. to rank institutions roughly similar to their own, resulting in "positive feedback in the creation of prestige whereby institutions that are prestigious today are more likely to have a high level of prestige tomorrow" (Brewer et al., 1999, p. 30). It is not incidental that 'new' universities are almost always ranked below the 'old' universities (Bowden, 2000).

In spite of all the criticism, prestige rankings do have their function. For some students prestige is indeed an important decision criterion. In higher education systems where prestige is associated with high selectivity, a diploma from a prestigious university has a strong signalling effect about the capacity and ambition of the particular student. This can be a valuable asset on the labour market. However, this approach to rankings implies that students are primarily concerned with the status of their degrees, rather than with what they learn. Rankings, then, degenerate into a popularity contests (Marginson, 2007).

While conceptually it is difficult to justify a discreet hierarchy of universities based on an aggregated score, from a practical perspective there seems to be a high demand for such rankings because of their simplicity and perceived certainty. The issue of aggregate vs. multi-dimensional rankings is now presented as one of the major dilemmas also in the international context (Marginson & van der Wende, 2007).

University vs. discipline rankings. A similar aggregation problem emerges also with respect to the unit of analysis: some rankings evaluate universities while others evaluate individual disciplines/programs in each university. Experts find program level rankings overwhelmingly sounder than discipline level rankings (Marginson, 2007; van Dyke, 2005; Dill & Soo, 2005). The quality of individual programs usually varies significantly across a university. Some universities may be particularly strong in one program (e.g. in sciences) while their other programs (e.g. humanities) are relatively weaker. Because students enter university to study a certain field, program level information is more helpful than university level information. Overall institutional rankings hide valuable information from potential students.

Both aggregation issues—aggregating different dimensions into a single score and aggregating program information for the entire university—are part of a larger conceptual approach of the purpose of the ranking. While almost all ranking systems claim to advise students, they differ in their view on what information students seek.

Rankings have been heavily criticised also from methodological and technical standpoints. While the purpose of this summary is not to list all potential problems, two aspects are important to keep in mind. These two issues are related to how universities are differentiated from each other and what indicators are included in the rankings.

Ranking vs. clustering. All rankings use quantitative scores. Most rankings take a mechanistic approach and rank universities based on the scores. This approach is problematic because actual differences in performance may be only marginal while the rank suggests that one university is clearly better than another. For example, a university ranked 10 could be virtually identical to one ranked, say, 17, but the miniscule differences are exacerbated when the differences are translated into discrete ranks. Empirical evidence has shown that inferences about performance differences across universities are often based on statistically insignificant differences (van Dyke, 2005). The mean value of entrance scores, for example, can be slightly higher in one university, but considering the overall variance of the entrance scores in the universities, the marginal difference is generated by a random error rather than a systematic difference. The conclusion that one university is better than the other would be in such case inaccurate.

An alternative to such a discreet ranking is to group universities based on their performance, without producing a specific rank for each university. The CHE ranking and the Australian The Good Universities Guide are amongst those who have adopted such a methodology. Universities are placed into groups—e.g. good, medium, and bad—based on their scores and in each group universities are listed (e.g. alphabetically). In-group universities have more or less comparable performances whereas universities in different groups differ substantially. This approach does not, therefore, suggest that one university is considerably better than another if the differences are miniscule or non-systematic.

Choice of indicators and measurement. One of the most important aspects in the ranking design is the choice of indicators. This choice is often based on the availability of data rather than on conscious decisions of what really reflects quality in education. Rankings typically use some combination of proxies, often making implicit assumptions about causal links between institutional factors and student outcomes. Commonly used measures of teaching quality are student-staff ratio, selectivity in enrolment, student entrance scores, resources available to students, and research quality. While each of these indicators is arguably associated with learning, none of the indicators measure learning directly. Lack of indicators that capture the actual outcomes of universities' teaching and the 'value-addedì of the educational process is identified as a problem in many ranking systems (Dill & Soo, 2005).

Effective teaching output measures are not easily available. Student surveys are one of the best alternatives to ambiguous proxies such as research quality or university inputs. Although student perceptions of the university are not objectively comparable because of different expectations, they do provide some information on 'customer satisfaction'.

The set of indicators is a fundamental issue in ranking design. Dill & Soo (2005) conclude that the set should satisfy the attributes of relevance, comprehensiveness, validity, and functionality. In short this means that indicators should reflect the dimensions that students truly consider when choosing a university. To build an unbiased picture, all critical dimensions of academic quality should be included in the set of indicators. Indicators should actually measure what they intend to measure and provide reliable information. And lastly, the measures should be robust so that they do not encourage gaming and manipulation is part of universities. The effect of rankings on universities is an important issue and deserves a more detailed discussion in the next section.

Effects of Rankings on Higher Education Systems

University rankings are meant to provide information on the relative performance of universities. However, university rankings are not only passive observers, but they have an effect on universities' behaviour and arguably also on how their users perceive academic quality. Ideally a ranking would encourage universities to improve their performance, but this positive impact can be achieved only under certain circumstances. Empirical evidence rather points to perverse and dysfunctional effects of university rankings. Current rankings are found to produce a lot of gaming and manipulation in the system. In the increasingly competitive higher education market, universities consider it ever more important to be ranked and remain at the top of the list. This, however, is not necessarily an expression neither of quality education nor of sincere interest in student learning. Universities will go to great lengths to improve their ranks. For example, universities have been found to boost their selectivity score by attracting more applications (Ehrenberg, 2002). This behaviour, costly for institutions and candidates alike, fails to help universities improve their performance. The effect of university ranking is deeper

than mere manipulation of information by universities. Ehrenberg (2002) also argues that university rankings are one reason why universities are becoming increasingly costly in the United States. Fierce competition for a higher rank calls for substantial investments, for example in student merit aid scholarships. Hence, rankings tend to influence strategic decisions and investments in universities.

An influential ranking can have a large effect also on the entire university system. Marginson (2007) argues that prestige rankings in Australia are making the higher education landscape more homogenous. Since prestige rankings are heavily based on research performance, universities are encouraged to concentrate on this area even if by their original mission they might be more teaching focused institutions. While universities become more homogenous, they may actually become more differentiated in terms of their performance. Better performing universities attract more financial resources and more qualified staff and increase their lead even further.

Hazelkorn (2006) studies how higher education institutions reacted to rankings. Almost without commenting on the methodology, institutional respondents took the outcomes seriously and many sought to improve their institution's position in the (world-wide) rankings. Over half her respondents found that rankings had a positive impact on their institution, mostly through increased (comparative) publicity and reputation in students' eyes. They helped finding academic partners, stimulated curriculum renewal and boosted staff morale. Moreover, the view was widespread among higher education institutions that stakeholders (students, research contractors, fellow higher education institutions, etc.) used rankings in their decision-making. At the same time, large majorities of her respondents found that rankings favoured established higher education institutions, led to more hierarchy in the system, were distortive, and emphasised research over education.

Espeland and Sauder (2007) study how the ranking of American Law Schools affects the behaviour of these institutions. They observe that rankings indeed produce gaming and manipulation, that they tend to affect institutional strategic decisions, and that they hinder heterogeneity among law schools. But Espeland and Sauder also argue that rankings produce a self-fulfilling prophecy. Rankings encourage schools to become more like what they measure, which then increases the validity of the measures.

Many of these issues are produced by specific ranking types—the prestige rankings. The dysfunctional (and socially costly) effects of such rankings add to their conceptual weakness. Their experience however reminds that rankings are not only a neutral observer but also a participant in the higher education system. Perverse effects of rankings can be avoided with a careful design. An ideal ranking would not only provide adequate and helpful information to students but encourage universities to serve their students better.

International Rankings

As a response to the increasingly global higher education market, several international university rankings have recently been launched of which two capture

most attention: The Shanghai's Jiao Tong University's *Academic Rankings of World Universities* (since 2003) and the *Times Higher Education Supplement*'s *World University Rankings* (since 2004). Comparing universities in different higher education systems adds another layer of complexity to the discourse. Both of the rankings are prestige rankings. They both produce an aggregated unique score for each university; they are conducted at a university rather than program level; they produce a discreet hierarchy of universities; and are primarily prestige oriented. The Jiao Tong University ranking is based on research excellence and includes indicators such as Nobel Prize winning scientists, highly cited scientists, and articles in the journals *Nature* and *Science.* The *Times Higher* ranking is heavily driven by a world survey of academics.

The problems of these rankings are similar to other prestige rankings. Marginson and van der Wende (2007) argue that a better approach to global rankings begins from the recognition that all rankings are partial in coverage and contain biases: "It is valid to engage in rankings provided they are tailored to specific and transparent purposes, and interpreted only in the light of those purposes" (p. 322).

Another major challenge in international rankings is the comparability of data. It is quite evident in national rankings that the choice of indicators is often driven not by conceptually justified measures but by the availability of data. In the national context there is often either a common source of comparative data on universities or norms what data universities should collect and report. Even then not all measures are equally relevant for all types of universities. Data issues are much more severe in the international context. Universities in different countries are subject to different regulations, expectations, and social norms. While selectivity of a university, for example, is an important factor in the U.S. rankings, the higher education system in the Netherlands or Germany is not structured around the notion of selectivity. The number of declined applications or even the academic ability of the incoming class would not carry the same meaning in these countries as in the U.S. where it represents 'student demand' and 'market value'. In an international ranking it is therefore even more crucial to develop a sound justification for any measure included in the ranking, to develop its causal link to educational quality, and to ensure that it measures the same thing in all countries.

The international experience with university rankings provides with many lessons about the implications of different ranking systems. While rankings have been heavily criticised from a conceptual and methodological standpoint, and for their potentially dysfunctional effects, the criticisms should be considered as a constructive input in the process of improving the quality and effectiveness of rankings.

The CHE ranking has been widely praised as the current best example of university rankings. Usher and Savino (2007) title the CHE ranking "the best practice" in higher education rankings and Marginson (2007) argues for a CHE-type ranking also for the international setting. The advantage of the CHE ranking lies in its conceptual and methodological design, which circumvents the most common problems mentioned above. It is an informational tool and provides

information on various aspects of universities' performance at the program level. Students can design their individual ranking based on criteria they themselves consider most relevant for their decisions. The information source has been made user-friendly with a web application. On the web students can prioritise their decision criteria, allowing the program to produce the list of the most suitable universities for them. Universities are presented in groups, not in ranks, and thereby the system does not exacerbate marginal or random performance differences. Finally, the indicators are not dominated by research excellence or prestige, but an important part of the ranking is student survey. For these reasons, the CHE methodology was chosen for this pilot project.

STEPS IN THE PILOT PROJECT

For the pilot project, it was decided to limit the number of different questionnaires to one for students and one for the participating faculties/institutions, besides collecting bibliometric research information for one of the areas.

Both in the Netherlands and in Flanders, participating higher education institutions were informed about the ranking exercise and their own activities for the ranking (surveys, delivering of data) by the relevant national partners.

From October 2006 to June 2007, the relevant data were to be collected in the Netherlands and in Flanders. With regard to faculty and institutional data, as much as possible questionnaires were 'pre-filled' with data from the SKI database, i.e. the database underlying the existing student information website in the Netherlands. This database proved to cover the information needs for the CHE methodology only partly, as had been expected based on our previous analysis of the commonalities and differences between the two.

The CHE preformed comparative analysis of the data from all participating countries according to its standard methods. Common indicators were calculated (numerical values, rank groups). However, in the end only two pilot programmes were willing to have their results included in the CHE database (and these institutions had wanted to do that even without the pilot project). All other data are treated as confidential.

RESPONSE AND RESULTS FOR PILOT PROGRAMMES

Flanders

For Flanders, the CHE EuroRanking was one of the very first pilot activities with regard to system-wide student information systems. Fourteen programmes signed up for the pilot project, but one withdrew before data collection started. In the thirteen actual pilot programmes, as is the standard CHE procedure, up to 500 students were approached via the organisational channels of the study programme to complete an online questionnaire on their opinions with regard to qualitative aspects of their study, together with opinions and data regarding their study situation (e.g. on their living quarters). With absolute response numbers often near

the lowest acceptable level (CHE accepts student opinions on study programmes only if at least 15 answers are received), usable results were obtained for eight study programmes. Response rates by students (about 10% net[2], or 440 responses) were lower than CHE is used to, even with applying the same procedures to increase response, with a reminder for non-respondents and with the same online questionnaire; still the Flemish response rate was higher than that in the Netherlands (see next section).

Three faculties returned the faculty questionnaire, although the pilot institutions had been involved in a long process of drafting a Dutch-language version of the questionnaire adapted to the Flemish higher education system.[3]

Results from Student Questionnaires. There are two major ways of looking at the students' responses. First, in absolute scores, and second in comparative ranking with the other programmes in the pilot.

In absolute terms, the student judgements almost all fall in the range of 2 to 3 on the 6-point scale (1 is 'very good/high', 6 is 'very bad/low') to which the originally used 1 to 10 scale results have been recalculated. Calculating back, this means that the average 'judgement overall' of 2.4 on the 6-point scale corresponds to 7.5, i.e. reasonable to good.

In terms of ratings, the Flemish programmes were compared with their counterparts in the large CHE database of German study programmes with some Austrian and Swiss as well. In the CHE method, student judgement indicators are only rated as 'top group' or 'bottom group' if they deviate significantly from the average judgement in a statistical sense; all that are not that far from the average are rated in the 'middle group'.[4] Here, the perhaps surprising result is that although many average judgements per study programme fall in the middle category, there are a comparatively large number of judgements in the 'bottom group', but none in the 'top group'.

The Netherlands

Response by Students and Study Programmes. Out of the twelve 'slots' in the matrix of study programmes for the pilot project in the Netherlands, eleven institutions eventually reacted positively to the invitation. One institution that originally had signalled interest withdrew for organisational reasons at a late moment. From the resulting eleven pilot programmes, again up to 500 students per programme were approached via the organisational channels of the study programme to complete an online questionnaire. As in Flanders, absolute numbers of returned responses were often near the lowest acceptable level, and usable results were obtained for eight study programmes. Response rates (7% gross, about 5% net) were clearly lower than CHE is used to. A possible explanation lies in the fact that there is already another student information system in the Netherlands, for which students are also surveyed (Studiekeuze 123). Moreover, the number of questionnaires for internal quality assurance schemes for which students are

approached may be larger in the Netherlands than in e.g. Germany or Austria, leading to 'evaluation fatigue' among Dutch students.

Although pilot programmes were volunteered by their higher education institutions, no more than four succeeded in filling out the institutional questionnaire. When asked for difficulties with the institutional questionnaire, responses were obtained from four (partially different) institutions. Main issues that were mentioned, included:

- Questions for data were in terms unfamiliar to our administration;
- Structure of questionnaire with some data per programme, some for the whole faculty, was confusing;
- Communication by researchers should have been better.

The first and second points are problems inherent in international data collection: organisation of study programmes in the higher education institution is largely dependent on national traditions, regulations and data collection needs, which do not easily transfer across borders. In one response, a possible solution was suggested, namely to organise visits by researchers would have been better so that terms and data could have been explained.

Results of Student Questionnaire. With regard to the absolute scores, the overall judgement of students across all study programmes in the pilot is 2.38 in the 1-6 scale, which in the original 1-10 scale in the questionnaire corresponds to 7.5, i.e. reasonable to good. There were practically no judgements on individual indicators where study programmes deviated significantly from the national average: with one exception all would fall in the 'middle group' if a Dutch-only CHE-type ranking would have been made. The one exception is one programme that scored in the 'top group' with regard to 'courses on offer'.[5]

In the comparative ranking view, the psychology study programme at the University of Maastricht[6] mostly scores in the top group (green on the CHE web site), while most other programmes predominantly end in the middle (yellow) and bottom groups (red on the CHE web site).

Looking for an Explanation for Low Ranking Results in the Pilot. The predominance of bottom group rankings for the Dutch and Flemish study programmes were not expected. In fact, in the CHE rankings, since their beginning, there is a 26%-50%-24% division between the top, middle and bottom group judgement by students.

For the Dutch and Flemish study programmes in the pilot, there are about 50% of comparative rankings in the bottom group. What may explain this result? There are several possibilities:

Option 1

The quality of these eight programmes is indeed significantly worse than the average quality of study programmes in Germany (and Austria and Switzerland).

- i) *Counterargument*: Anecdotal evidence holds that the objective study situation (e.g. facilities, student-staff ratio, 'crowding' of lecture halls) at Dutch universities and *hogescholen* as a rule is certainly not worse than what students encounter in Germany. There is a continuous net mobility of students from Germany to the Netherlands, and we may suppose that there is some degree of rationality in this movement.
- ii) *Counterargument*: The seven study programmes cut an average to relatively good figure among Dutch study programmes, according to their rankings in SK123.
- iii) *Counterargument*: There is reasonable agreement between the rankings in CHE and SK123 for three of the seven programmes, but for four others, the SK123 results are rather better than those in the CHE pilot (see Figure 1).[7]

Option 2

Dutch students responded less positively to the CHE pilot than to the SK123 questionnaire.

- i) *Counterargument*: In four cases, the general opinions are almost the same (see Figure 4, the cluster of programmes around 7.0), but in the other three cases, the CHE-pilot score was more positive than the one in the SK123 data.

Option 3

The projection of the Dutch 1 to 10 scale on the German/Austrian 6 to 1 scale is not correct.

- i) *Explanation*: The endpoints of the scales were given explicit meanings, in both countries this was 'very bad' to 'very good'. The recoding from one to the other proceeded from the assumption that 'very bad' to 'very good' has the same meaning across countries and cultures, and that the figures on the scales are equidistant. That is to say: the difference from 1 to 2 in the Netherlands has the same meaning as the difference between 5 and 6 or between 8 and 9, and these distances correspond to differences of 5/9 of a point on the German scale (see Figure 2 in the Appendix).
- ii) *Counterargument*: However, the scales had been chosen to be intuitively known to the respondents, because in Germany, Austria and Switzerland 6-point scales are in use for examination grading; in the Netherlands, a 10-point scale is used. This might imply that respondents use the interpretations of the grading scales. In that case, especially the cut-off between a positive and a negative meaning of a grade ought to be taken into account. Practically: what grade is needed in order not to fail? In Germany, only 5 and 6 are fail grades, so 4 to 1 are increasingly 'good'; in the Netherlands, 1 to 5 are fail grades, and 6 to 10 are 'good'. However, if the Dutch student responses would be recalculated to get to the same cut-off point, the ratings for Dutch study programmes would be even lower than they were now. It would have the contrary effect to what was sought (compare Figure 2 with Figure 3).

– iii) *Explanation*: It is possible that for reasons of tradition or culture, grades given in the Netherlands are lower than those in Germany. For instance, the general average of customer satisfaction with services in the Netherlands was 7.6 on a scale of 1 to 10 (76% of the maximum attainable; surveyed 2007-07-02; www.tevredenheidsindex.nl, retrieved 2007-09-28), which is hardly different from the 7.5 that students in the Netherlands and Flanders gave to their study programmes overall. We found no hard evidence, but it is possible that German customers value services higher than Dutch ones. In other words: a 76% score in the Netherlands might indicate the same level of satisfaction as, for instance, an 80% score in Germany. This needs further research before any empirical statements can be made. However, it points to the need to calibrate scales between countries when making cross-national comparisons. Calibration should of course take place on other data than higher education in order to avoid tautologies.

Option 4

Dutch students are more demanding from their study programmes than German students.

– i) *Estimate*: this may be true. An indication is that with regard to the total judgement, foreign students (mostly German) were more positive than Dutch students (Table 1); due to the large standard deviation in the small sample of foreign students, this result was not statistically significant.
– ii) *Explanation*: Perhaps such a difference derives from Dutch students having attitudinally adapted to paying tuition fees of ca. € 1,500 per year ('we pay, so we expect something decent in return'), while in Germany such an adaptation to the then-recently introduced tuition fee of at most € 500 per semester (€ 1,000 per year) has not (yet) taken place.

Table 1. Comparison of mean general opinions between students from Dutch and foreign parents

Respondents' parents	*N*	*Mean*	*Std. Deviation*
Dutch	258	7,48	1,188
foreign	56	8,21	1,486

In sum, the 'technical' option 3 can be refuted. Option 1, i.e. that the Dutch study programmes have lower quality than the German ones was deemed improbable, on arguments supported by the more objective data, too. Finally, option 4 about different levels of expectation is left standing, and this is connected to option 3 argument iii). Both point to international differences in expectation, satisfaction and how to express them. The difficulty with different expectation levels is that they cannot easily be corrected in a technical manner. A possibility worth pursuing is to target data collection much more to the (international) target population for which an international student

information tool is intended in the first place: instead of eliciting average opinions of all students, one might focus on the opinions of foreign students only. (This assumes that foreign students have homogeneous expectation levels, which may be warranted to some extent if they hail predominantly from one country[8], but is a heroic assumption if they come from many different regions.) As a preliminary indication, this test was made on the data from the Dutch study programmes in the pilot. Foreign students did tend to give a higher overall judgement, but the difference failed to reach statistical significance.

LESSONS FOR FUTURE STUDENT INFORMATION SYSTEMS

From this project, we drew a number of lessons that might be useful for other projects that try to apply (ranking) methodologies developed in a certain country's context to another. We present them here as a list of points for ease of reference.

- I. Design different international student information systems (SISs) for different target groups.
  - a. If the target group of students consists mainly of students in a 'core' country choosing an (undergraduate) study programme, adding some foreign study programmes to a national SIS may be the most appropriate solution.
  - b. When it comes to top-level second or third-cycle programmes, Europe-wide populations of excellent students may be the target group and a dedicated SIS may need to be designed.
    - i. This implies finding international publication channels for the SIS.
    - ii. Realise that there may be different choice models applying to 'national' vs. 'foreign' study choice.
- II. Make an international SIS only with volunteering study programmes in higher education institutions from the different countries. Do not try to force whole national SISs to become linked.
  - a. Make clear to study programmes and higher education institutions what is attractive about an international SIS, as compared to other (international) marketing instruments (e.g. institutional rankings).
- III. Work incrementally, with continued pilot projects, i.e. add a limited number of study programmes in a limited number of knowledge areas and a limited number of countries per year. This approach is advocated to get to grips with the complications of different knowledge areas and countries while keeping resources needed for the SIS to a manageable level.
- IV. Minimise additional data collection. Where possible, co-operate with existing national data collection schemes, e.g. in the framework of national SISs or national accounting systems.
- V. When collecting data on purpose for an international SIS, use the same methodology in all countries.
  - a. A common core set of indicators must be defined, to determine which (non-core) data may be taken in perhaps somewhat different forms from

national SISs and other data sources in contrast to which must be collected additionally (the core indicators).
  - b. In additional data collection for core indicators, use scales to which all respondents give similar meanings.
- VI. Do further research for designing methods to overcome national or cultural differences in students' replies to questionnaires.
  - a. Compare opinions of national vs. international students about a study programme: they may have different levels and types of expectations.

THE AFTERMATH: WHAT HAPPENED AFTER THE PILOT PROJECT?

Once the pilot project was completed, including dissemination seminars and similar activities, the issue of follow-up presented itself.

Regarding Flanders, one can be short: the experiment with providing this type of 'ranking' information was not valued positively and neither did the Flemish higher education institutions wish to join the CHE rankings on a regular basis, nor did they want to set up their own student-information system. Whether this was due to the conviction among the higher education community that in such a relatively small higher education system tacit information would be sufficient to guide prospective students, or to the fact that the pilot project had been stimulated by the Department for Education of the Flemish government rather than by the higher education organisations themselves, remains an unclarified question to date.

In the Netherlands, cooperation between Studychoice123 and the CHE ranking was intensified in the following years. Some of the lessons drawn up in the previous section guided the steps towards cooperation. In particular, some small-scale studies were set up to designing methods for overcoming cultural differences in students' replies to questionnaires. Simultaneously, negotiations took place between the two organisations to try to find a common basis for future questionnaires to students, e.g. for the newly-introduced master-level programmes. However, it appeared that the value of maintaining time series of data, the philosophies behind the design of questionnaires and the consideration that national student bases (and student recruitment) were more important than international ones, resulted in the two systems remaining separate. Moreover, the Dutch SIS management was occupied with integrating student questionnaires for SIS within the country, which in 2010 led to the Studychoice123 questionnaires becoming the (almost) single national student satisfaction questionnaire, used for several purposes, including internal quality management of a large number of higher education institutions. Yet, internationalisation of student movements and the desire of especially university managers to benchmark themselves at a larger scale than within their own country led to all universities of the Netherlands (but not many non-university colleges) also taking part in the CHE rankings since 2009.

Rankings, then, are increasingly going international—the CHE ranking also obtained more participants in especially the medical field from several Central European countries in the same period. The process, however, remains not only incremental, as we suggested in our 'lessons' section, but also unpredictable due to

all the other contextual factors that play a role in adopting policy changes in the world of higher education.

APPENDIX

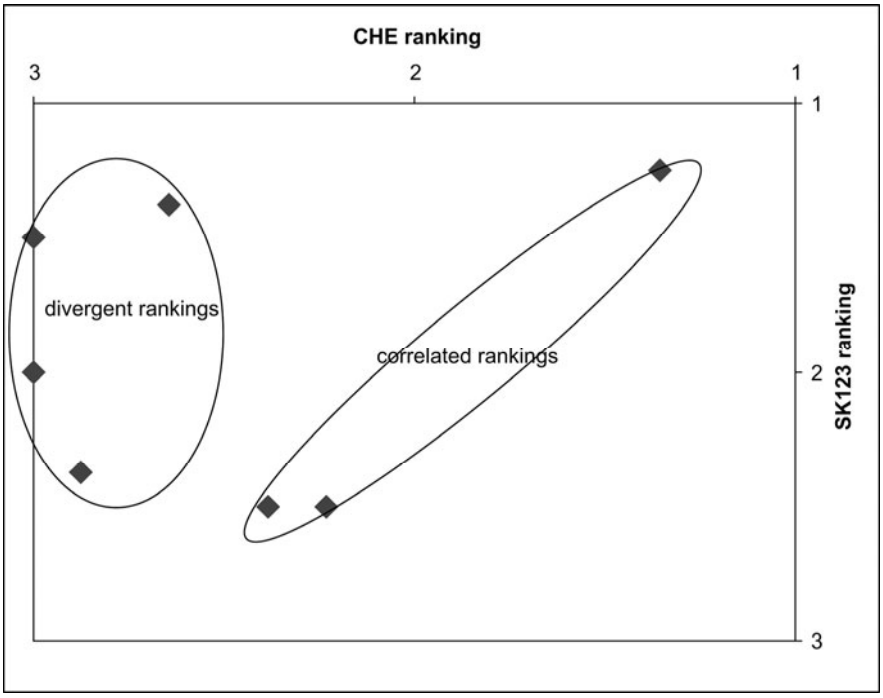

Figure 1. Correlation between CHE pilot and SK123 rankings of study programmes in the Netherlands, averaged over 8 (SK123) and 13 (CHE) indicators.

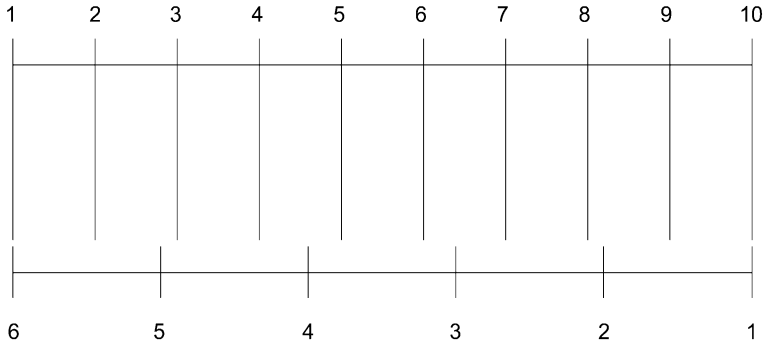

Figure 2. Equidistant projection of Dutch (top) on German (bottom) scales.

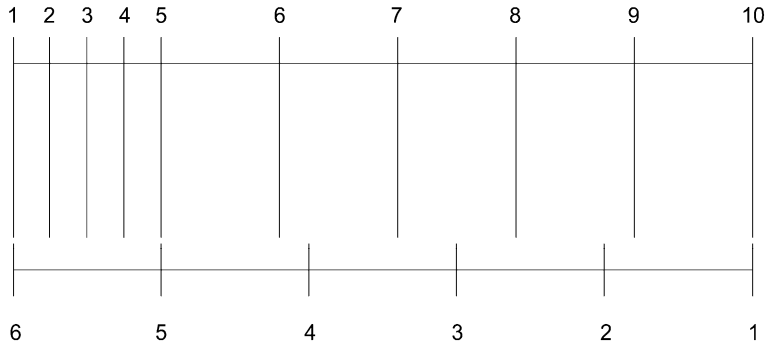

Figure 3. Projection of Dutch (top) on German (bottom) scales with same 'cut-off' or fail point.

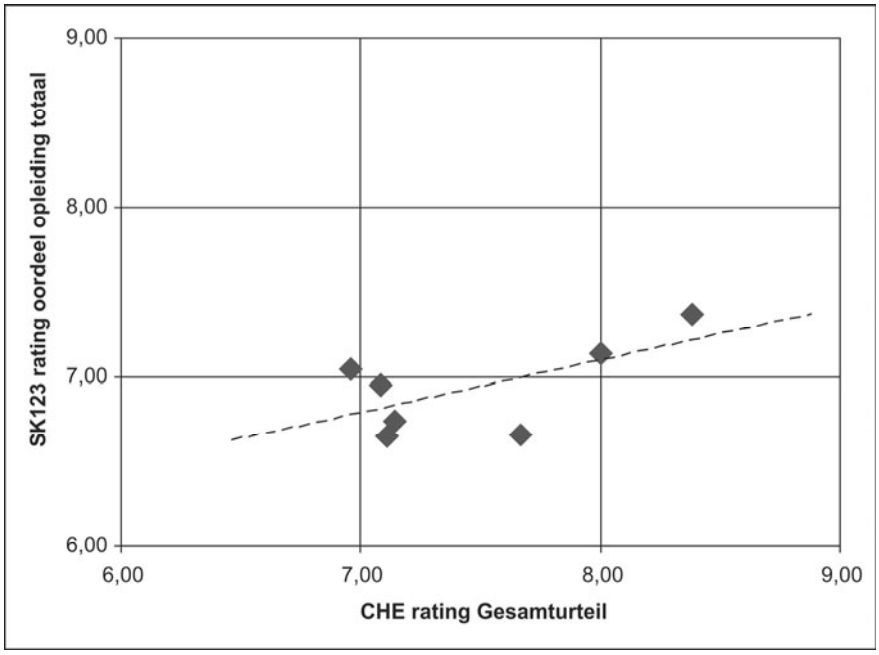

Figure 4. Correlation between mean general student opinions on study programmes in the Netherlands in CHE-pilot and SK123, calculated to scale from 1 = very bad to 10 = very good.

NOTES

[1] This project has been funded with support from the European Commission. This paper reflects the views only of the authors. The Commission cannot be held responsible for any use which may be made of the information contained therein.

[2] In the net response rate, only Bachelor-phase students of the second and higher years were regarded.

[3] This included both the faculty and student questionnaires. Some questions were added to the pilot core questions that were of special interest in Flanders.

[4] The principle of the CHE groups per indicator is to distinguish programmes that are significantly higher or lower than the average for all study programmes of the same kind on that indicator, i.e. outside the 95% confidence interval (Berghoff et al., 2007, pp. 51-52); all others are interpreted as 'middle group'.

[5] Please see the previous section on Flanders, where it was explained that this exercise was not methodologically sound enough to give much weight to findings, but was made to get an impression of what the pilot would look like in the national context. See also below on the comparison with the broader Dutch base of Studiekeuze123-data.

[6] This is one of the two Dutch programmes entered into the public CHE website. For that reason, the name can be mentioned here.

[7] The figure is methodically risky: it greatly reduces data by taking the average of ranking groups of 16 indicators from the CHE pilot (horizontal axis) and of 8 indicators in the SK123 on more or less similar issues (vertical axis). Top, middle and bottom group definitions are taken from the respective ranking sites at face value.

[8] Even within a single, small country expectation levels may differ significantly. For instance, in Dutch ranking publications it appears that students in the urban area of Holland are more critical than their colleagues in other parts of the Netherlands.

REFERENCES

Berghoff, S., Federkeil, G., Giebisch, P., Hachmeister, C.-D., Hennings, M., & Müller-Böling, D. (2007). *CHE Hochschulranking: Vorgehensweise und Indikatoren*. Gütersloh: CHE.

Bowden, R. (2000), Fantasy Higher Education : University and College League Tables, *Quality in Higher Education, 6*(1), 41–60.

Brewer, D., S. Gates, & C. A. Goldman (1999). *In Pursuit of Prestige: Strategy and Competition in U.S. Higher Education*. New Brunswick, NJ: Transaction Press.

Brooks, R. (2005). Measuring university quality. *Review of Higher Education, 29*(1), 1–21.

Clarke, M. (2002), Some Guidelines for Academic Quality Rankings. *Higher Education in Europe, 27*(4), 443–459.

Dill, D.D. & Soo M. (2005). Academic Quality, League Tables, and Public Policy: A Cross-National Analysis of University Ranking Systems. *Higher Education, 49*(4), 495–534.

Dyke, N. van (2005). Twenty Years of University Report Cards. *Higher Education in Europe, 30*(2), 103–125.

Eccles, C. (2002). The Use of University Rankings in the United Kingdom. *Higher Education in Europe, 27*(4), 423–432.

Ehrenberg, R. G. (2000). *Tuition Rising: Why College Costs So Much*. Cambridge, Mass: Harvard University Press.

Ehrenberg, R. G. (2002). Method or Madness? Inside the 'USNWR' College Rankings. Paper presented at the Wisconsin Center for Advancement of Postsecondary Education Forum on the Abuse of College Rankings. Madison, Wisconsin, 20–21, November.

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

Federkeil, G. (2002), Some Aspects of Ranking Methodology The CHE-Ranking of German Universities. *Higher Education in Europe, 27* (4), 389–97.

Frank, R. H. & P. J. Cook (1995). *The Winner-Take-All Society: How More and More Americans Compete For Ever Fewer and Bigger Prizes, Encouraging Economic Waste, Income Inequality, and an Impoverished Cultural Life*. New York: The Free Press.

Gater, D.S. (2002). *A Review of Measures Used in U.S. News & World Report's "America's Best Colleges"*. Gainesville, FL: Lombardi Program on Measuring University Performance, University of Florida.

Hazelkorn, E. (2006). Impact and Influence of League Tables and Ranking Systems on Institutional Decision-Making. Paper presented at the HRK/OECD conference 'Institutional Diversity: Rankings and Typologies in Higher Education', Bonn.

Marginson, S. (2007). Global University Rankings: Implications in General and for Australia. *Journal of Higher Education Policy and Management, 29*(2), 131–42.

Marginson, S. & Wende, M. van der (2007). To Rank or To Be Ranked: The Impact of Global Rankings in Higher Education. *Journal of Studies in International Education, 11*(3–4), 306–29.

Merisotis, J.P. (2002). On the Ranking of Higher Education Institutions. *Higher Education in Europe, 27*(4), 361–363.

Pascarella, E.T. & P.T. Terenzini (1991). *How College Affects Students: Findings and Insights from Twenty Years of Research*. San Francisco: Jossey-Bass.

Usher, A. & Savino, M. (2006). *A World of Difference: A Global Survey of University League Tables.* Educational Policy Institute.

Don F. Westerheijden, Gero Federkeil, Leon Cremonini, Frans Kaiser, Maarja Beerkens-Soo
CHEPS – Center for Higher Education Policy Studies
University of Twente