Ranking of rankings: benchmarking twenty-five higher education ranking systems in Europe

Ingo Stolz · Darwin D. Hendel · Aaron S. Horn

Abstract The purpose of this study is to evaluate the ranking practices of 25 European higher education ranking systems (HERSs). Ranking practices were assessed with 14 quantitative measures derived from the Berlin Principles on Ranking of Higher Education Institutions (BPs). HERSs were then ranked according to their degree of congruence with the BPs. Additionally, the three domains of methodology, transparency, and consumer-friendliness were proposed to underlie the BPs, and the measures were also applied for assessing HERSs regarding these domains. Results indicate that the cooperating HERSs by CHE (Germany), AQA (Austria) and swissUp (Switzerland) exhibit the highest congruence with the BPs. However, no European HERS demonstrates good overall congruence with the BPs, mainly due to methodological shortcomings. Results further show that HERSs compiled and published by non-profit research entities seem to perform better than the majority of HERSs published by for-profit news outlets. International comparisons indicate that HERSs published in German-speaking countries and the Netherlands tend to exhibit a higher congruence with the BPs. Overall, this study hopes to stimulate the improvement of ranking practices through benchmarking with existing exemplary models. The quantitative assessment tool further promises to be useful in explaining relative stability or change of higher education institutions in HERSs, as well as in helping to explain resource allocation pressures within higher education institutions.

Keywords University ranking · Higher education · Berlin Principles · Europe · Comparative · Benchmarking

Introduction

Higher education ranking systems (HERSs) find themselves in an ever-greater position of influence (Marginson 2007). They have "cemented the notion of a world university

market" (Marginson and van der Wende 2007, p. 306) in which higher education systems and higher education institutions are measured according to their relative standing on a global scale, thus introducing the notion of competition among higher education institutions as a new paradigm in most countries (Altbach 2006). Also, the assessment of the standing of a particular higher education system and/or higher education institution in this market is again referenced back to HERSs (Hazelkorn 2007; Marginson 2007), often due to a lack of alternative evaluative systems, as in the case of Europe (Federkeil 2008). Thus, HERSs have become established as seemingly unquestioned tools for assessing quality (Federkeil 2008) and university excellence (Taylor and Braddock 2007), for promoting accountability (IHEP 2007), classification (McCormick 2008), comparison (McDonough et al. 1998), evaluation (Bellon 2007; Breimer 2007) and information distribution (Bellon 2007; Yorke 1997), as well as for initiating strategy development on the part of higher education systems and higher education institutions (Hazelkorn 2007; Marginson and van der Wende 2007). Consequently, at least as long as "everyone wants a world-class university" (Altbach 2003), HERSs "clearly are here to stay" (IHEP 2007, p. 2) and "will continue to be published in ever-increasing numbers" (Bowden 2000, p. 58).

Given this position of influence, discussions about HERSs are waged from a wide variety of perspectives. However, the following two issues can be considered as problematic in these discussions: (1) a selection bias when analysing the diverse ranking practices of HERSs that are applied internationally; and (2) the lack of benchmarking and/or ranking efforts of those systems by using standards of good ranking practice. First, existing comparative analyses of HERSs (Cheng and Liu 2008; Dill and Soo 2005; Usher and Savino 2006, 2007; van Dyke 2005) consider only a fraction of those systems published internationally, without proposing a rationale for inclusion or exclusion. There appears to be a bias of selection, however, towards HERSs published in English. Cheng and Liu (2008), as well as Dill and Soo (2005), include only HERSs published in English, and van Dyke (2005) includes two non English-HERSs in a total of ten. Usher and Savino (2006, 2007) analyze the most diverse set of HERSs, but the majority of them are also published in English. Such an emphasis on HERSs published in English does not reconcile with the international territory of HERSs. Salmi and Saroyan (2007) found the majority of HERSs to be published in non-English speaking countries, targeted to their particular domestic higher education market. Further, Hendel and Stolz (2008) found the overwhelming majority of European HERSs to be published in a language other than English. Therefore, the seemingly existing consensus about the most studied ranking systems might be linguistically biased. Such a bias might lead to a skew in the judgement about international ranking practices and to disregarding promising exemplars as well as problematic trends associated with these practices.

Second, no systematic efforts of HERS-benchmarking have been added to the discourse about HERSs, a discourse that is too often free of empirical analyses (Müller-Böling 2001a). The question therefore arises: "Who ranks the university rankers?" (Enserink 2007). Cheng and Liu (2008) came closest to answering this question. They derived 14 criteria from the BPs, and applied those for evaluating the ranking practices of eight HERSs. However, their particular operationalization of the BPs did not enable the assessment of degrees of compliance with the BPs, solely assigning a 'true' or 'false' statement about whether compliance was achieved by a particular HERS regarding each of their 14 criteria.

This study addresses the two aforementioned issues by ranking 25 of the 32 HERSs published in Europe since 2003. First, a linguistic selection bias was circumvented by establishing a geographical criterion for inclusion. Europe was chosen as this geographical

criterion based on the observation that "the discussion of rankings and the related discussion of typologies in Europe is promising" (Marginson and van der Wende 2007, p. 326). We, therefore, propose that by choosing European HERSs as research objects, conclusions can be drawn that are especially enlightening from an international comparative perspective. After all, European HERSs have to position themselves in an increasingly international market, characterized by the Bologna Agreement, high student and faculty mobility due to programs of the European Union (e.g., Erasmus), and the geographical closeness of nations that have the stated goal to attract the best minds of the world.

Second, the goal of this study is to assess the ranking practices used by European HERSs according to objective and quantitative measures. The *Berlin Principles on Ranking of Higher Education Institutions* (IREG 2006: BPs) are used as standards from which these measures were derived (Table 1). By selecting the BPs as appropriate standards for assessing HERSs, the authors follow the example of Cheng and Liu (2008), and agree with the observation that the BPs are "useful for the improvement and evaluation of rankings around the world" (IHEP 2007, p. 3). The following section will now establish a rationale for translating BPs into objective and quantifiable criteria. By doing so, a rationale for the general usefulness of the BPs for ranking HERSs will be given. Further, the method for selecting the European HERSs under study will be explained.

Method

Quantifying the Berlin Principles

The need for a set of best practices to guide the design of HERSs has been broadly stated in the literature (Altbach 2006; Clarke 2002; Merisotis and Sadlak 2005; Proulx 2006). In 2006, members of the International Ranking Expert Group (IREG), founded in 2004 by the UNESCO European Center for Higher Education (UNESCO-CEPES), established a "set of principles of quality and good practice" (IREG 2006, p. 1) in order to create a framework "that ultimately will lead to a system of continuous improvement and refinement of the methodologies used to conduct" (p. 1) HERSs. International experts from Asia, Europe and North America, representing higher education institutions, HERSs, governmental and non-governmental agencies as well as research institutes and foundations, participated in establishing the principles in Berlin, Germany (UNESCO-CEPES 2006).

In order to use these BPs for the purpose of benchmarking and ranking European HERSs, 10 of the 16 BPs were quantified into objective and measurable criteria. Thus, for each BP, one or two scales was established, leading to a total of 14 measurement scales used for benchmarking and ranking European HERSs. Each scale was scored on a scale from 5 to 1; a score of 5 indicating excellent congruity of a particular ranking practice with a particular BP; a score of 4 good congruity; a score of 3 fair congruity; a score of 2 poor congruity, and a score of 1 no congruity (see Table 2). Six BPs were excluded for the purpose of this study. The rationales regarding how each of the scales was constructed are provided in the following, as well as the reasoning for excluding six BPs.

BP 1 was not included for the purpose of benchmarking and ranking European HERSs. This particular principle refers to the problem that the lack of alternative methods for comparing higher education institutions often unintentionally legitimizes the use of HERSs for exclusively assessing the overall quality of such institutions (Clarke 2002). BP 1 agrees with the statement that "no ranking system can accurately measure quality" (Merisotis

Table 1 The Berlin Principles on Ranking of Higher Education Institutions

Number	Principle statement
1	*Be one of a number of diverse approaches to the assessment of higher education inputs, processes, and outputs.* Rankings can provide comparative information and improved understanding of higher education, but should not be the main method for assessing what higher education is and does. Rankings provide a market-based perspective that can complement the work of government, accrediting authorities, and independent review agencies
2	*Be clear about their purpose and their target groups.* Rankings have to be designed with due regard to their purpose. Indicators designed to meet a particular objective or to inform one target group may not be adequate for different purposes or target groups
3	*Recognize the diversity of institutions and take the different missions and goals of institutions into account.* Quality measures for research-oriented institutions, for example, are quite different from those that are appropriate for institutions that provide broad access to underserved communities. Institutions that are being ranked and the experts that inform the ranking process should be consulted often
4	*Provide clarity about the range of information sources for rankings and the message each source generates.* The relevance of ranking results depends on the audiences receiving the information and the sources of that information (such as databases, students, professors, employers). Good practice would be to combine the different perspectives provided by those sources in order to get a more complete view of each higher education institution included in the ranking
5	*Specify the linguistic, cultural, economic, and historical contexts of the educational system being ranked.* International rankings in particular should be aware of possible biases and be precise about their objective. Not all nations or systems share the same values and beliefs about what constitutes "quality" in tertiary institutions, and ranking systems should not be devised to force such comparisons
6	*Be transparent regarding the methodology used for creating the rankings.* The choice of methods used to prepare rankings should be clear and unambiguous. This transparency should include the calculation of indicators as well as the origin of data
7	*Choose indicators according to their relevance and validity.* The choice of data should be grounded in recognition of the ability of each measure to represent quality and academic and institutional strengths, and not availability of data. Be clear about why measures were included and what they are meant to represent
8	*Measure outcomes in preference to inputs whenever possible.* Data on inputs are relevant as they reflect the general condition of a given establishment and are more frequently available. Measures of outcomes provide a more accurate assessment of the standing and/or quality of a given institution or program, and compilers of rankings should ensure that an appropriate balance is achieved
9	*Make the weights assigned to different indicators (if used) prominent and limit changes to them.* Changes in weights make it difficult for consumers to discern whether an institution's or program's status changed in the rankings due to an inherent difference or due to a methodological change
10	*Pay due attention to ethical standards and to the good practice recommendations articulated in these Principles.* In order to assure the credibility of each ranking, those responsible for collecting and using data and undertaking on-site visits should be as objective and impartial as possible
11	*Use audited and verifiable data whenever possible.* Such data have several advantages, including the fact that they have been accepted by institutions and that they are comparable and compatible across institutions
12	*Include data that are collected with proper procedures for scientific data collection.* Data collected from an unrepresentative or skewed subset of students, faculty, or other parties may not accurately represent an institution or program and should be excluded
13	*Apply measures of quality assurance to ranking processes themselves.* These processes should take note of the expertise that is being applied to evaluate institutions and use this knowledge to evaluate the ranking itself. Rankings should be learning systems continuously utilizing this expertise to develop methodology

Table 1 continued

Number	Principle statement
14	*Apply organizational measures that enhance the credibility of rankings.* These measures could include advisory or even supervisory bodies, preferably with some international participation
15	*Provide consumers with a clear understanding of all the factors used to develop a ranking, and offer them a choice in how rankings are displayed.* This way, the users of rankings would have a better understanding of the indicators that are used to rank institutions or programs. In addition, they should have some opportunity to make their own decisions about how these indicators should be weighted
16	*Be compiled in a way that eliminates or reduces error in original data, and be organized and published in a way that errors and faults can be corrected.* Institutions and the public should be informed about errors that have occurred

Note: From "Berlin Principles on Ranking of Higher Education Institutions", by IREG 2006, Berlin, Germany: International Ranking Expert Group

2002, p. 478), but rather provides one of many possible views on the quality of higher education institutions. After all, HERSs have to reduce complexity when assessing institutional quality in order to provide information consumers can understand. This specific emphasis of BP 1 on the limited role that HERSs can and should play in assessing the 'true' quality of higher education institutions led to the exclusion of BP 1. This study uniquely examines the quality of ranking practices, not the specific role that HERSs play when assessing educational and institutional quality in concert with other evaluative efforts.

BP 2 combines the two aspects of clarity of purpose and clarity of the target group into one single principle. By making this combination, the authors of the BPs propose a close relationship between purpose and target group when designing a HERS. The necessity of such closeness has also been proposed in the research literature; whereas the purpose of a particular HERSs has been seen as directly derived from the target group. McDonough et al. (1998) state that the core purpose of a HERS should be "to place rankings in a broader decision-making context for students, parents, and professional advisors who are relying on them" (p. 533). Lebherz et al. (2005) agree with this proposition by suggesting that HERSs have "to synthesize and present information in order to alleviate the decision-making process of the targeted audience" (p. 188, translated by first author). Therefore, the establishment of a criterion for BP 2 not only had to reflect the dual aspect of clarity of purpose and clarity of the target group, but also how the purpose of a particular HERS was derived from the target group. As a result, the authors chose to create two sub-scales, whereas the sub-scale for purpose directly builds on the sub-scale for the target group. Sub-scale 2A assesses the degree to which a particular HERS shows clarity about which target group it is serving and which impact(s) it is having, including how this impact on the target group then impacts the broader higher education system. Sub-scale 2B assesses the degree to which this clarity of the target group translates into the clarity of purpose for serving the target group. It is hereby proposed that a HERS serves this purpose better, and increases the relevance of the ranking results for the target group, when data directly collected from the target group is used for the computation of the overall ranking score. Also, the unique message this data reveals should be openly accessible to the target group, rather than possibly becoming aggregated with other data sources when an overall score is computed. Further, it is proposed that the relevance of the purpose is further increased if the target group can easily orient itself within the mechanisms of the ranking scheme. Therefore, sub-scale 2B includes an assessment of the degree to which a target group can identify itself in

Table 2 Scales of assessment for 10 of the 16 Berlin Principles on Ranking of Higher Education Institutions

BP		Scale
2	A	**5**—Information given about all of the following: Target group; intended impact on target group; scope of information provided to target group; impact that information provided to target group might have beyond direct ranking system-target group interaction **4**—Information given about three of the above; **3**—Two of the above; **2**—One of the above; **1**—None of the above
	B	**5**—All of the following dispositions can be found: Data directly collected from the target group (i.e. through surveying) is used when computing an overall ranking score; the distinct message that this target group data reveals is identifiable when an overall ranking result is presented; a score of at least 4 has been achieved for Principle 3; an overall score of at least 4 has been achieved for Principle 15 **4**—Three of the above dispositions can be found; **3**—Two of the above; **2**—One of the above; **1**—None of the above
3		**5**—Separate rankings/scores according to all four of the following differentiations are computed: Type of higher education institution (i.e. research university; teaching college); program area (i.e. social sciences; engineering); level of program (i.e. graduate; undergraduate; doctoral); organizational feature of institution (public vs. private) **4**—For three of the above; **3**—For two of the above; **2**—For one of the above; **1**—For none of the above
4	A	**5**—All four of the following sources of information used in the data collection process: Bibliometrics, students, professors, employers **4**—Three of the above used; **3**—Two of the above; **2**—One of the above; **1**—None of the above/ remains unclear
	B	**5**—The particular messages generated by all four different sources of information mentioned for 4A are made explicit in the ranking **4**—The messages of three of the sources mentioned for 4A are made explicit; **3**—Two; **2**—One; **1**—Messages remain unclear
6	A	**5**—Methodology given allows for recalculation of overall score on part of consumer **4**—All of the following methodological issues explained: Entity that has been measured for assessing particular indicator;process of aggregating an indicator score; process of weighting indicator scores in order to calculate ranking score **3**—Two of the methodological issues explained (as outlined for a score of 4); **2**—One of the above as outlined for a score of 4; **1**—None of the above as outlined for a score of 4
	B	**5**—Information given about origin of data for all indicators used **4**—For at least 75% of indicators used; **3**—For at least 50%; **2**—For at least 25%; **1**—for less than 25%, no assessment possible because issue not addressed by ranking author(s)
7	A	**5**—At least one indicator for all five dimensions: Funding; human resources, academic outcomes; third mission and governance **4**—At least one indicator for four of the five dimensions; **3**—Three of the five dimensions; **2**—Two of the five dimensions; **1**—One of the five dimensions
	B[a]	**5**—All indicators are valid **4**—at least 75% are valid; **3**—at least 50%; **2**—at least 25%; **1**—less than 25%, or no assessment possible because issue not addressed by ranking author(s)
8		**5**—100% of overall score calculated on basis of output measures **4**—At least 75% of the overall score; **3**—at least 50%; **2**—at least 25%; **1**—less than 25%, no assessment possible because issue not addressed by ranking author(s)
9	A	**5**—All weights assigned to the indicators used in the ranking are clearly stated **4**—Weights assigned are stated clearly for all sets of indicators used (i.e. category), but not for each indicator individually **3**—For at least 75% of sets of indicators (i.e. category) and/or indicators used; **2**—for at least 50%; **1**—for less than 50%, no assessment possible because issue not addressed by ranking author(s)
	B[b]	**5**—No indicators used have been changed relative to the ranking most recently published **4**—Less than 5% of indicators used; **3**—less than 10%; **2**—less than 15%; **1**—more than 15%

Table 2 continued

BP		Scale
11		**5**—100% of the overall score calculated on basis of data gathered by third party institutions (collected on all universities included in the ranking) **4**—At least 75% of the overall score; **3**—At least 50%; **2**—At least 25%; **1**—Less than 25%, no assessment possible because issue not addressed by ranking author(s)
12	A	**5**—Description of data sampling procedures given for all indicators included in the ranking **4**—Given for at least 75% of indicators; **3**—For at least 50%; **2**—For at least 25%; **1**—For less than 25%, no assessment possible because issue not addressed by ranking author(s)
	B[c]	**5**—Sampling procedure sound for all indicators included in the ranking **4**—For at least 75% of indicators; **3**—For at least 50%; **2**—For at least 25%; **1**—For less than 25%, no assessment possible because issue not addressed by ranking author(s)
15	A	**5**—Definitions for all indicators used in the ranking are given **4**—For at least 75% of indicators; **3**—For at least 50%; **2**—For at least 25%; **1**—For less than 25%, no assessment possible because issue not addressed by ranking author(s)
	B	**5**—Consumer can alter both weights assigned to indicators and the selection of indicators used in computing the ranking results for all of the following three levels: institutional, departmental, program **4**—alter both weights assigned to indicators and the selection of indicators for two of the three levels; **3**—for one of the three levels **2**—Consumer can alter either weights assigned to indicators or the selection of indicators used in computing the ranking results for at least one of the following three levels: institutional, departmental, program **1**—Consumer has no choice in how ranking is displayed

Note: [a] Sub-scale 7B was not assigned a score due to methodological reasons (see "Method" section); [b] sub-scale 9B was not assigned a score, because a longitudinal design would have been needed; however, such a design is outside the scope of this particular study; [c] sub-scale 12B was not assigned a score due to the lack of access to according data. The overall scores for the three Berlin Principles marked with either [a,b,] or [c] consist only of the score given for the respective part A of each of the principles

a more detailed way according to institutional/departmental/program affiliation (reference to BP 3). Sub-scale 2B further includes an assessment of the ease of understanding what has been measured, and of choosing the relevance of these measurements for the target group's own purpose (reference to BP 15).

BP 3 asks for the recognition of the diversity of institutions and the diversity of their missions and goals. This principle directly reflects the argument made in the literature that differences in institutional focus and institutional mission need to be considered (Merisotis 2002). This has especially become translated into the demand for differentiating between rankings according to different institutional types, departments, and programs (Clarke 2002). Therefore, the scale established for BP 3 assesses the degree to which a particular HERS offers separate rankings and/or ranking scores according to these institutional, departmental or program differentiations.

BP 4 again combines two aspects into one principle: the clarity of the range of information sources used for computing a HERS, and the clarity of the message each of these information sources creates. The authors chose to establish two sub-scales for reflecting this dual aspect of this principle. Regarding the clarity of the range of information sources, BP 4 refers to databases, students, professors, and employers as possible information sources. Therefore, sub-scale 4A assesses whether a particular HERS uses one or more of these information sources. Congruent with the explanations given by the BPs, it is hereby proposed that a ranking score based on a broader set of perspectives results in a more complete view of a particular higher education institution. Special attention had to be given

to the fact that HERSs sometimes do not collect data themselves, but use data compiled by third-party institutions. If this is the case for a particular HERS, sub-scale 4A assesses the degree to which data provided by third-party institutions was established based on one or more of these four information sources. Third-party institutions are herby understood as all institutions, except higher education institutions and HERSs, that grant open access to the data that has become used by a particular HERS (this understanding of a third-party institution also applies to BP 11). Sub-scale 4B then assesses whether the messages that the information sources generate become clear and explicit for each of the information sources separately.

BP 5 was excluded for the purpose of this study. It refers to the special obligation of HERSs that rank institutions/departments/programs internationally to consider linguistic, cultural, economic, and historical differences of educational systems. However, only three European HERSs rank internationally (*The Leiden Ranking*; *Champions League*; *World University Ranking*). Therefore, BP 5 would only have had relevance for a great minority of the European HERSs. As a result, the decision was made to exclude BP 5 for the purposes of this study.

BP 6 asks for transparency of the methodology used for creating a HERS, and is a reflection of the plethora of articles that strongly demand such transparency of the methodology (Marginson and van der Wende 2007; Sadlak et al. 2008). The authors of the BPs hereby considered transparency of methodology to contain two aspects: the transparency of the method of how indicators and ranking results were calculated, as well as the transparency of the origin of the data. Therefore, two sub-scales have been created for BP 6. Sub-scale 6A refers to the transparency of calculating indicators and ranking results. The proposition is made that the possibility of recalculating the overall ranking score(s) would prove the best transparency achievable. In case this is not possible, an assessment was made whether readers of a particular HERS can follow the process of deriving the data, creating indicator scores, and weighting those scores in order to calculate an overall ranking score. Sub-scale 6B refers to the origin of the data. The proposition is made that greater knowledge about the information sources leads to greater transparency regarding the origin of the data.

In order to apply the scales for BP 6, but also those of BP 7, 12, and 15, the use of the term 'indicator' had to be specified. Unfortunately, the use of this term is understood quite differently by the various HERSs (see also: Hendel and Stolz 2008), sometimes referring to a set of sub-indicators framing a particular indicator, sometimes rather meaning a category (e.g., teaching quality) that is assessed by using various measures. In this particular study, the smallest measurable entity was considered as an indicator. Therefore, no macro-indicators or categories were considered as indicators, even though particular HERSs might have used this particular term. As a consequence, more indicators might have been attributed to a HERS by the authors of this study than the particular HERS might proclaim. This adaptation was necessary to make the HERSs comparable to each other, especially when applying BPs 6, 7, 12, and 15.

BP 7 refers to the need for relevant and valid indicators. The discussions about which indicators are relevant and valid as proxy measures for university quality are still contested (Altbach 2006; Bellon 2007; Bowden 2000; Brooks 2005; Federkeil 2008; Hazelkorn 2007), also hinting at the underlying disagreements about what constitutes a quality university (Birnbaum 2006). These contestations and disagreements in the field of HERS were only partly resolved by the authors of this study. As a result, only scores for the relevance of indicators of a particular HERS were created, even though two sub-scales were

established for assessing both relevance and validity of indicators, reflecting the dual aspect of BP 7.

Sub-scale 7A assesses the degree to which a particular set of indicators of a particular HERS can be considered as relevant for representing quality as well as academic and institutional strength. For making this assessment, the evaluative framework established by the PRIME Network of Excellence in their Observatory of the European University (PRIME 2006) was chosen. This evaluative framework proposes five dimensions with which higher education institutions should be analyzed: Funding, human resources, academic outcomes, third-mission dimensions and governance. The funding dimension refers to the budget and the sources of funding, such as the total amount of grants received. The human resources dimension includes all aspects regarding researchers, staff and students. Publications and citations are part of the academic outcomes dimensions. The third mission dimension refers to the relationship between higher education institutions and non-academic partners, including employment rates of graduates, number of patents and consultancy work. Finally, the governance dimension addresses management issues of higher education institutions and "the process by which the university converts its inputs (funding and human resources) into research outputs (academic outcomes and third mission)" (Bellon 2007, p. 136). The authors propose that this evaluative framework allows for a broad view on higher education institutions, a view that reflects their quality as well as academic and institutional strength in a relevant way. The authors further propose that the relevance of a ranking result is stronger the more comprehensive these five dimensions are assessed by the indicators of a particular HERS.

Sub-scale 7B assesses the degree to which indicators can be considered as valid. Given the contested and divergent perspectives on validity in the research field, the authors did not strive for resolving this contestation and divergence at this point; the BPs themselves do neither give guidance on how this issue could be resolved. Therefore, establishing an understanding of what constitutes a valid indicator, or a valid set of indicators, is still such a great task for the field of higher education, that the authors of this study felt unable to assign scores for sub-scale 7B. Given the contestations and disagreements in the field, the authors were also unsure about whether even a panel of HERS-experts could at this point assess the validity of indicators, without coming to widely divergent judgements. Nevertheless, the issue of valid indicators must be addressed in order to make the ranking of higher education institutions not only valid, but ultimately more credible. The authors are aware that by not scoring European HERSs, the meaningfulness of the results of this study might be lessened. However, choosing not to score the sub-scale for validity ultimately constitutes a request for assistance that should be considered within the field of higher education. By identifying valid indicators, the field of higher education might be able to reverse the trend of having become considerably shaped by results of HERSs, and rather actively shape HERSs as tools for assessing and improving higher education institutions, in concert with other means through which such assessments and improvements can be achieved. The International Ranking Expert Group might be the appropriate forum for this task. However, this expert group has so far provided neither a theoretical rationale nor a practical guideline for achieving validity. Providing both rationales and guidelines might be the most important challenge for this expert group ahead.

Principle 8 stresses the importance of using outcome indicators, rather than the current emphasis on input indicators. This demand for outcome indicators agrees with the notion as presented in the literature that output measures are preferable to input and throughput measures (Müller-Böling 2001b; Terenzini and Pascarella 1994). However, BP 8 is also a reaction to the fact that "only a few valid, reliable and really comparable indicators of

outputs exist" (Federkeil 2008, p. 229). In order to assess BP 8, Hendel and Stolz' (2008) model for categorizing indicators has been used. Thus, every single indicator became attributed to one of the sub-categories of the meta-categories Input, Throughput, Output, or Research. According to this model, Input-indicators encompass indicators measuring the beginning characteristics of students, faculty, and staff (e.g. entry exams, teaching faculty with PhDs), financial resources, program design components, and amenities (e.g. library). Throughput indicators measure management systems, teaching quality, student support systems, and program/course proceedings. Output-indicators encompass indicators assessing such aspects as career prospects, reputation, degrees awarded, and student satisfaction. It is hereby assumed that the use of the term 'Output' agrees with the understanding of the term 'Outcome' as used by the BPs. Finally, research indicators measure publications, citations, and patents. After having attributed indicators to the four categories, the proportion of Output-indicators as the basis for an overall ranking score became obvious. The assessment of the proportion was usually made by adding the relative weights of all indicators in the Output category. However, if no weights were assigned to the indicators (i.e., for the HERSs of CHE, swissUp and AQA) or were not given by a particular HERS (i.e., the HERS of Spiegel), the amount of Output-indicators was taken as proportion of the total number of indicators used.

BP 9 encompasses the dual aspect of making the weights assigned to indicators prominent, and to limit changes to these weights. This has become translated into two subscales. Sub-scale 9A assesses the degree to which indicators and/or categories of indicators are assigned a clearly stated weight; or whether it was clearly stated that no such weights were used for the purpose of a particular HERS. Sub-scale 9B then assesses the degree to which these weights assigned to indicators have been changed in comparison to the according HERS most recently published. However, this particular study did not adopt a longitudinal design that would have been needed for scoring sub-scale 9B. Therefore, no score was attributed to sub-scale 9B.

BP 10 is excluded for the purpose of this study. BP 10 refers to the issue of ethical standards when compiling HERSs, mainly in the form of the objectivity and impartiality of those responsible for a particular HERS. This particular principle has been excluded based on two reasons. First, none of the rankings provides concrete information that would have allowed for measuring the degree of compliance with ethical standards. Second, the decision for exclusion has also been made based on the proposition that even though HERSs have the responsibility for ethical conduct, the process of ensuring compliance with ethical standards might have to be initiated by an external supervising body. It would be difficult to measure and credibly validate ethical conduct for HERSs themselves. Therefore, ensuring ethical conduct on the part of HERSs has not only to do with conducting internal mechanism ethically, but also with external frameworks, procedures, and supervising tasks that ensure such ethical conduct. This crucial role of external stakeholders, frameworks, and processes, which ultimately cannot directly be determined by a particular HERS, also has contributed to the decision for not using BP 10 when assessing internal ranking practices of HERSs.

BP 11 asks for the use of audited and verifiable data. This principle has become translated into an assessment of the degree to which HERSs use data provided by third party institutions (see BP 4 for definition). This translation is based on the proposition that specialized third party institutions, like ministries of education, governmental statistical offices or publication/citation databases, are much more likely than HERSs to be able to provide the resources and special expertise for auditing and verifying data. Further, third party institutions also most often grant open access to the data that HERSs ultimately use.

Thus, a mechanism for open audit and verification is often built into third party institution data. As a result, the use of data provided by third-party institutions has become a proxy for assessing the use of audited and verifiable data.

BP 12 refers to the need for collecting data using appropriate scientific techniques. The challenge of establishing a scale for assessing this particular principle is resolved in a similar way as for BP 7. Sub-scale 12A assesses the degree to which data sampling procedures have been described by a particular HERS for the indicators used. Therefore, this first sub-scale undertakes a purely quantitative judgment about whether the set of indicators conforms to a particular requirement, similar to sub-scale 7A. Sub-scale 12B then qualitatively assesses whether these procedures can be considered scientifically sound. Unfortunately, even though descriptions about data sampling procedures were given in various degrees by the HERSs under study, those descriptions generally did not provide enough information to make a scientifically valid judgment about the scientific rigor of these procedures. Therefore, a scoring for sub-scale 12B was not undertaken. This inability to score BP 12B should encourage HERSs to give more detailed insights into processes of data collection.

BP 13 and BP 14 were not used for ranking HERSs. The vast majority of HERSs did neither provide enough information for assessing whether a particular HERS applies measures of quality assurance to the process of computing a particular HERS, nor for assessing the application of internal organizational measures for enhancing ranking practices. As a consequence, an evaluation of the degree to which HERSs use measures as demanded by BPs 13 and 14 did not occur. This lack of assessment of BPs 13 and 14 again might lessen the meaningfulness of the ranking of HERSs as undertaken in this study. However, this lack of assessment nevertheless provides an overarching negative judgment of current practices of European HERSs that either have not established such measures or do not feel the need to provide information about the way how these measures have become applied for ensuring the quality of a particular HERS.

BP 15 asks HERSs to provide both understanding and choice to the consumer. First, the consumer should be able to get a clear understanding of the indicators used for the computation of a particular HERS. The literature notes that many HERSs produce a single score, making it "difficult for students to distinguish among institutions based on the characteristics they find most important" (IHEP 2007, p. 2). Second, the consumer also should get the chance to decide which indicators are important and how important they are. Thus, HERSs that can be "re-weighted" (Longden and Yorke 2006, p. 15) have been recommended in discussions about HERSs. Two sub-scales were created for BP 15. Sub-scale 15A is based on the proposition that clear definitions of indicators enhance the understanding of indicators sets. Therefore, sub-scale 15A assesses the degree to which definitions for the indicators used are given. Sub-scale 15B then assesses the degree to which consumers can select indicators and alter weights assigned to these indicators in order to produce ranking results that presumably serve their own particular purposes in a better way. Also, it is assessed whether this selection of indicators and indicator weights is possible on all ranking levels (institution/department/program).

BP 16 could not be included for the purpose of this study. The vast majority of HERSs did not provide enough information in order to make an assessment of BP 16 possible. Again, HERSs have to be asked for granting a more detailed access to their ranking processes.

Scoring process

The IREG (2006) did not assign degrees of importance to the 16 principles. Therefore, the authors of this paper also chose not to attribute such degrees of importance. As a result, the particular scores for each of the 10 BPs quantified for the purpose of this study receive the same weight when calculating an overall final score for a particular HERS. The final score for a particular HERS, therefore, is computed as the average of the 10 individual BP-scores. For each of the BPs 2, 4, 6, 7, 9, 12 and 15, two sub-scales were used in order to enable an assessment that more closely reflects the content of the particular BPs. The individual BP-score for these BPs was computed by taking the average of the two sub-scale scores respectively. Due to the exclusion of the sub-scales 7B, 9B, and 12B, the individual scores for BPs 7, 9 and 12 are based only on the scores of the scales for 7A, 9A and 12A respectively.

Categorizing the Berlin Principles

Also, sub-scores are presented to highlight the congruence of a particular HERS with the BPs in the realms of methodology, transparency and consumer friendliness. In agreement with Cheng and Liu (2008), the authors of this paper also propose the issues of methodology and transparency as 'latent variables' underlying the BPs; however, the aspect of consumer friendliness was added as an additional distinct third sub-score. Consumer friendliness is hereby defined as the potentiality given to each reader of a particular HERS to make rankings results more relevant, based on a clear understanding of the measurements used.

The methodology score of a particular HERS was computed by taking the average scores of the following scales: 2B, 3, 4A, 7A, 8 and 11; the transparency-score by taking the average score of 2A, 4B, 6A, 6B, 9A and 12A; the consumer friendliness-score by taking the average score of 15A and 15B. This computation makes it obvious that the individual BPs can not necessarily be attributed to one category only; some BPs rather encompass both aspects of methodology and transparency.

HERSs in Europe

All HERSs published in the geographical area of Europe in 2003 and/or after have been considered for this study. Thirty-two of those systems in each of the following 16 countries were indentified: Austria, The Czech Republic, France, Germany, Italy, Ireland, the Netherlands, Norway, Poland, Portugal, Russia, Slovakia, Spain, Sweden, Switzerland and the United Kingdom (see Table 3).

The systems were identified based on a review of the published literature, both in English and in the particular national languages, extensive Internet searches on institutional rankings within each of the countries and by establishing a network of experts regarding HERSs in Europe (mainly through e-mail). For identifying HERSs, those systems were defined as having all of the following characteristics: (1) they weight data sets collected about higher education institutions using previously established criteria; (2) they make the results openly available to a wide variety of stakeholders, but not necessarily in the format of a league table; and (3) they do not invest primarily in interpretive and/or evaluative tasks, but rather provide information so that stakeholders themselves, external to the HERSs, can engage in the process of interpreting and evaluating the results.

Table 3 Higher education ranking systems (HERSs) in Europe

Country	Name of HERS	Responsible entity[a]
Austria	Hochschulvergleich	Austrian Agency for Quality Assurance (AQA)
Czech Republic	Lidove Noviny[b]	SC&C Agency (SC&C)
Germany	Humboldt-Ranking	Alexander von Humboldt-Stiftung (Humboldt)
	HochschulRanking	Centrum für Hochschulentwicklung (CHE)
	Hochschul-Ranking	*Focus*
	Die besten Hochschulen	*Karriere*
	Studentenspiegel	*Spiegel*
France	Universités: Enquête[b]	*Vie Universitaire*
Ireland	The Sunday Times University League Table	*The Sunday Times*
Italy	La Grande Guida All'Universita'	*La Repubblica*
The Netherlands	De beste studies	*Elsevier*
	Keuzegids Hoger Onderwijs	Keuzegids Hoger Onderwijs
	The Leiden Ranking	Center for Science and Technology Studies (CSTS)
Norway	stud.mag[b]	NIFU STEP/TNS Gallup
Poland	Ranking szkól wyższych	*Perspektywy*
	Ranking wyższych uczelni	*Polityka*
	Ranking szkól wyższych	*Wprost*
Portugal	Ranking de Universidades[b]	Universidade of Aveiro
Russia	[b]	RatER
Slovakia	Ranking (VS/HEI)	Academic Ranking and Rating Agency (ARRA)
Spain	Excelencia	Centro de Investigaciones Sociologicas (CIS)
Sweden	Högskolerankingen	Sydsvenska Industri- och Handelskammaren (SIH)
	Hur mycket citeras svenska publikationer?	Vetenskapsrådet
Switzerland	Champions League	Center for Science and Technology Studies (CEST)
	Ranking swissUp	Stiftung zur Förderung der Ausbildungsqualität (swissUp)
United Kingdom	University league table[b]	*Daily Telegraph*
	The table of tables[b]	*Daily Telegraph*
	Financial Times League Table	*Financial Times*
	The Guardian University Guide	*The Guardian*
	The Sunday Times University League Table	*The Sunday Times*
	The Times Good University Guide	*The Times*
	World University Ranking	*Times Higher Education Supplement* (THES)

[a] Daily, weekly, and monthly newspapers/journals responsible for a higher education ranking system are italicized

[b] Indicates that a particular higher education ranking system has not been included in this study due to the lack of access to according data

Not included were HERSs that consider only particular types of institutions (e.g., business schools, medical schools), or only a single field of study (e.g., economics). The results of the scoring process of BP 3, and partly of BP 2, would not have been comparable to HERSs that encompass a wider set of institutions and fields, because the single-focus HERSs would necessarily have scored as inferior. Further, HERSs that have been discontinued before 2003 were excluded. According to these criteria, no HERSs could be identified in Albania, Andorra, Belarus, Belgium, Bosnia, Bulgaria, Croatia, Denmark, Estonia, Finland, Greece, Hungary, Iceland, Latvia, Lithuania, Luxemburg, Macedonia, Malta, Moldavia, Montenegro, Romania, San Marino, Serbia, Slovenia, Turkey and Ukraine. We acknowledge that we may have missed particular HERSs that would have met our criteria.

The necessary data about the HERSs have been derived from four different sources of information: (1) the online portal of a particular HERS; (2) in-print publications presenting/and or accompanying a HERS; (3) separate in-print publications about a HERS; and (4) direct contact with the ranking authors. For the following HERSs, the online portal was the sole source of data: The HERSs published by Focus and Karriere in Germany; both the Irish and British *The Sunday Times University League Table*; *La Grande Guida All'Universita'*; *The Leiden Ranking*; all three Polish HERSs; and the *Financial Times League Table*. In-print publications presenting and/or accompanying HERSs have been used mainly for the rankings by AQA (Gaertner et al. 2007) and CHE (Berghoff et al. 2007); *Humboldt-Ranking* (Alexander von Humboldt-Stiftung 2005); *Studentenspiegel* (Studentenspiegel 2006); the HERSs by Keuzegids Hoger Onderwijs and Elsevier in the Netherlands (Elsevier 2006; Steenkamp et al. 2006); for the HERSs in Slovakia (ARRA 2005); Sweden (Karlsson and Wadskog 2006; Sydsvenska Industri-och Handelskammaren 2007); Switzerland (CEST 2004; swissUp 2006a, b); *The Guardian University Guide* (Macleod 2007); *The Times Good University Guide* (O'Leary et al. 2006a); and *The World University Ranking* (O'Leary et al. 2006b). Separate publications have been considered for the Spanish HERS (de Miguel et al. 2001; de Miguel et al. 2005). Additionally, contacts with the ranking authors have been established in the case of the Keuzegids Hoger Onderwijs and Elsevier as well as for the Slovak HERS.

Native speakers were consulted for translating the materials of the Italian, Polish and Swedish HERSs (the HERS by Vetenskapsrådet was translated by a native English speaker with Swedish language skills). The remaining non-English materials were translated by the authors of this paper.

The following HERSs could not be included in the study due to the lack of access to data that would have been comprehensive enough to rank relative to the BPs: the Czech ranking system, published in the journal Lidove Noviny, compiled by the CS&C Agency; the French *Universités: Enquête* by Vie Universitaire; the Portuguese ranking by the *Universidade de Aveiro*; the Russian ranking published by RaTER; and the two British rankings by *The Daily Telegraph*. The remaining 25 European HERSs—representing 11 countries—were included in this study.

Results

Table 4 presents the results of the scoring process. A final score, as well as separate subscores for methodology, transparency and consumer friendliness, were computed. A score of 5 indicates an excellent congruence between the ranking practices of a particular HERS and the best practices proposed by the BPs; a score of 4 is interpreted as indicating good

congruence, a score of 3 as fair congruence, a score of 2 as poor and a score of 1 as no congruence. The HERSs are sorted according to their final scores. Thus, a league-table has been created which ranks the HERSs published in Europe.

The three collaborating HERSs published by CHE, AQA and swissUp occupy the first three slots in the Ranking of Rankings, due mainly to their high scores in the areas of transparency and consumer friendliness, but also due to their above-average scores in the methodology principles. Seven HERSs then rank closely together, with final scores well above the three-point threshold with scores between 3.30 and 3.50. The HERSs published by CSTS and Elsevier score just above three points. More than half of the HERSs score lower than three points, with a lowest score of 2.10 for the HERS published by the German magazine *Karriere*.

No HERS published in Europe achieves good overall congruence with the best practices proposed by the BPs. The CHE-ranking comes closest with a final score of 3.90. The fact that no HERS achieves good overall congruence is due mainly to the finding that the HERSs score low on methodological congruence. Both the best sub-score for methodology (3.33) as well as the median for this particular sub-score across all 25 HERSs (2.33) are surprisingly low. This relatively weak performance in the area of methodology did not allow any HERS to rise to a final score of at least 4.00. Thus, mainly methodological improvements are needed in order to create higher congruence of ranking practices of current European HERSs with the best practices as proposed by the BPs.

The greatest overall congruence of European HERSs with the BPs has been achieved in the area of transparency. Eleven HERSs achieve scores of 4 and higher in this realm, and only 24% of them fall under the 3-point threshold (in comparison to 80% regarding the methodology sub-score). However, both the range between the lowest and highest transparency sub-score as well as the according standard deviation (range = 2.83; SD = 0.75) is higher than for the methodology sub-score (range = 1.83; SD = 0.48). Even though many HERSs excelled in accounting for good transparency practices, other HERS are still performing well below the proposed standards. Then again, no single HERS excelled in the area of methodology, with basically all HERSs having to make improvements in this area.

Regarding consumer friendliness, most rankings achieved a score of 3, with three scoring higher and six scoring lower. A better overall performance regarding consumer friendliness was mainly precluded by the fact that only the three cooperating HERSs of CHE, AQA and swissUp allow the consumer to rank and weight indicators according to individual preference. Thus, the remaining 22 HERSs received a score of 1 for the scale 15B; however, the overall performance regarding scale 15A was good with an average score of 4.10 across all HERSs. Therefore, HERSs generally should integrate mechanisms that allow consumers to select indicators and indicator weights in order to increase congruence with the BPs in the area of consumer friendliness. Overall, the lack of congruence of European HERSs with the BPs in the realm of methodology and consumer friendliness becomes obvious, as did the relatively better results regarding the standards of transparency.

Discussion

After having described the more obvious results regarding methodology, transparency and consumer friendliness, the purpose of this section is to provide some deeper observations that can be made based on the Ranking of Rankings. We show that there are signs of relationships between the governance structure of the entity publishing a particular HERS

as well as the country in which it is published, and the position of this particular HERS in the Ranking of Rankings.

Only one of the six best-scoring HERSs is published by a commercial news-outlet (the weekly magazine *Focus* in Germany in fifth place). Even though the majority of HERSs considered for the scoring process are published by daily, weekly and monthly papers and/ or magazines, the HERSs that are compiled by non-commercial research entities seem to perform better regarding their congruence with the BPs. However, exceptions to this finding are the HERSs by CEST (rank 10), CSTS (rank 11), Vetenskapsrådet (rank 13), SIH (rank 16) and CIS (rank 21). Nevertheless, the rather controversial question can be posed of whether HERSs are more appropriately administered by specialized non-commercial entities rather than by for-profit news-outlets. Answering this question is outside the scope of this study. However, future research might be able to provide an answer by comparing even more HERSs, using the assessment scale in Table 2.

A second interesting finding is that 8 of the 12 most highly ranked HERSs are published either in German-speaking countries or in the Netherlands. On first sight, the ranking position of a particular HERS does not seem to be attributed to the country in which it is published and/or the particular higher education system in which it is embedded. This finding is in agreement with previous studies that found no holistic categorizations that could be made based on the country in which HERSs are published (Hendel and Stolz 2008; Usher and Savino 2006, 2007). However, Hendel and Stolz (2008) found that HERSs published in the Netherlands and German-speaking countries utilize distinct ranking practices. Specifically, the indicator scores used in HERSs published in the German-speaking countries as well as in the Netherlands are computed with data collected through self-administered student surveys (with the exception of HERSs by *Focus* and CSTS). This is especially notable due to the fact that no other HERS published in another European country collects any data using this method. Also, throughput-indicators are dominant in two Dutch HERSs (the one-indicator *Leiden Ranking* assesses only the category research); all other HERSs emphasize the use of input- and output-indicators and largely neglect throughput-indicators. Further, only HERSs in the German-speaking countries offer the possibility for the consumer to select indicators and indicator weights according to their individual preferences.

Arguably, two of the three distinct ranking practices in German-speaking countries and the Netherlands mentioned above are not even reflected by any scale used for the scoring process of this study, and the aspect of choice regarding indicators and indicator weights is only reflected by the sub-scale 15B. Therefore, we posit that their influence is not broadly reflected in the scoring results. Nevertheless, this particular observation might very well be another example of the notion of a "dark matter" (Usher and Savino 2007, p. 13) within HERSs. Usher and Savino (2007) state that this kind of unexplained force exerts "a gravitational pull on all ranking schemes such that certain institutions or types of institutions (the Harvards, Oxford and Tsinghuas of the world) rise to the top regardless of the specific indicators or weightings used" (p. 13). One might ask whether such a particular force of unknown nature is also present in the Ranking of Rankings. Investigating whether an underlying (latent) variable accounts both for the partially distinct ranking practices in German speaking countries and the Netherlands and the relatively good performance of HERSs published in these countries relative to the BPs is a worthwhile question to analyze. Such an investigation might unearth so far unmeasured political or cultural variables that may influence the observed results; going back to de Wit's (2003) notion that a true understanding of the "complex whole of formal and informal decision-making rules and procedures [that take] the form of interrelated networks based on variable functional and

Table 4 Ranking of higher education ranking systems (HERSs) in Europe

Country	Name of HERS	Responsible entity	Methodology score[a]	Transparency score[b]	Consumer friendliness score[c]	Final score[d]
Germany	HochschulRanking	CHE	3.33	4.67	4.50	**3.90**
Austria	Hochschulvergleich	AQA	3.00	4.67	4.50	**3.75**
Switzerland	Ranking swissUp	swissUp	3.17	4.50	4.50	**3.75**
Germany	Hochschul-Ranking	Focus	3.33	3.33	3.00	**3.50**
The Netherlands	Keuzegids Hoger Onderwijs	KHO	3.00	4.33	1.00	**3.45**
Slovakia	Ranking (VS/HEI)	ARRA	2.67	4.00	3.00	**3.45**
United Kingdom	The Guardian University Guide	The Guardian	2.50	4.17	3.00	**3.45**
United Kingdom	The Times Good University Guide	The Times	2.67	4.00	3.00	**3.45**
Italy	La Grande Guida All'Universita'	La Repubblica	2.33	4.00	3.00	**3.30**
Switzerland	Champions League	CEST	2.50	4.00	3.00	**3.30**
The Netherlands	The Leiden Ranking	CSTS	2.17	3.83	3.00	**3.10**
The Netherlands	De beste studies	Elsevier	2.67	4.00	1.00	**3.05**
Sweden	Hur mycket citeras svenska publikationer?	Vetenskapsrådet	2.00	3.67	3.00	**2.95**
United Kingdom	The Sunday Times University League Table	The Sunday Times	2.33	3.33	2.50	**2.90**
Ireland	The Sunday Times University League Table	The Sunday Times	1.83	3.50	3.00	**2.85**
Sweden	Högskolerankingen	SIH	2.00	3.67	3.00	**2.85**
Germany	Humboldt-Ranking	Humboldt	1.50	4.17	3.00	**2.80**
Poland	Ranking szkól wyższych	Perspektywy	2.17	3.00	3.00	**2.60**
Poland	Ranking szkól wyższych	Wprost	2.17	2.83	1.50	**2.60**
United Kingdom	Financial Times League Table	Financial Times	2.17	2.67	3.00	**2.60**
Spain	Excelencia	CIS	2.50	1.83	3.00	**2.55**
Germany	Studentenspiegel	Spiegel	1.83	3.50	3.00	**2.50**
Poland	Ranking wyższych uczelni	Polityka	2.33	2.83	1.00	**2.50**

Table 4 continued

Country	Name of HERS	Responsible entity	Methodology score[a]	Transparency score[b]	Consumer friendliness score[c]	Final score[d]
United Kingdom	World University Ranking	THES	1.83	2.50	2.00	**2.25**
Germany	Die besten Hochschulen	Karriere	2.33	2.33	1.00	**2.10**
Median			**2.33**	**3.67**	**3.00**	**2.95**

Note: HERSs are sorted according to their highest final score. [a] The methodology score was calculated as the average score for scales 2B, 3, 4A, 7A, 8 and 11; [b] the transparency score as the average score for scales 2A, 4B, 6A, 6B, 9A and 12A; [c] the consumer friendliness score as the average score for scales 15A and 15B. [d] The final score was computed as the average score of the respective overall scores of BPs 2, 3, 4, 6, 7, 8, 9, 11, 12, 15 (see Table 2 for scales used)

territorial connections"(de Wit 2003, p. 176) is needed. Such an understanding might lead to a better ability to concisely measure the quality of higher education institutions within their particular political or cultural contexts.

In the end, it remains outside the scope of this study to propose definitive answers to the still open question of whether differences in the results of HERSs go back to "largely a statistical fluke" (Usher and Savino 2006, p. 3) or are truly "attributable to differences in aim, higher education systems and cultures, and the availability and reliability of data" (van Dyke 2005, p. 103). The observations made in this study can suggest only the existence of underlying systematic differences. However, these observations are rather speculative in nature. Further research is needed to both substantiate these observations as well as to identify variables that account for particular ranking practices and their congruence with the BPs.

Conclusion

The purpose of this paper was to benchmark and rank the European HERSs according to their congruence with the BPs. We have shown that no HERS achieves good overall congruence with the BPs, mainly due to the finding that ranking practices in the realm of methodology did not conform with the best practices as proposed by the BPs, which is also true in the area of consumer friendliness. However, almost half of the HERSs achieved good congruence regarding transparency, and only 24% of the HERSs fell under the three-point threshold indicating "fair" congruence in that regard. It therefore has been suggested that HERSs especially have to make up ground in the realm of methodology and in the area of providing choices to their consumers.

Also, speculations about the relationship between governance structures of the entities publishing HERSs as well as the country in which they are published with their position in the Rankings of Rankings have been made. These speculations hinted to the notion of (an) underlying variable(s) that account(s) for particular ranking practices. Further research is needed to establish an understanding of what good ranking practices are fundamentally made of and how they come about. A better understanding in this realm could explain possible tensions between ranking practices actually applied by HERSs within their political, cultural, and/or educational contexts and the best practices as proposed by the BPs.

Hopefully, the results of this study, as well as the fact that a quantitative assessment tool was introduced for the first time, will stimulate improving HERS through benchmarking with existing exemplary models. Further, the quantitative assessment tool that evolved promises to be useful in assisting to explain relative stability or change of higher education institutions in HERSs; as well as in helping to explain resource allocation pressures within higher education institutions.

Despite this hope and promise, the authors are aware that this study, also due to its' limitations, can only be one small contribution in the greater effort of enhancing practices of HERSs. Even though many scales were derived from the BPs in a rather straightforward manner, and should be uncontroversial, others will remain hotly debated within the field. As a consequence, similar ranking efforts might be inclined to use a different set of scales, and might as a result come to slightly different conclusions. Nevertheless, the authors provided a clear rationale for their choices of scales. These choices, despite their controversial nature regarding some issues, might ultimately lead to a more open and focused dialogue about the various aspects that constitute a quality HERS.

Some limitations of this study are at the same time limitations of the BPs. Especially the lack of practical guidance in how HERSs can conform to the principles stated has to be noted. This lack of guidance has especially become obvious regarding BPs 7, 10, and 12. The authors therefore agree with the statement that "it would be useful to rankers if efforts could be made to give examples of good practice for every item of the Berlin Principles" (Cheng and Liu 2008, p. 208). The International Ranking Expert Group is in a unique position to discuss and present such examples. Again, this study might offer a contribution to these discussions.

References

Alexander von Humboldt-Stiftung. (2005). Das Humboldt-Ranking. http://www.humboldt-foundation.de/de/aktuelles/presse/doc/2006ranking72.pdf. Accessed December 29 2006.

Altbach, P. G. (2003). The costs and benefits of world class universities. *International Higher Education*, 33. http://www.bc.edu/bc_org/avp/soe/cihe/newsletter/News33/text003.htm. Accessed September 27 2008.

Altbach, P. G. (2006). The dilemmas of ranking. *International Higher Education*, 42, 2–4. http://www.bc.edu/bc_org/avp/soe/cihe/newsletter/Number42/p2_Altbach.htm. Accessed September 27 2008.

ARRA. (2005). *Assessment of public universities and their faculties.* http://www.arra.sk/oldweb/ARRA-Report2005-EN.pdf. Accessed December 13 2006.

Bellon, B. (2007). Evaluation practices and methodologies: Lessons for university ranking. In J. Sadlak & N. C. Liu (Eds.), *The world-class university and ranking: Aiming beyond status* (pp. 123–145). Bucharest, Rumania: UNESCO-CEPES.

Berghoff, S., Federkeil, G., Giebisch, P., Hachmeister, C.-D., Hennings, M., & Müller-Böling, D. (2007). *CHE—HochschulRanking: Vorgehensweise und Indikatoren—2007.* Guetersloh, Germany: Centrum für Hochschulentwicklung.

Birnbaum, R. (2006). No world-class university left behind. *International Higher Education*, 47. http://www.bc.edu/bc_org/avp/soe/cihe/newsletter/Number47/p7_Birnbaum.htm. Accessed September 27 2008.

Bowden, R. (2000). Fantasy higher education: University and college league tables. *Quality in Higher Education, 6*(1), 47–60.

Breimer, D. (2007). *Rankings: How should universities respond?* Paper presented at the 2nd Leiden University International Symposium on Ranking, Leiden, the Netherlands.

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

CEST. (2004). *CEST scientometrics research portfolios: Universities and colleges participating in the champions league—Diagrams and profiles, 1998–2002.* Bern, Switzerland: Center for Science and Technology Studies of the Swiss Federal Government.

Cheng, Y., & Liu, N. C. (2008). Examining major rankings according to the Berlin Principles. *Higher Education in Europe, 33*(2/3), 201–208.

Clarke, M. (2002). Some guidelines for academic quality rankings. *Higher Education in Europe, 27*(4), 443–459.

de Miguel, J., Cais, J., & Vaquera, E. (2001). *Excelencia: Calidad de las universidades espanolas.* Madrid, Spain: Centro de Investigaciones Sociologicas.

de Miguel, J., Vaquera, E., & Sanchez, J. (2005). Spanish universities and the ranking 2005 initiative. *Higher Education in Europe, 30*(2), 199–215.

de Wit, K. (2003). The consequences of European integration for higher education. *Higher Education Policy, 16*(2), 161–178.

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), 495–533.

Elsevier. (2006, October). Studeren. *Elsevier Thema.*

Enserink, M. (2007). Who ranks the university rankers? *Science, 317*(5841), 1026–1028.

Federkeil, G. (2008). Rankings and quality assurance in higher education. *Higher Education in Europe, 33*(2/3), 219–231.

Gaertner, R., Birke, B., & Kohler, A. (2007). *Hochschulranking 2007—Ingenieurwissenschaften, Geisteswissenschaften, Erziehungswissenschaften und Psychologie: Bericht zu Ergebnissen der österreichischen Universitäten, Verfahren und Methodik.* Vienna, Austria: Austrian Agency for Quality Assurance.

Hazelkorn, E. (2007). The impact of league tables and ranking systems on higher education decision making. *Higher Education Management and Policy, 29*(2), 87–110.

Hendel, D. D., & Stolz, I. (2008). A comparative analysis of higher education ranking systems in Europe. *Tertiary Education and Management, 14*(3), 173–189.

IHEP. (2007). *College and university ranking systems: Global perspectives and American challenges.* Washington, DC: Institute for Higher Education Policy.

IREG. (2006). *Berlin Principles on Ranking of Higher Education Institutions.* Berlin, Germany: International Ranking Expert Group.

Karlsson, S., & Wadskog, D. (2006). *Hur mycket citeras Svenska publikationer? Bibliometrisk översikt över Sveriges vetenskapliga publicering mellan 1982 och 2004.* Stockholm, Sweden: Vetenskapsrådet.

Lebherz, C., Mohr, C., Henning, M., & Sedlmeier, P. (2005). Wie brauchbar sind Hochschul-Rankings? Eine empirische Analyse. In U. Teichler & R. Tippelt (Eds.), *Hochschullandschaft im Wandel* (pp. 188–208). Germany; Switzerland, Beltz: Weinheim; Basel.

Longden, B., & Yorke, M. (2006). *Institutional rankings, marketing, and the needs of intending students.* Paper presented at the Annual EAIR Forum, Rome, August 30 to September 1.

Macleod, D. (2007). *The guardian university guide 2008.* London, UK: Guardian Newspapers.

Marginson, S. (2007). Global university comparison: the second stage. *Griffith University/IRU symposium on international trends in university rankings and classifications.* http://www.griffith.edu.au/conference/university-rankings/pdf/simon-marginson-paper.pdf. Accessed May 15 2008.

Marginson, S., & van der Wende, M. (2007). To rank or to be ranked: The impact of global rankings in higher education. *Journal of Studies in International Education, 11*(3/4), 306–329.

McCormick, A. C. (2008). The complex interplay between classification and ranking of colleges and universities: Should the Berlin Principles apply equally to classification? *Higher Education in Europe, 33*(2/3), 209–218.

McDonough, P. M., Lising, A., Walpole, A. M., & Perez, L. X. (1998). College rankings: Democratized college knowledge for whom? *Research in Higher Education, 39*(5), 513–537.

Merisotis, J. P. (2002). Summary report of the invitational roundtable on statistical indicators for the quality assessment of higher/tertiary education institutions: Ranking and league table methodologies. *Higher Education in Europe, 27*(4), 476–480.

Merisotis, J., & Sadlak, J. (2005). Higher education rankings: Evolution, acceptance, and dialogue. *Higher Education in Europe, 30*(2), 97–101.

Müller-Böling, D. (2001a). Hochschulranking—Aussagefähigkeit, Methoden, Probleme: Eine Einführung. In D. Müller-Böling, S. Hornbostel, & S. Berghoff (Eds.), *Hochschulranking: Aussagefähigkeit, Methoden, Probleme; wissenschaftliche Tagung Berlin 25/26. September 2000* (pp. 5–6). Guetersloh, Germany: Bertelsmann-Stiftung.

Müller-Böling, D. (2001b). Qualität an deutschen Hochschulen: Der Beitrag des Hochschulrankings. In Schweizerische Akademie der Geistes- und Sozialwissenschaften (Ed.), *Welche Qualität in den Sozialwissenschaften? Ansätze, Erfahrungen, Perspektiven: Eine Tagung des Wissenschaftspolitischen Rates für Sozialwissenschaften (CPS/WRS), des Schwerpunktprogramms 'Zukunft Schweiz', der Schweizerischen Evaluationsgesellschaft SEVAL und der Schweizerischen Akademie der Geistes- und Sozialwissenschaften, Gurten/Bern, 16./17. Mai 2001* (pp. 125–142). Bern, Switzerland: Schweizerische Akademie der Geistes- und Sozialwissenschaften.

O'Leary, J., Hindmarsh, A., & Kingston, B. (2006a). *The Times Good University Guide 2007.* London, UK: Times Books.

O'Leary, J., Quacquarelli, N., & Ince, M. (2006b). *Guide to world's top universities: Exclusively featuring the complete THES/QS world university rankings.* London, UK: QS Quacquarelli Symonds, Blackwell Publishing.

PRIME. (2006). *Observatory of the European university: Methodological guide.* Lugano, Switzerland: PRIME Network of Excellence.

Proulx, R. (2006). *World university rankings: The need for a new paradigm.* Paper presented at the 28th annual EAIR Forum, Rome, August 30–September 1.

Sadlak, J., Merisotis, J., & Liu, N. C. (2008). University rankings: Seeking prestige, raising visibility and embedding quality—The editor's view. *Higher Education in Europe, 33*(2/3), 195–199.

Salmi, J., & Saroyan, A. (2007). League tables as policy instruments: Uses and misuses. *Higher Education Management & Policy, 19*(2), 31–68.

Steenkamp, F., de Looper, H., & Bliekendaal, M. (2006). *Keuzegids Hoger Onderwijs 2006–2007: Verzamelgids.* Leiden, The Netherlands: Keuzegids Hoger Onderwijs.

Studentenspiegel. (2006). Methodik des Studentenspiegels. http://www.studentenspiegel1.de/methodik.pdf. Accessed February 14 2007.

swissUp. (2006a). Definitionen der gerankten Indikatoren: Ranking 2005 + 2006. http://www.ranking swissup.ch/pdf/Def_Indikatoren_Ranking06+05_D.pdf. Accessed January 29 2007.

swissUp. (2006b). swissUp Ranking 2006, Bemerkungen zur Methodologie. http://www.rankingswissup.ch/ pdf/methodo06_D_DEF.pdf. Accessed January 29 2007.

Sydsvenska Industri-och Handelskammaren. (2007). Högskolerankingen. http://www.handelskammaren. com/backoffice/bakpress/ladda/H%C2%A8gskolerankingen%202007%20slutlig.pdf. Accessed May 14 2008.

Taylor, P., & Braddock, R. (2007). International university ranking systems and the idea of university excellence. *Journal of Higher Education Policy and Management, 29*(3), 245–260.

Terenzini, P. T., & Pascarella, B. T. (1994). Living with myths: Undergraduate education and what we can do about it. *Change, 26*(1), 28–32.

UNESCO-CEPES. (2006). Second meeting of the International Rankings Expert Group: List of participants. http://www.cepes.ro/hed/meetings/berlin06/list.htm. Accessed October 3 2008.

Usher, A., & Savino, M. (2006). *A world of difference: A global survey of university league tables*. Toronto, Canada: Educational Policy Institute.

Usher, A., & Savino, M. (2007). A global survey of university ranking and league tables. *Higher Education in Europe, 32*(1), 5–15.

van Dyke, N. (2005). Twenty years of university reports cards. *Higher Education in Europe, 30*(2), 103–124.

Yorke, M. (1997). A good league table guide? *Quality Assurance in Education, 5*(2), 61–72.