Multilevel-Statistical Reformulation of Citation-Based University Rankings: The Leiden Ranking 2011/2012

Lutz Bornmann
Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstrasse 8, D-80539 Munich, Germany. E-mail: bornmann@gv.mpg.de

Rüdiger Mutz
Professorship for Social Psychology and Research on Higher Education, ETH Zurich, D-GESS, Mühlegasse 21, 8001 Zürich, Switzerland. E-mail: ruediger.mutz@gess.ethz.ch

Hans-Dieter Daniel
Professorship for Social Psychology and Research on Higher Education, ETH Zurich, D-GESS, Mühlegasse 21, 8001 Zürich, Switzerland; Evaluation Office, University of Zurich, Mühlegasse 21, 8001 Zurich, Switzerland. E-mail: daniel@evaluation.uzh.ch

Since the 1990s, with the heightened competition and the strong growth of the international higher education market, an increasing number of rankings have been created that measure the scientific performance of an institution based on data. The Leiden Ranking 2011/2012 (LR) was published early in 2012. Starting from Goldstein and Spiegelhalter's (1996) recommendations for conducting quantitative comparisons among institutions, in this study we undertook a reformulation of the LR by means of multilevel regression models. First, with our models we replicated the ranking results; second, the reanalysis of the LR data showed that only 5% of the $PP_{top10\%}$ total variation is attributable to differences between universities. Beyond that, about 80% of the variation between universities can be explained by differences among countries. If covariates are included in the model the differences among most of the universities become meaningless. Our findings have implications for conducting university rankings in general and for the LR in particular. For example, with Goldstein-adjusted confidence intervals, it is possible to interpret the significance of differences among universities meaningfully: Rank differences among universities should be interpreted as meaningful only if their confidence intervals do not overlap.

Introduction

There is a long history of competition among universities; they compete for students, professors, and financial means. Since the 1990s, with the heightened competition and the strong growth of the international higher education market, an increasing number of rankings have been created that measure the scientific performance of an institution based on data (Shin, Toutkoushian, & Teichler, 2011). "Recently, the competition has been accelerated in many countries as governments develop initiatives to build world-class universities that can compete more effectively with other leading institutions across the globe. Although there are concerns with using rankings as a tool for measuring the quality of a university, many institutional leaders and policymakers still often rely on rankings to inform their policy-making" (Shin & Toutkoushian, 2011, pp. 1–2; see also Bornmann, in press).

In the rankings, research institutions and universities are ranked on an implicit or explicit dimension of quality based on different criteria (mainly bibliometric indicators, number of research awards, and assessments by survey respondents). At present, important global research rankings include the Academic Ranking of World Universities ("Shanghai Ranking"; http://www.arwu.org/), the Times Higher Education World University Rankings (http://www.timeshighereducation.co.uk/world-university-rankings/), and the SCImago Institutions Rankings (SIR; http://www.scimagoir.com/). Whereas the Shanghai Ranking and the Times Higher Education World University Rankings are performance rankings, the SCImago Institutions Rankings do not claim to be a league table (SCImago Reseach Group, 2012). An up-to-date overview of existing classifications and rankings in higher education and research and an evaluation

Received July 16, 2012; revised October 6, 2012; accepted October 8, 2012

© 2013 ASIS&T • Published online 7 June 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.22857

and criticism of current rankings is provided by van Vught and Ziegele (2012; see also Buela-Casal, Gutiérrez-Martínez, Bermúdez-Sánchez, & Vadillo-Muñoz, 2007; Chen & Liao, 2012; Hazelkorn, 2011). A critical comment on "the application of insufficiently developed bibliometric indicators" (p. 133) in university rankings, like the Shanghai Ranking, can be found in van Raan (2005).

The Leiden Ranking 2011/2012 (LR) was published early in 2012. Similar to the SIR, the LR uses only a set of bibliometric indicators (number of publications, mean citation score, mean normalized citation score, and proportion of the publications of a university that belong to the top 10% publications [$PP_{top10\%}$]) to rank the universities (Waltman et al., 2012). The LR is in fact a set of eight different rankings according to the selection criteria decided by the end user (LR in previous editions used a different set of indicators). The indicator values for the individual universities can be downloaded as an Excel file at the website of the Centre for Science and Technology Studies (CWTS) of Leiden University, which conducts the LR (http://www.leidenranking.com/leidenranking.zip). Even though Waltman et al. (2012) used basic statistical concepts, such as mean (i.e., mean citation), percentile (i.e., $PP_{top10\%}$), and bootstrapping (i.e., stability interval), advanced statistical analyses provide interesting insights into the scientific performance of the universities that basic concepts cannot produce. In this article, we reformulate the LR by means of multilevel regression models (Bornmann, Mutz, Marx, Schier, & Daniel, 2011; Mutz & Daniel, 2007). The reformulation follows Goldstein and Spiegelhalter's (1996) key points for conducting quantitative comparisons among institutions:

> We shall pay particular attention to the specification of an appropriate statistical *model*, the crucial importance of uncertainty in the presentation of results, techniques for *adjustment* of outcomes for confounding factors and finally the extent to which any reliance may be placed on explicit rankings. (p. 390)

In a report by the Joint IMU/ICIAM/IMS-Committee on Quantitative Assessment of Research, Adler, Ewing, Taylor, and Hall (2009) found:

> The article by Goldstein and Spiegelhalter is valuable to read today because it makes clear that the overreliance on simple-minded statistics in research assessment is not an isolated problem. Governments, institutions, and individuals have struggled with similar problems in the past in other contexts, and they have found ways to better understand the statistical tools and to augment them with other means of assessment. (p. 14)

Of the bibliometric indicators used by the LR (see above), in this study we performed a statistical analysis of the indicator $PP_{top10\%}$. Waltman et al. (2012) regard this indicator "as the most important [citation] impact indicator in the Leiden Ranking" (p. 2425). $PP_{top10\%}$ is the proportion of the publications of a university that belong to the top 10% most frequently cited publications; a publication ($P_{top10\%}$) belongs to the top 10% most frequently cited if it is cited more frequently than 90% of publications published in the same field and in the same year (Bornmann, de Moya Anegón, & Leydesdorff, 2012; Bornmann et al., 2011; Leydesdorff, Bornmann, Mutz, & Opthof, 2011).

On the basis of the $PP_{top10\%}$, we examine the following research questions in this study:

1. How well do the ranks and standard errors predicted by our statistical model agree with the ranks and stability intervals in the LR in Waltman et al. (2012)?
2. Are there real differences in citation impact among universities that justify a clear ranking regarding scientific performance? That is, do different rank numbers reflect considerable actual differences in performance?
3. To what extent do differences in citation impact among the universities (considered in the LR) also reflect systematic differences among the countries in which the universities are located? Do certain countries have, on average, better universities than other countries (see Bornmann & de Moya Anegón, 2011)?
4. To what extent can differences among the universities (considered in the LR) be explained, first, by country-specific factors in the area of the economy (per capita gross domestic product [GDP] in the country), geography (total area of a country), and population (number of people in a country) and, second, by a university-specific factor: the size of the university (measured via number of publications)?

Materials and Methods

Data and Variables

The LR ranks only universities; other types of research institutions are not included (the SIR, for example, assesses universities and other research-focused institutions using Scopus data). The bibliographic data used in the LR are based on publications in Thomson Reuters's Web of Science (WoS) database. The ranking considers publications from the period 2005–2009 of the WoS document types: Article, Letter, and Review. Publications in all subject categories are included except for the arts and humanities. The publications are assigned to the different universities via the affiliation addresses provided by the authors. The LR considers only the 500 universities with the largest WoS publication output: "Together, the universities have produced 3.4 million publications in the period 2005–2009. This is 61.3% of all WoS publications in this period. The 500 universities included in the LR are located in 41 different countries" (Waltman et al., 2012). The countries having more than 30 universities included in the LR are China ($n = 31$), the United Kingdom ($n = 36$), Germany ($n = 39$), and the United States ($n = 127$). For a detailed description of the data collection of the LR, see Waltman et al. (2012) and http://www.leidenranking.com/methodology.aspx.

In the data analysis of this study, we included the frequencies of $P_{top10\%}$ reported by the LR and the total number of publications of each university. Because the LR comes with two series of indicators, one series based on full counting and one based on fractional counting, we considered here the $P_{top10\%}$ and the total number of publications based on full counting. We used the two indicators in our analysis to build the dependent variable: For each university, based on individual publications, we produced a binary variable to mark a publication as highly cited (belonging to the top 10%) or not highly cited. For example, for Aarhus University, there are a total of 13,020 publications and a $PP_{top10\%}$ of 14.7% in the data of the CWTS for the LR. Hence, the binary data sheet for this university is made up of 13,020 rows or cells, of which in 1,914 (14.7%) cells there is a 1 (belonging to the top 10%) and in 11,106 a 0 (does not belong to the top 10%).

The $PP_{top10\%}$ of the universities is on average 12.5% (Table 1). The proportion of the universities' publications belonging to highly cited papers is therefore on average 2.5% higher than the expected proportion (10%). Across the universities, the proportion ranges from 3% (MIN) to 26.4% (MAX). To be able to explain the differences in citation impact ($PP_{top10\%}$) among the universities (and countries) in the regression model, we included as covariates the total number of publications of a university, the number of universities per country, the GDP per capita of a country based on purchasing power parity (PPP), the number of residents in a country, and the proportion of residents younger than age 15 years (Table 1). The GDP (PPP) per capita is the gross domestic product converted to international dollars using purchasing power parity rates, which makes it possible to compare different countries (www.worldbank.org). All covariates refer to possible effects of size or volume that can be supposed to have an influence on the citation impact of the universities (see Austrian Science Fund, 2007). The covariates are also utilized to create a covariate-adjusted ranking of the universities.

The publication output of the universities in the LR (the first covariate listed in Table 1) ranges from about 3,230 to 61,623 publications in the period 2005 to 2009. The average publication output of the universities is about 9,000 publications. The larger a university is (i.e., the greater its publication output), the greater its impact in the scientific community. Accordingly, for the universities we expect to find a positive correlation between citation impact and output. Several studies have already shown that there is a correlation between the output and the impact of the publications of research units (see, e.g., Abramo, D'Angelo, & Costa, 2010; Hemlin, 1996).

With regard to GDP (PPP) per capita (the second covariate in Table 1), we assume that, in a country where more money is available generally, there will be more funding for research and thus higher-level research can be conducted than in countries with limited financial means. Miranda and Lima (2010) point out that "the knowledge evolution, as seen through the evolution of major scientific discoveries and impacting technological inventions, is exponentially correlated to the GDP" (p. 92). As Table 1 shows, GDP (PPP) per capita in the countries included in the LR ranges from 3,652 to 61,882; the median value is 30,254. Common sense suggests that the universities in a country with a high GDP (PPP) per capita also have, on average, a higher citation impact than universities in a country with a low GDP (PPP) per capita.

Number of residents, proportion of residents younger than age 15 years (the third and fourth country-level covariates in Table 1), and total area of the country (the fifth

TABLE 1. Summary statistics.

	N	*M*	STD	MIN	Q1	MED	Q3	MAX
Dependent variable								
$PP_{top10\%}$	500	0.125	0.040	0.03	0.096	0.124	0.148	0.264
Covariates								
University level								
Total number of publications	500	8,997	6,262	3,230	4,575	6,984	11,055	61,623
Country level								
GDP (PPP) per capita 2011 (current international $)[a]	41	30,118	13,839	3,652	20,031	30,254	39,438	61,882
Number of residents (millions)[b]	41	106.4	279.4	2.1	7.9	22.7	63.3	1,345.9
Proportion of residents younger than 15 years (percent)[b]	41	18.2	4.8	13.1	15.1	16.0	20.1	29.7
Total area of the country (1,000 km^2)[b]	41	1,931.9	3,831.5	0.7	77.5	312.7	783.6	17,098.2
Number of universities in each country[c]	41	151.8	326.2	4	31	62	144	2,049

N = number of units; *M* = mean; STD = standard deviation; MIN = minimum; Q1 = first quartile (25%); MED = median (50%); Q3 = third quartile (75%); MAX = maximum.

[a] From www.worldbank.org on October 4, 2012 (code: NY.GDP.PCAP.PP.CD). For "Taiwan" (not included in the worldbank database), the GDP (PPP) per capita was retrieved from https://www.cia.gov/library/publications/the-world-factbook/index.html. For "New Zealand," the GDP (PPP) per capita 2010 instead of the missing value in the worldbank database was used, retrieved from www.worldbank.org on October 4, 2012.

[b] From CIA World Factbook, retrieved from https://www.cia.gov/library/publications/the-world-factbook/index.html on October 4, 2012.

[c] From "Universities Worldwide," retrieved from http:// univ.cc (list of universities with university status).

covariate in Table 1) also differ greatly across countries. The number of residents ranges from 2.1 to 1,345.9 (million), the proportion of residents younger than age 15 years varies between 13.1 and 29.7, and the total area ranges from 0.7 to 17,098.2 (1,000 km^2). Because a larger population and a larger total area usually also mean a larger pool of potential (excellent) scientists, we expect positive correlations with citation impact. A high proportion of residents younger than age 15 years is an indicator for developing nations, which should therefore be associated with a low number of highly cited papers. As with these covariates, we also expect to find a positive correlation between number of universities (the sixth covariate in Table 1) and citation impact. The more universities in which research is conducted in a country, the greater the number of excellent papers that we can expect. Because not all universities of a country are included in the LR (see above), we searched the numbers of universities per country at http://univ.cc ("Universities Worldwide"; searched August 20, 2012). The compiled numbers are strongly correlated to the numbers of universities per country in the LR ($r = 0.92$).

Because the covariates are scaled very differently (e.g., km^2, $), all covariates were *z*-transformed with a mean value of 0 and a standard deviation of 1.0. The intercept in the statistical model, where the regression line crosses the y-axis, thus represents the value of a fictitious university for which all covariates are exactly average.

Statistical Procedures

The data of the LR are seen as a snapshot of the universities at a particular moment in time (February 2012, when the data were downloaded) and are thus a sample of all possible points in time. The data of all possible points in time make up the population. In the statistical analyses, we calculated again, first, the university ranking following the CWTS method and the corresponding stability intervals/standard errors (Waltman et al., 2012) and, second, the university ranking and standard errors using multilevel logistic regression (Bornmann et al., 2011).

Stability intervals/standard errors. The CWTS computes for each university the $PP_{top10\%}$ and ranks the universities according to the value of this indicator (in descending order). Additionally, stability intervals are calculated; they indicate "a range of values of an indicator that are likely to be observed when the underlying set of publications changes. . . . The larger the stability interval of an indicator, the lower the stability of the indicator" (Waltman et al., 2012, p. 2429). The stability interval for each university is calculated by resampling from the publication data at hand (bootstrapping). One thousand samples are drawn from the total set of publications with replacement (multiple occurrence of the same publication is possible), whereas the size of the samples equals the size of the total set of publications for each university. The 2.5% and 97.5% percentiles of the distribution of the $PP_{top10\%}$ provide the stability interval (95% interval). The standard deviation of the distribution is the standard error of $PP_{top10\%}$.

In addition to the stability intervals, in this study the standard error (SE) of the $PP_{top10\%}$ for each university with probability $p = PP_{top10\%}$ was calculated using the following formula:

$$SEPP_{top10\%} = \sqrt{(p * (1-p))/N} \tag{1}$$

The 95% confidence interval amounts to $PP_{top10\%}$
± 1.96 SE $PP_{top10\%}$.

Multilevel logistic approach. In multilevel analysis, the hierarchical structure of data is explicitly considered. The single publications are clustered in universities, whereas j ($j = 1 \dots N$) denotes the level-2 units ("universities") and i ($i = 1 \dots n_j$) the level-1 units ("publications"). Because the dependent variable y_{ji} is binary (1 = publication i belongs to the top 10% papers, 0 = publication i does not belong to the top 10% papers), ordinary multilevel models for continuous data without borders (i.e., 0/1) are not adequate. Instead, generalized linear mixed models are favored, especially the multilevel logistic model for binary data, which can be specified by the following three components (Hox, 2010):

1. The probability distribution for p_{ji} ($=\Pr[y_{ji} = 1]$) is a Bernoulli distribution (1, μ) with mean μ.
2. A linear multilevel regression part generates a latent (unobserved) predictor ν_{ji} of the binary outcome y_{ji}: $\nu_{ji} = \beta_0 + \Sigma\beta_r x_{rj} + u_{0j}$, whereas u_{0j} is a normally distributed random effect $u_{0j} \sim N(0, \sigma^2_{u0})$ with the variance σ^2_{u0}, x_r the rth level-2 covariate, and β_0 and β_r the intercept and slope parameters, respectively.
3. The link function that links the expected value of the dependent variable y with the latent predictor ν is here the logit function: $\nu = \text{logit}(\mu) = \log(\mu/[1 - \mu])$. The logit link function transforms probabilities that vary from 0 to 1 to logits, which continuously vary between $-\infty$ and $+\infty$ with a variance of $\pi^2/3 = 3.29$.

With respect to an empty model without any covariates and without any level-1 information, the relative frequency p_j ($PP_{top10\%}$ papers) can be modeled instead of the raw data y_{ji}. The multilevel logistic model for the observed proportions p_j of $PP_{top10\%}$ papers is

$$p_j = \text{logistic}\left(\beta_0 + u_{0j}\right) \quad u_{0j} \sim N\left(0, \sigma^2_{u0}\right), \tag{2}$$

where "logistic" means the logistic transformation of p_i ($\text{logistic}[x] = e^x/[1 + e^x]$), which is nothing but the inverse logit link function. In the case of *r* level 2 covariates, the multilevel model (equation 2) can be augmented as follows.

$$p_j = \text{logistic}\left(\beta_0 + \sum \beta_r x_{rj} + u_{0j}\right) \quad u_{0j} \sim N\left(0, \sigma^2_{u0}\right). \tag{3}$$

Finally, the variability of random effects can be further decomposed in one part that represents the variability of citation impact *among* countries ($k = 1 \dots K$) σ^2_{u0k} and in another part that represents the variability of citation impact *within* countries $\sigma^2_{u0j(k)}$. Thus, Equation 2 can be reformulated in a three-level model (publication, university, country), as follows:

$$p_j = \text{logistic}\left(\beta_0 + u_{0j(k)} + u_{0k}\right) \quad \begin{aligned} u_{0j(k)} &\sim N\left(0, \sigma^2_{u0j(k)}\right) \\ u_{0k} &\sim N\left(0, \sigma^2_{u0k}\right). \end{aligned} \tag{4}$$

In multilevel modeling, it is crucial to know what proportion of the whole variability is explained by true differences among universities and what proportion is due to within-university variability or random fluctuations. The latent variable model offers a solution by transforming the within-university variance to the logistic scale (Goldstein, Browne, & Rasbash, 2002). The unobserved latent variable ν follows a logistic distribution with variance $\pi^2/3 = 3.29$ (level-1 variance), as mentioned above. This makes it possible to formulate an intraclass correlation coefficient (ICC) by using σ^2_{u0}, as follows:

$$\text{ICC} = \sigma^2_{u0} / (\sigma^2_{u0} + 3.29), \tag{5}$$

which indicates the average correlation of two publications' citation impact within universities. An ICC near 0 indicates in combination with a statistically nonsignificant variance component σ^2_{u0} that the universities differ only by chance concerning their publications' citation impact. In this case, any fixed rankings of universities do not make sense.

Because the level-1 variance is arbitrarily fixed (3.29), any meaningful explanation of level-1 variance inevitably changes the level-2 variance components. Therefore, the parameters of the models are corrected to allow different models to be compared (Bauer, 2009).

Besides opportunities for testing parameters, multilevel models also offer empirical Bayes estimates, which consider the accuracy of estimates (Hox, 2010). The lower the information for a university (e.g., small sample size, low variance among universities, σ^2_{u0}) and therefore its accuracy is, the more the predicted value of this university is shrunken to the overall mean $PP_{top10\%}$. In the LR 2011/2012, the number of publications for each university is so high that the accuracy of the mean values is also very accurate, and the $PP_{top10\%}$ are not shrunken toward the overall mean, respectively.

Results

Ranking of Universities and Countries

Table 2 shows the results of the multilevel regression models. In the first model, the 2-level null model (M_1), none of the covariates listed above is included. Also, country is not yet included as a third level in addition to university as the second level. The results of M_1 yield information on how large the differences among universities in $PP_{top10\%}$ are. The variance component of the random intercept in M_1 is 0.149. Null does not lie in the 95% confidence interval of the random intercept's variance component (CL95 0.132, Cu95 0.169). The loglikelihood ratio test is used to test whether M_1 fits the data as well as the most restricted model M_0 without random intercept (not shown in Table 2). The test value at χ^2_{LR} ($df = 1$) = 71,316.0 is very high and statistically significant ($p < 0.05$), so the two models strongly differ. Both results (null does not lie in the confidence interval, and the loglikelihood ratio test is statistically significant) speak for systematic differences in the $PP_{top10\%}$ indicator among the

TABLE 2. Results of three multilevel logistic regressions for binary outcomes (publications belonging to the top 10% most-cited papers or not).

		M_1: 2-level null model				M_2: 3-level null model				M_3: 3-level model with covariates			
Terms	Par	Est	Cl95	Cu95	Rescal	Est	Cl95	Cu95	Rescal	Est	Cl95	Cu95	Rescal
Fixed effects													
Intercept	β_0	–1.997	–2.031	–1.963	–1.077	–2.235	–2.359	–2.112	–1.198	–2.063	–2.184	–1.942	–1.112
University: Total number of publications	β_1									0.098	0.081	0.115	0.053
Country: Number of universities	β_2									0.026	–0.174	0.227	0.014
GDP (PPP) per capita 2011	β_3									0.312	0.231	0.394	0.168
Number of residents	β_4									0.177	0.076	0.278	0.096
Proportion of residents younger than 15 years	β_5									0.05	–0.011	0.111	0.027
Total area of the country	β_6									–0.141	–0.244	–0.039	–0.076
Random effects													
University (j(k))	$\sigma^2_{uj(k)}$	0.149	0.132	0.169	0.043	0.042	0.037	0.048	0.012	0.032	0.280	0.037	0.009
Country (k)	σ^2_{u0k}					0.149	0.097	0.258	0.043	0.043	0.027	0.080	0.013
Residual	σ^2_{ε}	3.29	—	—	0.957	3.29	—	—	0.945	3.29	—	—	0.955
BIC			7,143.65				6,636.94				6,508.79		

Par = parameter; Est = estimate; CL95 = lower 95% confidence limit; Cu95 = upper 95% confidence limit; Rescal = rescaled parameter; BIC = Schwarz Bayesian information criterion.

DOI: 10.1002/asi

universities. However, the share of the universities in the $PP_{top10\%}$ total variance is not very high. The rescaled parameters in Table 2 show that only 0.043 or 4.3% of the variance in citation impact can be attributed to differences among the universities. The rest, namely, 0.957 or 95.7%, is allotted to variance within the universities (i.e., to departments, research groups, and individual researchers).

In model M_2 in Table 2, the country is included as a further level (level 3) in addition to the university. M_2 is an improvement over M_1. For one thing, the Schwarz Bayesian information criterion (BIC) of M_2 is clearly lower than that of M_1 (6,636.94 vs. 7,143.65). For another, the intercept of M_2 is closer to the value of 10% (the value that would be expected for $PP_{top10\%}$) than the intercept of M_1 is: For M_2 with –2.235, there is a $PP_{top10\%}$ probability of 0.097 ($e^{-2.235}/(1 + e^{-2.235})$; for M_1, the probability is 0.12 ($e^{-1.997}/(1 + e^{-1.997})$. The intercept of the $PP_{top10\%}$ probability for M_2 is also better than the actual measured value for the universities of $M = 0.125$ (see Table 1). As the random effects in Table 2 show, $(0.012 + 0.043) = 0.055$ or 5.5% of the $PP_{top10\%}$ total variance is attributable to differences between the universities and 0.945 or 94.5% to differences within the universities. This result agrees with the result in M_1. Beyond that, however, in M_2 78.2% of the systematic total variance between the universities can be explained by differences among the countries. The country in which a university is located therefore plays a considerable role in its citation performance.

Figure 1 shows, as the outcome from M_2, the ranking of the 50 best of the total 500 universities. The predicted probabilities are shown for the individual universities as Goldstein-adjusted confidence intervals (Hox, 2010, p. 25). In contrast to a pure ranking of universities, this has the advantage of providing indications of important differences between the universities. Only differences in rank between the universities whose Goldstein-adjusted confidence intervals do not overlap should be interpreted as meaningful (see Figure 1). This means that Massachusetts Institute of Technology (MIT) differs in its citation performance from all other universities statistically significantly; this is also true for Harvard University, with one exception, Princeton University. The difference between Harvard and Princeton is so small that it should not be interpreted as meaningful. Figure 1 also shows that the differences among most of the universities are so small that they can only rarely be interpreted as meaningful. We can assume that not only in the LR but also in many other university rankings the differences among most of the universities are meaningless.

In our next analysis, we compared the results of the LR with the results of M_2 (or of Figure 1). To do so, we calculated, based on the individual universities, correlations between the ranking results of the LR (http://www.leidenranking.com/ranking.aspx; using full counting, leaving in non-English publications, and without normalizing for university size) and the M_2. Table 3 shows the correlation coefficients. The coefficients point out that the results

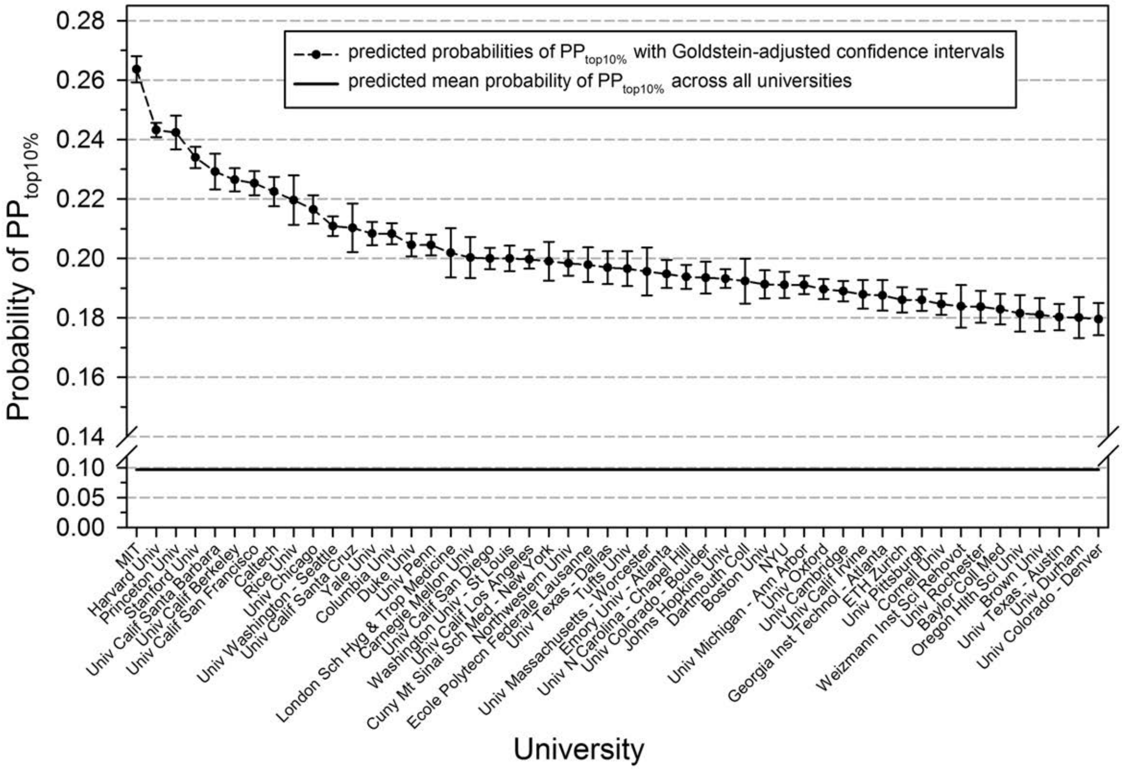

FIG. 1. The 50 highest ranked universities among the total of 500 universities, ranked from left to right according to decreasing $PP_{top10\%}$ probabilities.

TABLE 3. Correlations (Kendall's Tau) of university rankings ($N = 500$).

	Leiden Ranking	M_2 (multilevel unadjusted)	M_3 (multilevel adjusted)
Leiden Ranking	1.00	0.99	0.47
M_2 (multilevel unadjusted)	0.99	1.00	0.46
M_3 (multilevel adjusted)	0.47	0.46	1.00

of LR correlate very highly with M_2. As the comparison of Figure 1 with the LR shows, although the ranks of some of the universities are different (for instance, Harvard University ranks 2 in Figure 1 but ranks 3 in the LR), differences in rank position are only minimal overall.

Based on the $PP_{top10\%}$ for each university, Figure 2 shows a ranking of the countries that are included in the LR with at least one university. Here, again, differences among the countries should be interpreted as meaningful only if their confidence intervals do not overlap. For countries such as Russia and Serbia, the probability of $PP_{top10\%}$ is lower than 10% (that is, they have fewer papers in this category than would be expected in terms of their paper output); countries such as Switzerland and the United States have a higher probability. If we compare the ranking of the countries in Figure 2 with the SCImago Country Rank (http://www.scimagojr.com/ based on Scopus data, countries ordered by citations per paper for papers published between 1996 and 2010), we find similar ranks for the individual countries. In the SCImago Country Rank, the top five ranks are held by Switzerland (21.77 citations per document), Denmark (20.42 citations per document), the United States (20.18 citations per document), Netherlands (20.05 citations per document), and Sweden (19.09 citations per document).

Covariate-Adjusted Rankings of Universities

In the last model (M_3) in Table 2, we tested the extent to which the covariates described above have an influence on the $PP_{top10\%}$ of the universities. In terms of the BIC, M_3 outperforms M_2 with a lower value (6,508.79 vs. 6,636.94). Table 2 shows for each covariate the 95% confidence interval of the parameters (Cl95 und Cu95), which takes into account the dependence of measurement within universities. If, for a covariate, the null no longer lies in the confidence interval, the parameter is statistically significant. As the results in Table 2 show, this is the case for all of the covariates except for two. This means that all covariates except two have a statistically significant effect on the universities' performance: The larger the publication output of a university and the greater the number of inhabitants, total area, and GDP (PPP) per capita, the higher the citation impact of a university. Approximately 25% of the $PP_{top10\%}$ variance between the universities using the rescaled variance component $\sigma^2_{ui(j)}$ of M_2 and M_3 ($100 * [0.012 - 0.009]/0.012 = 25\%$) and 70% of the $PP_{top10\%}$ variance between the countries ($100 * [0.043 - 0.013]/0.043 = 70\%$) are explained by the covariates. In light of these results, it makes sense to take these covariates into consideration with university rankings, to obtain covariate-adjusted results.

Figure 3 shows the covariate-adjusted ranking of universities and the differences among the universities, with the assumption that all of the universities have the same mean in each of the covariates included (see the means in Table 1). If we compare the results in Figure 3 with the results in Figure 1, it is interesting to see that the results differ greatly. The results of the correlation analysis also make this clear: There is only a moderate correlation ($r = 0.47$) between the results of the LR (http://www.leidenranking.com/ranking.aspx; using full counting, leaving in non-English publications, and without normalizing for university size) and the results of the covariate-adjusted multilevel model M_3 (see Table 3). When comparing the ranks of the different universities, MIT, for example, ranked first in performance in Figure 1 but is at rank 6 in Figure 3. The universities with the best covariate-adjusted performance in Figure 3 are the London School of Hygiene and Tropical Medicine (London School Hyg Trop Med), the Weizmann Institute of Science (Weizmann Inst Sci Rehovot; in Israel), the Hong Kong University of Science and Technology (Hong Kong Univ Sci Technol; in China), the University of Cape Town (Univ Cape Town; in South Africa). These four universities differ (covariate adjusted) in citation impact statistically significantly from most other universities. Overall, the differences among the universities become smaller.

Comparison of the Stability Interval With Other Possible Ways to Calculate Standard Errors

Finally, we tested the extent to which Waltman et al.'s (2012) procedure for constructing the stability interval (stability interval with the aid of bootstrapping) yields results similar to the results of other less data-intensive computing procedures. To do so, we compared the SE yielded by bootstrapping with (a) the SE of a binary probability (see the formula in Materials and Methods) and with (b) the SE calculated in the multilevel analysis (M_2). The results show that the different procedures correlate perfectly ($r = 1.0$), with a nearly identical mean of 0.004; that is, the SEs calculated in the different ways are nearly identical.

Discussion

Starting out from Goldstein and Spiegelhalter's (1996) recommendations for the conducting of quantitative comparisons among institutions, in this study we undertook a reformulation of the LR (Waltman et al., 2012) with the following major findings and recommendations: First, our results show that approximately 5% of the $PP_{top10\%}$ total variance can be attributed to differences among universities; approximately 95% is attributable to variance within the universities. Moreover, about 80% of the $PP_{top10\%}$

DOI: 10.1002/asi

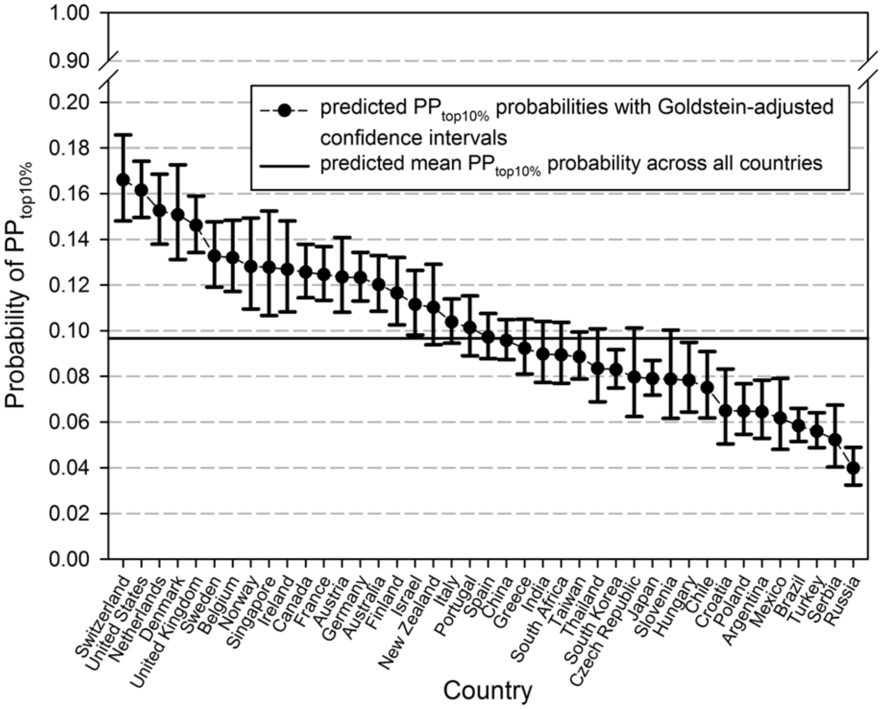

FIG. 2. Ranking of all countries with at least one university in the Leiden Ranking, ranked from left to right according to decreasing $PP_{top10\%}$ probabilities.

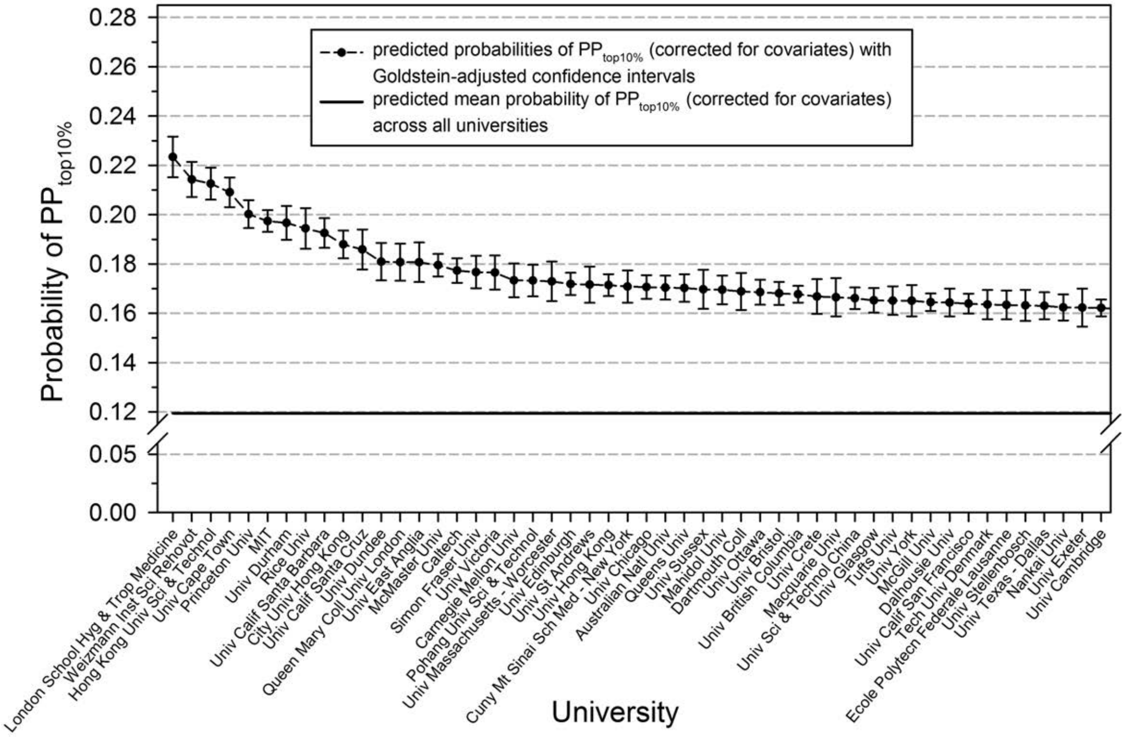

FIG. 3. Covariate-adjusted ranking of the 50 highest ranked universities among the total of 500 universities, ranked from left to right according to decreasing $PP_{top10\%}$ probabilities.

variance between the universities is explained by differences between the countries in which the universities are located (about 20% of the variance is explained by differences between universities within the countries). Thus, the country in which a university is located plays an important role in the performance of a university. The low percentage of the variance of $PP_{top10\%}$ that is explained by differences between universities as well as the high percentage of variance of $PP_{top10\%}$ that can be explained by differences between the countries and within the universities make it appear questionable to conduct rankings at the level of the universities (Daniel & Fisch, 1990). Apparently, the framework conditions in the different countries and the performance capability of smaller research units play a more important role in explaining differences in research performance. For this reason, we recommend that, in the future, rankings should be $PP_{top10\%}$ rankings based on a three-level model including universities *and* countries. In any case, all university rankings should be complemented by a country ranking (see Williams, de Rassenfosse, Jensen, & Marginson, 2012).

Second, university rankings are very important when it comes to research policy. When a university improves its position in a ranking, this is used in marketing and is promotionally effective. Because most rankings do not work with appropriate confidence intervals, any minimal improvement in rank can be interpreted as a significant increase in performance. However, with Goldstein-adjusted confidence intervals, it is possible to interpret the significance of differences over time meaningfully: Rank differences among universities should be interpreted as meaningful only if their confidence intervals do not overlap. The universities differ statistically significantly only when there is no overlapping. Thus, we recommend that not only in the LR but also in other university rankings Goldstein-adjusted confidence intervals should be used to present the results.

Third, the covariates examined in this study explain approximately 25% of the $PP_{top10\%}$ variance between the universities within countries and 70% of the $PP_{top10\%}$ variance between the countries. We chose as covariates mainly possible effects of size, such as size of a university or country, and the financial means of a country. The differences in the results between the covariate-adjusted university ranking and the LR in Waltman et al. (2012; or the unadjusted ranking based on multilevel analysis) support the conclusion that, in the future, university rankings should always be conducted also using covariates that explain the variance (in addition to a variant without the covariates; see also Jovanovic et al., 2012). We use a specific set of covariates in this study. In other studies (based on data from other university rankings), variables such as subject area profile and language of publication (not considered in this study) may be more important than demographic and economic measures (outside the control of universities).

Together with random country effects, using a value-added model (Ballou, Sanders, & Wright, 2004; Raudenbush, 2004), an expectancy value or reference value could be calculated for each university that shows the expected $PP_{top10\%}$ for a university with comparable framework conditions (e.g., country, GDP per capita). Using these reference values, it is possible to ascertain whether a certain university achieves its reference value (sufficient performance), exceeds it (excellent performance), or does not achieve it (nonsufficient performance).

Fourth, Waltman et al. (2012) use a very data-intensive computing procedure to calculate stability intervals. Because the stability intervals are nearly identical to the common standard error of a probability, for the LR the bootstrapping procedure would not be necessary, and the usual SE could be used.

The methods introduced here for conducting a university ranking offer a lot of advantages, but they do have limitations. A prerequisite for the use of the methods is the independence of the underlying data. Because scientists at universities are more often publishing jointly, the independence of the data is not always ensured. The independence of the data could be taken into consideration in the statistical analysis if the CWTS published the data set at the level of individual publications and not (as currently) in aggregated form at the level of the individual universities (Bornmann et al., 2011). A second limitation of our study concerns the indicator $PP_{top10\%}$. We treat the indicator as a binary indicator; that is, a publication is in the top 10% or not. However, CWTS uses a fractional counting approach for publications that are at the top 10% threshold. For example, for a set with 41 publications of the same subject category, in which two publications have 15 citations each, four publications have 14 citations each, and 35 publications have 10 citations each, the citation threshold for the 10% most-cited publications is not clear (14 or 15 citations). According to the fractional counting approach, the four publications with 14 citations would be assigned to the 10% most-cited publications with a weighting of 0.525 (and to the remaining 90% with a weighting of 0.525). Thus, the assumption of the $PP_{top10\%}$ as a binary indicator is not completely valid in this study.

There is another, less important, third limitation of this study, namely, with publications of the document type "letter." In the LR, these publications have a weight of 0.25. That means the publication counts provided by the LR are usually noninteger, even in the case of full counting. A fourth limitation refers to the country ranking. For a valid ranking of countries, data for all universities in all countries worldwide would be necessary. Countries with many universities, which are not included in the LR (only the 500 universities with the largest WoS publication output are considered), might outperform other countries because of cumulative numbers of papers in the $PP_{top10\%}$. Thus, the ranking of countries in this study reflects (mainly) the different amounts of the countries' contributions to the performance of their universities, as far as they are included in the LR.

In addition to the implications that follow from our results mentioned above, in the future we would like to see university rankings that also take into account the range of disciplines/fields at a university (van Vught & Ziegele,

DOI: 10.1002/asi

2012). The results of Bornmann, de Moya Anegón, and Mutz (in press) point out that, in university rankings based exclusively on bibliometric data, universities that focus on disciplines with a high citation volume (e.g., life sciences) have an advantage over other universities that have mainly disciplines with a lower citation volume (e.g., engineering). Although the indicators, which are used in the rankings, are as a rule field and age normalized, this effect is visible. In the World Report 2012 of SCImago, an excellent citation impact, a field- and age-normalized citation impact higher than 1.75, "has been obtained by 428 institutions . . . mainly highly specialized Health Research Institutions" (SCImago Reseach Group, 2012). Because a university should not receive a better or worse rank because of its range of disciplines, a ranking should be conducted specifically for a discipline, or the range of disciplines at a university should be taken into consideration in the statistical model (Mutz & Daniel, 2012).

References

Abramo, G., D'Angelo, C.A., & Costa, F.D. (2010). Testing the trade-off between productivity and quality in research activities. Journal of the American Society for Information Science and Technology, 61(1), 132–140.

Adler, R., Ewing, J., Taylor, P., & Hall, P.G. (2009). A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS). Statistical Science, 24(1), 1–28. doi: 10.1214/09-sts285

Austrian Science Fund. (2007). Rethinking the impact of basic research on society and the economy. Vienna, Austria: Austrian Science Fund.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. Journal of Educational and Behavioral Statistics, 29(1), 37–65.

Bauer, D.J. (2009). A note on comparing the estimates of models for cluster-correlated or longitudinal data with binary or ordinal outcomes. Psychometrika, 74(1), 97–105. doi: 10.1007/S11336-008-9080-1

Bornmann, L. (in press). On the function of university rankings. Journal of the American Society of Information Science and Technology.

Bornmann, L., & de Moya Anegón, F. (2011). Some interesting insights from aggregated data published in the World Report SIR 2010. Journal of Informetrics, 5(3), 486–488. doi: 10.1016/j.joi.2011.03.005

Bornmann, L., Mutz, R., Marx, W., Schier, H., & Daniel, H.-D. (2011). A multilevel modelling approach to investigating the predictive validity of editorial decisions: do the editors of a high-profile journal select manuscripts that are highly cited after publication? Journal of the Royal Statistical Society Series A (Statistics in Society), 174(4), 857–879. doi: 10.1111/j.1467-985X.2011.00689.x

Bornmann, L., de Moya Anegón, F., & Leydesdorff, L. (2012). The new Excellence Indicator in the World Report of the SCImago Institutions Rankings 2011. Journal of Informetrics, 6(2), 333–335. doi: 10.1016/j.joi.2011.11.006

Bornmann, L. de Moya Anegón, F., & Mutz, R. (in press). Do universities or research institutions with a specific subject profile have an advantage or a disadvantage in institutional rankings? A latent class analysis with data from the SCImago ranking. Journal of the American Society of Information Science and Technology.

Buela-Casal, G., Gutiérrez-Martínez, O., Bermúdez-Sánchez, M., & Vadillo-Muñoz, O. (2007). Comparative study of international academic rankings of universities. Scientometrics, 71(3), 349–365.

Chen, K.-H., & Liao, P.-Y. (2012). A comparative study on world university rankings: A bibliometric survey. Scientometrics, 92(1), 89–103. doi: 10.1007/s11192-012-0724-7

Daniel, H.-D., & Fisch, R. (1990). Research performance evaluation in the German university sector. Scientometrics, 19(5-6), 349–361.

Goldstein, H., & Spiegelhalter, D.J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. Journal of the Royal Statistical Society Series A (Statistics in Society), 159, 385–409.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in generalized linear multilevel models. Understanding Statistics, 1, 223–232.

Hazelkorn, E. (2011). Rankings and the reshaping of higher education. The battle for world-class excellence. New York: Palgrave Macmillan.

Hemlin, S. (1996). Research on research evaluations. Social Epistemology, 10(2), 209–250.

Hox, J.J. (2010). Multilevel analysis: techniques and applications (2nd ed.). New York, NY: Routledge.

Jovanovic, M., Jeremic, V., Savic, G., Bulajic, M., & Martic, M. (2012). How does the normalization of data affect the ARWU ranking? Scientometrics, 93(2), 319–327. doi: 10.1007/s11192-012-0674-0

Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables in citation analysis one more time: Principles for comparing sets of documents. Journal of the American Society for Information Science and Technology, 62(7), 1370–1381.

Miranda, L.C.M., & Lima, C.A.S. (2010). On trends and rhythms in scientific and technological knowledge evolution: A quantitative analysis. International Journal of Technology Intelligence and Planning, 6(1), 76–109.

Mutz, R., & Daniel, H.D. (2007). Development of a ranking procedure by mixed Rasch model and multilevel analysis—Psychology as an example. Diagnostica, 53(1), 3–16. doi: 10.1026/0012-1924.53.1.3

Mutz, R., & Daniel, H.-D. (2012). Skewed citation distributions and bias factors: Solutions to two core problems with the journal impact factor. Journal of Informetrics, 6(2), 169–176. doi: 10.1016/j.joi.2011.12.006

Raudenbush, S.W. (2004). What are value-added models estimating and what does this imply for statistical practice? Journal of Educational and Behavioral Statistics, 29(1), 121–129.

SCImago Reseach Group. (2012). SIR World Report 2012. Granada, Spain: University of Granada.

Shin, J.C., & Toutkoushian, R.K. (2011). The past, present, and future of university rankings. In J.C. Shin, R.K. Toutkoushian, & U. Teichler (Eds.), University rankings: Theoretical basis, methodology and impacts on global higher education (vol. 3, pp. 1–16). Dordrecht, Netherlands: Springer.

Shin, J.C., Toutkoushian, R.K., & Teichler, U. (Eds.). (2011). University rankings: Theoretical basis, methodology and impacts on global higher education. Dordrecht, Netherlands: Springer.

van Raan, A.F.J. (2005). Fatal attraction: conceptual and methodological problems in the ranking of universities by bibliometric methods. Scientometrics, 62(1), 133–143.

van Vught, F.A., & Ziegele, F. (Eds.). (2012). Multidimensional ranking: The design and development of U-Multirank. Dordrecht, Netherlands: Springer.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E.C.M., Tijssen, R.J.W., van Eck, N.J., Wouters, P., et al. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. Journal of the American Society for Information Science and Technology 63(12), 2419–2432. doi: 10.1002/asi.22708

Williams, R., de Rassenfosse, G., Jensen, P., & Marginson, S. (2012). U21 ranking of national higher education systems. Melbourne, Australia: University of Melbourne.

DOI: 10.1002/asi