UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

A longitudinal analysis of university rankings

Selten, F.; Neylon, C.; Huang, C.-K.; Groth, P.

[Link to publication](#)

The MIT Press

RESEARCH ARTICLE

A longitudinal analysis of university rankings

Friso Selten[1] iD, Cameron Neylon[2] iD, Chun-Kai Huang[2] iD, and Paul Groth[1] iD

[1]Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[2]Centre for Culture and Technology, Curtin University, Perth, Australia

ABSTRACT

Pressured by globalization and demand for public organizations to be accountable, efficient, and transparent, university rankings have become an important tool for assessing the quality of higher education institutions. It is therefore important to assess exactly what these rankings measure. Here, the three major global university rankings—the Academic Ranking of World Universities, the Times Higher Education ranking and the Quacquarelli Symonds World University Rankings—are studied. After a description of the ranking methodologies, it is shown that university rankings are stable over time but that there is variation between the three rankings. Furthermore, using principal component analysis and exploratory factor analysis, we demonstrate that the variables used to construct the rankings primarily measure two underlying factors: a university's reputation and its research performance. By correlating these factors and plotting regional aggregates of universities on the two factors, differences between the rankings are made visible. Last, we elaborate how the results from these analysis can be viewed in light of often-voiced critiques of the ranking process. This indicates that the variables used by the rankings might not capture the concepts they claim to measure. The study provides evidence of the ambiguous nature of university rankings quantification of university performance.

1. INTRODUCTION

Over the past 30 years, the public sector has been subject to significant administrative reforms driven by an increased demand for efficiency, effectiveness, and accountability. This demand sparked the creation of social measures designed to evaluate the performance of organizations and improve accountability and transparency (Romzek, 2000). Universities, as part of this public sector, have also been subject to these reforms (Espeland & Sauder, 2007). One of the measures taken in the higher education domain to serve this need for accountability and transparency is the popularization of university rankings (URs). URs are "lists of certain groupings of institutions […] comparatively ranked according to a common set of indicators in descending order" (Usher & Savino, 2006, p. 5).

The idea of comparing universities dates to the 1980s when the *US News & World Report* released the first ranking of American universities and colleges. The process however gained major attention in 2003 with the release of the Shanghai league table (Stergiou & Lessenich, 2014). Many new URs have been established since then, with the most notable being the THE-QS and the Webometrics Ranking of World Universities in 2004, the NTU (HEEACT) Ranking, and the CWTS Leiden Ranking in 2007. In 2009, the THE and QS rankings split, and they have published separate rankings since 2010. These rankings make it easy to quantify the

achievements of universities and compare them to each other. Universities therefore use the rankings to satisfy the public demand for transparency and information (Usher & Savino, 2006). Furthermore, rankings were met with relief and enthusiasm by policy-makers and journalists, and also by students and employers. Students use them as a qualification for the value of their diploma, employers to assess the quality of graduates, and governments to measure a university's international impact and its contribution to national innovation (Hazelkorn, 2007; Van Parijs, 2009). Also, the internationalization of higher education has increased the demand for tools to assess the quality of university programs on a global scale (Altbach & Knight, 2007). From this perspective the increasing importance of URs can be understood because they provide a tool for making cross-country comparisons between institutions. The impact of rankings is unmistakable. They affect the judgments of university leaders and prospective students, as well as the decisions made by policy-makers and investors (Marginson, 2014). For certain politicians, having their country's universities at the top of the rankings has become a goal in itself (Billaut, Bouyssou, & Vincke, 2009; Saisana, d'Hombres, & Saltelli, 2011).

University rankings also quickly became subject to criticism (Stergiou & Lessenich, 2014). Fundamental critiques of the rankings are twofold. First, some researchers question whether the indicators used to compute the ranking are actually a good proxy for the quality of a university. It is argued that the indicators that the rankings use are not a reflection of the attributes that make up a good university (Billaut et al., 2009; Huang, 2012). In addition, researchers reason that URs can become a self-fulfilling prophecy; a high rank creates expectations about a university and this causes the university to remain at the top of the rankings. For example, prior rankings influence surveys that determine future rankings, they influence funding decisions, and universities conform their activities to the ranking criteria (Espeland & Sauder, 2007; Marginson, 2007).

Other criticisms focus on the methodologies employed by the rankings. This debate often revolves around the weightings placed on the different indicators that comprise a ranking. The amount of weight placed on certain variables is decided by the rankings' designers, but research has shown that making small changes to the weights can cause a major shift in ranking positions. Therefore, the position of a university is largely influenced by decisions made by the rankings' designers (Dehon, McCathie, & Verardi, 2009; Marginson, 2014; Saisana et al., 2011). Also, the indicator normalization strategy used when creating the ranking can influence the position of a university (Moed, 2017). Normalization is thus, next to the assignment of weightings, a direct manner in which the ranking designers influence ranking order. Furthermore, it has been suggested that rankings are biased towards universities in the United States or English-speaking universities, for example, by using a subset of mostly English journals to measure the number of publications and citations (Pusser & Marginson, 2013; Van Raan, 2005; Vernon, Balas, & Momani, 2018). Last, there is evidence that suggests that there are major deficiencies present in the collection of the ranking data; that is the data used to construct the rankings are incorrect (Van Raan, 2005).

The aim of this research is to better understand what it is that URs measure. This is studied by examining the data that are used to compile the rankings. We assess longitudinal patterns, observe regional differences, and analyze whether there are latent concepts that underly the data used to build the rankings: Can the variables used in the ranking be categorized into broader concepts? The relation between the result of these analyses and the various criticisms described above will also be discussed. Three rankings are analyzed in this study: the Academic Ranking of World Universities (ARWU), the Times Higher Education World University Ranking (THE) and the Quacquarelli Symonds World University Rankings (QS). These rankings are selected because they are seen as the most influential and they claim international coverage. They are

also all general in that they measure the widest variety of variables, as they focus not only on research but also on teaching quality (Aguillo et al. 2010; Scott, 2013).

1.1. Related Work

We take a data-driven approach to our analysis, which is somewhat uncommon in the literature. The most notable works that study URs using such an approach are Aguillo et al. (2010), Dehon et al. (2009), Docampo (2011), Moed (2017), Safón (2013), and Soh (2015).

Aguillo et al. (2010) study the development of the ARWU and the, THE-QS rankings (at that time still publishing a ranking together). This research shows that rankings differ quite extensively from each other, but that they do not change much over the years. This is also confirmed by the research of Moed (2017), which shows that, when analyzing five URs (besides the ARWU, THE, and QS this paper also considers the Leiden and U-Multirank rankings), only 35 universities appear in the top 100 of every ranking. Furthermore, this research examines relations between similar variables that the rankings measure. This analysis proves that citation measures between the different rankings in general are strongly correlated. Also, variables that aim at measuring reputation and teaching quality show moderate to strong correlation (Moed, 2017). Where Moed (2017) explores the relation between the ranking variables using correlations, these relations have also been analyzed using more sophisticated techniques: principal component analysis (PCA) and exploratory factor analysis (EFA). Dehon et al. (2009) use this first technique to study the underlying concepts that are measured by URs. Their research provides insights into the ARWU ranking by showing that the 2008 edition of this ranking measured two distinct concepts: the volume of publications and the quality of research conducted at the highest level. This is also found by Docampo (2011), who applies PCA to data from the ARWU and shows that the extracted components can be used to assess the performance of a university at a country level. Safón (2013) and Soh (2015) both apply EFA to URs. Safón (2013) shows that the ARWU ranking measures a single factor, while the THE ranking measures three distinguishable factors. Likewise, the study by Soh (2015) suggests that the ARWU ranking only measures academic performance, while the THE and QS rankings also include nonacademic performance indicators.

We take inspiration from this prior work, but move beyond it by performing our analysis longitudinally, over three rankings, using multiple analysis approaches as well as performing geographic and sample comparisons. Specifically, the contribution of this paper is fourfold:

1. It describes the evolution of, and gives a comparison between, the three major URs over the past 7 years.
2. It shows the results of a multi year robust PCA and EFA of the UR data, expanding on the work of Dehon et al. (2009), Safón (2013), and Soh (2015).
3. It provides evidence that URs are primarily measuring two concepts and discusses the implications of this finding.
4. It demonstrates a new visualization of how the position of specific (groups of) universities in the rankings changes over time.

The structure of this paper is as follows. First, a general explanation of the rankings methodologies and data collection is given in Sections 2 and 3. Then, our exploratory analysis of the ranking data is discussed in Section 4. This section also studies longitudinal stability and cross-ranking similarity. This is followed by the presentation of our analysis of the latent concepts underlying the rankings using PCA and EFA (Section 5). Finally, the implications of the results and limitations of the study are discussed.

2. RANKING METHODOLOGIES

This section briefly outlines what concepts the rankings use to compute ranking scores and the variables they use to evaluate these concepts. In the next three sections, after each concept the weight assigned to this concept when calculating a university's overall ranking score is indicated between parenthesis.

2.1. ARWU Ranking

The Academy Ranking of World Universities aims to measure four main concepts: Quality of Education (Alumni, 0.1), Quality of Faculty (Award, 0.2; HiCi, 0.2), Research Output (NS, 0.2; PUB, 0.2) and Per Capita Performance (PCP, 0.1). Quality of Education is operationalized by counting the number of university graduates that have won a Nobel Prize or Fields Medal. Awards won since 1911 are taken into account, but less value is assigned to prizes that were won longer ago. Quality of Faculty is similarly measured by counting (since 1921) the Nobel Prizes in physics, chemistry, medicine, and economics, and Fields Medals in mathematics won by staff working at the university at the time of winning the prize. In addition, the number of staff members that are listed on the Highly Cited Researchers list compiled by Clarivate Analytics is used as an input variable. Research output is measured using the number of papers published in *Nature* and *Science* and the total number of papers indexed in the Science Citation Index-Expanded and Social Science Citation Index. The per capita performance variable is a construct of the other five measured variables and—depending on the country a university is in—this construct is either divided by the size of the academic staff to correct for the size of a university or is a weighted average of the five other variables (Academic Ranking of World Universities, 2018). For a more in-depth overview of the ARWU methodology see the official ARWU website (http://www.shanghairanking.com), or the articles from Billaut et al. (2009), Dehon et al. (2009), Docampo (2011), Vernon et al. (2018), and Marginson (2014).

2.2. THE Ranking

The Times World Ranking of Universities is constructed from the evaluation of five different concepts: Teaching (0.3), Research (0.3), Citations (0.3), International Outlook (0.075), and Industry Income (0.025). Half of the Teaching indicator is constructed using a survey that aims to measure the perceived prestige of institutions in teaching. The other half is made up by the staff-to-student ratio, doctorate-to-bachelor's ratio, doctorates-awarded-to-academic-staff ratio and institutional income. Research is mostly measured using a survey that seeks to determine a university's reputation for research excellence among its peers. Furthermore, research income and research productivity (the number of publications) are taken into account when constructing the Research variable. Citations are measured by averaging the number of times a university's published work is cited. The citation measure is therefore normalized with regard to the total papers produced by the staff of the institution. Furthermore, data are normalized to correct for differences in citation rates; for example, in the life sciences and natural sciences average citations are much higher than in other research areas, such as arts and humanities. Data on citations have been provided by Elsevier using the Scopus database since 2015. Prior to 2015 this information was supplied by WoS. International Outlook is measured by evaluating the proportion of international students and international staff and the amount of international collaborations. Industry Income is measured by assessing the research income that an institution earns from industry (THE World University Ranking, 2018). For a more in-depth overview of the THE

methodology, see the official THE website (www.timeshighereducation.com), or the articles by Vernon et al. (2018) and Marginson (2014).

2.3. QS Ranking

The QS evaluates six different concepts: Academic Reputation (0.4), Employer Reputation (0.1), Faculty/Student Ratio (0.2), Citations per faculty (0.2), International Faculty Ratio (0.05), and International Student Ratio (0.05). Academic Reputation is based on a survey of 80,000 individuals who work in the higher education domain. Employer Reputation is measured by surveying 40,000 employers. The Faculty/Student Ratio variable measures the number of students per teacher and is used as a proxy to assess teaching quality. Citations are measured, using Elsevier's Scopus database, by counting all citations received by the papers produced by the institution's staff across a 5-year period and dividing this by the number of faculty members at that institution. As in the THE ranking, since 2015 the citation scores are normalized within each faculty to account for differences in citation rates between research areas. International Faculty and Student ratios subsequently measure the ratio of international staff and ratio of international students (QS World University Ranking, 2018). For a more in-depth overview of the QS methodology see the official QS website (www.topuniversities.com), or articles from Huang (2012), Docampo (2011), Vernon et al. (2018), and Marginson (2014).

The three rankings use overlapping concepts (teaching quality, research quality) but diverse input variables to evaluate these concepts—see Table 1. Next to these overlapping concepts the rankings also have unique characteristics. Noticeable is the inclusion of internationality in the THE and QS ranking. This is absent from the ARWU ranking. Also, the THE is the only ranking to include a university's income from industry. Furthermore, the THE and QS rankings apply corrections to normalize the citation scores with respect to the size of a university, while the ARWU includes uncorrected counts to measure research quality and quantity. This ranking only corrects for university size for institutions in specific countries using the PCP variable. In general, it can be stated that the methodologies of the THE and QS ranking are quite similar. They use comparable concepts for assessing the quality of a university and similar methodologies for measuring them. The ARWU ranking, while partly measuring the same concepts, uses different variables and input data to operationalize these concepts.

3. DATA COLLECTION

Data for this study have been collected from the official websites of the three URs. Data have been retrieved for all variables that form the rankings described in the previous section by scraping the university ranking websites.

Table 1. Comparing the indicators in the three rankings

	ARWU	THE	QS
Teaching Quality	Alumni & PCP	Teaching	Faculty Student Ratio
Research Quality	Award & HiCi	Research & Citations	Citations per faculty
Research Quantity	NS & PUB	–	–
Internationality	–	International outlook	International faculty ratio & International student ratio
Industry Reputation	–	Industry	Employer reputation

Table 2. Number of universities measured per year

Year	ARWU	λ	THES	λ	QS	λ	All
2012	500	500	400	364	869	392	324
2013	500	498	400	367	903	400	326
2014	500	497	401	381	888	395	326
2015	500	498	800	763	918	96	413
2016	500	497	981	981	936	140	405
2017	500	497	1,103	1,103	980	129	414
2018	500	497	1,258	1,258	1,021	498	419

This research focuses on the ranking years 2012 to 2018 because for these years it was possible to obtain data from the website of all selected rankings. This ensures that for all years analyzed, official data about all three rankings are available. Table 2 shows the number of universities present in the rankings per year. The lambda (λ) column shows the number of universities present in each respective ranking for which all data measured by the rankings is available. The last column (All) shows the number of institutions that are present in all three of the rankings in that specific year.

Different rankings use different names for universities, and also within a ranking name changes were observed over the year. Subsequently, to compare universities between years and rankings it was necessary to link all universities to their associated Global Research Identifier Database entry (GRID) (Digital-science, 2019). Records were linked using data retrieved from Wikidata (Wikidata Contributors, 2019). Wikidata includes the IDs that are assigned by the three rankings for many universities alongside the related GRID. By linking an institution's unique ranking ID to the Wikidata database and extracting the relevant GRID, it was possible to match almost all universities. This linkage proved effective; manual inspection of several universities did not detect mismatches. A small number of missing GRIDs were linked by hand.

4.   EXPLORATORY ANALYSIS

A comparison of the changes in overall positions of universities' rankings is now presented. Two distinct aspects are assessed: changes in the rankings over time and the dissimilarities of the three rankings in the same year with respect to each other.

Three different measurements are used to evaluate these relationships. The first is the number of overlapping universities (O) (the number of universities that are present in both rankings). The second is the Spearman rank correlation coefficient (F), which measures the strength of the association between overlapping universities (Gravetter & Wallnau, 2016). To assess the relationship between rankings including nonoverlapping universities, a third test, the inverse rank measure (M), as formulated by Bar-Ilan, Levene, and Lin (2007), is calculated. This test is also used to compare rankings in the research of Aguillo et al. (2010). The M-measure assesses ranking similarity while factoring in the effect of nonoverlapping universities. This is accomplished by assigning nonoverlapping elements to the lowest rank position +1. In the case of two URs with size k, if a university appears in ranking A but does not appear in ranking B, then the university is assigned to rank $k + 1$ in ranking B. The M-measure subsequently calculates a normalized difference between the two rankings (Aguillo et al., 2010). The resulting M-scores

should be interpreted as follows: Below 0.2 can be considered weak similarity, between 0.2 and 0.4 low similarity, between 0.4 and 0.7 medium similarity, between 0.7 and 0.9 high similarity and above 0.9 very high similarity (Bar-Ilan et al., 2007). Some universities were assigned the same position in the rankings because of a tie in their scores. These universities were assigned to the mid position (i.e., two universities that are ranked fifth are both assigned to place 5.5).

4.1. Longitudinal Ranking Stability

First, changes within rankings over the past 7 years are reviewed. In Table 3 the number of overlapping institutions, Spearman correlation coefficients and the M-measure scores are listed for the top 100 institutions in each ranking. This table shows, for each ranking, the years from 2013 to 2018, as indicated in the left column. For each of these years, each ranking is

Table 3. Similarity between ranking years (O: Overlap; F: Spearman correlation coefficient, M: M-measure)

| Measure	2012			2013			2014			2015			2016			2017		
	O	F	M	O	F	M	O	F	M	O	F	M	O	F	M	O	F	M
ARWU																		
2013	98	1.0	0.97															
2014	92	0.98	0.97	93	0.98	0.96												
2015	93	0.97	0.97	93	0.98	0.95	97	0.99	0.99									
2016	89	0.91	0.91	89	0.93	0.93	92	0.95	0.93	91	0.97	0.93						
2017	87	0.91	0.9	87	0.91	0.9	90	0.94	0.92	89	0.96	0.92	95	0.97	0.94			
2018	85	0.89	0.89	85	0.89	0.90	87	0.93	0.91	86	0.95	0.91	94	0.96	0.93	95	0.97	0.97
THE																		
2013	92	0.98	0.93															
2014	91	0.96	0.89	93	0.98	0.94												
2015	83	0.88	0.89	86	0.88	0.86	85	0.90	0.85									
2016	85	0.90	0.78	87	0.90	0.75	85	0.92	0.74	89	0.97	0.84						
2017	85	0.91	0.71	86	0.89	0.7	85	0.91	0.68	89	0.96	0.77	96	0.98	0.91			
2018	85	0.91	0.69	82	0.90	0.67	81	0.92	0.66	86	0.95	0.76	91	0.97	0.88	94	0.98	0.95
QS																		
2013	95	0.98	0.91															
2014	94	0.96	0.88	97	0.98	0.9												
2015	93	0.90	0.81	93	0.91	0.87	93	0.92	0.82									
2016	91	0.90	0.8	90	0.90	0.82	90	0.91	0.81	95	0.98	0.93						
2017	87	0.88	0.79	87	0.89	0.8	87	0.90	0.79	91	0.97	0.9	94	0.99	0.96			
2018	86	0.88	0.77	84	0.88	0.78	84	0.89	0.77	89	0.96	0.87	92	0.98	0.93	93	0.98	0.95

Note: All Spearman correlations (F) were significant: $p < .001$.

compared to the data for the years 2012 to 2017 (first row) of the same ranking. Here a comparison of only the top 100 universities is presented because the ARWU ranking assigns a singular rank only to universities in the top 100 of the ranking.

These analyses show that all three rankings are stable over time. A large portion of universities are overlapping for every year. Furthermore, all Spearman correlation coefficients are significant, with large effect sizes. This signifies that there is not much change in the ranking positions of overlapping universities between years. This is also demonstrated in Figure 1. The M-measures provide some more insights into the changes in the rankings over time. For the ARWU ranking this measurement shows strong similarities. Even when comparing the ranking from 2012 with that from 2018 the M-measure is very high, and only 15 universities do not overlap. The THE ranking is more volatile. For example, similarities between the 2018 and 2017 rankings and those from 2012, 2013, and 2014 are less strong. This may be connected with the shift from using Web of Science (WoS) to Scopus as a source of citation data between 2014 and 2015, and if so is indicative of a sensitivity to data sources. However, when considering that the number of universities ranked by the THE ranking is three times higher in 2018 than in the earlier years, the relationship between them is still quite high. Also, the M-measure between consecutive years shows strong similarities. The change that is present is thus subtle. The QS ranking is also very similar over the years. Consecutive years show very high similarity. But the latest ranking also shows high similarity with all previous years, with an M score of 0.77 indicating high similarity when comparing the 2012 ranking with the one from 2018.

Overall, our conclusion is that the rankings are very stable over time. The top 100 institutions of all rankings are significantly correlated between all years and the M-measure also shows very strong similarities between most years. From the three rankings the THE showed most (albeit subtle) change over time; it is the only ranking in which the M-measure showed a medium similarity between some years. From these results the conclusion can be drawn that universities in the top 100 are largely fixed. There are not many new institutions that enter, and consequently few institutions that drop out of, the top 100. Additionally, within the top 100 of each ranking there is little change in position between years. A comparison where more institutions are taken into account can be found in Section A of the supplementary material; see Selten (2020). In general, the results of this analysis do indicate that rankings are stable beyond the top 100. However, as is explained there, these results should be interpreted with care. This stability can be explained by the fact that the rankings use rolling averages to measure publications and citations. Furthermore, the ARWU ranking includes prizes won since 1911. In all rankings, subsequent ranking years are thus partly based on the same data. The fact that it is hard to move positions
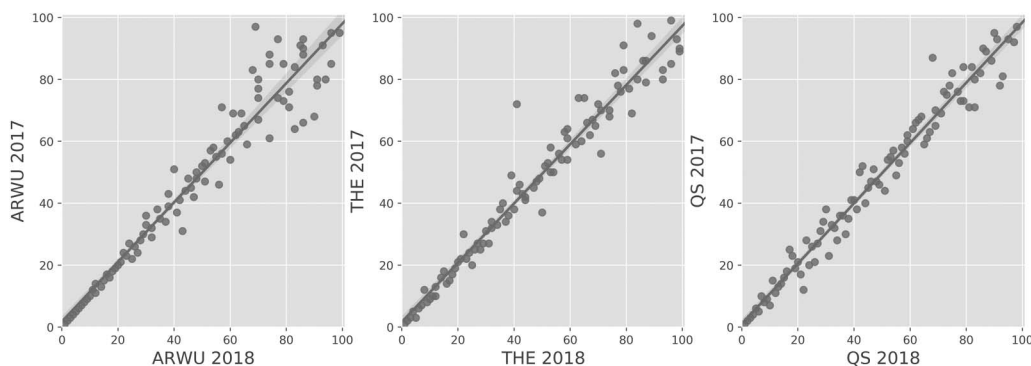
Figure 1. Similarity between ranking years.

at the top of the rankings, despite the differences between them, is also consistent with the idea that the rankings may be having a circular effect, reinforcing the positions that universities hold. This effect is likely to be strongest in the THE and QS rankings because these use reputation surveys, which are likely to be influenced by previous rankings. However, research by Safón (2019) shows that previous rankings might also influence research performance, indicating there might an additional circular effect present in the ARWU ranking. These processes are further elaborated on in Section 6.1.

4.2. Similarity Between Rankings

Next, we review the similarity between rankings. For each year, the three rankings are compared to each other. The same three measurements are used to test these relationships. However, as well as analyzing the top 100 universities, the similarities between the top 50 and 50 to 100 range of the rankings are independently examined. The results of this analysis are shown in Table 4. Comparisons are given for each year analyzed between the THE and ARWU, and between the QS and ARWU, and the QS and THE rankings. We observe no large discrepancies between years in how similar the rankings are with respect to each other. This is as expected because each ranking does not change much over time.

There is, however, a difference between the top 50 and positions 51 to 100. The overlap measurement in the top 50 of each ranking shows that 60–70% of the universities overlap between rankings. The ranks of these overlapping universities are also significantly correlated. However, the M-measure shows medium similarity, caused by the relatively high amount of nonoverlapping universities. In the 50 to 100 range the similarity between the rankings is very weak. Not even half of the universities overlap and the correlations between the rankings in all years, except for the correlation between the ARWU and THE rankings in 2013, are not significant. The M-measure also shows weak to very weak similarities between rankings in this range.

In the top 100 the THE and ARWU rankings and THE and QS rankings overlap for more than 70 universities. Between the ARWU and QS there is a little less overlap. This also results in an M-measure that is lower than that between the QS and THE and the ARWU and THE rankings. However, all M-measures can be classified as being of medium strength. Furthermore, the Spearman correlation is significant for all comparisons for the top 50 and top 100. The M-measure indicates more similarity at the top 100 level than at the top 50 level. This is caused by the fact that the M-measure assigns more importance to the top of the rankings, and when comparing the top 100 range there are fewer nonoverlapping universities (i.e., universities that are in the top 50 of one ranking but not in the top 50 of the other ranking are likely to be in the top 100 of the other ranking).

Generally, the top 50 and top 100 between all rankings are quite similar. The M-measure points out medium relationships, but the correlations between the ranks of overlapping universities are strong and significant. The 50 to 100 range displays much more difference between the rankings. Not even half of the universities are overlapping, the ranks of overlapping universities are not significantly correlated, and the M-measures show very weak similarity; this is also visible in Figure 2. Finally, no two rankings were clearly more similar to each other than to one other ranking. Comparing these two sets of plots clearly demonstrates that different years of the same ranking are very similar. There is much more variance when comparing ranks of similar universities in different rankings, especially amongst the higher ranking positions. In Section B of the supplementary material (see Selten, 2020) we show a similar analysis for the top 400 institutions. These results need to be interpreted with care but show that in the top 400 there is also

Table 4. Similarity between different rankings (O: Overlap; F: Spearman correlation coefficient; M: M-measure)

	Top 50						50–100						Top 100					
	ARWU			THE			ARWU			THE			ARWU			THE		
Measure	O	F	M	O	F	M	O	F	M	O	F	M	O	F	M	O	F	M
2012																		
THE	37	0.83***	0.57				19	0.22	0.13				70	0.81***	0.59			
QS	28	0.62***	0.48	36	0.77***	0.49	16	0.0	0.17	15	−0.27	0.05	64	0.68***	0.51	74	0.74***	0.53
2013																		
THE	35	0.82***	0.58				21	0.46*	0.13				73	0.83***	0.60			
QS	29	0.66***	0.51	36	0.73***	0.52	15	0.28	0.10	14	−0.02	0.08	63	0.69***	0.53	72	0.72***	0.56
2014																		
THE	38	0.83***	0.60				22	−0.19	0.15				71	0.78***	0.62			
QS	32	0.72***	0.48	36	0.70***	0.50	16	0.22	0.16	15	0.06	0.21	62	0.65***	0.50	72	0.71***	0.54
2015																		
THE	35	0.75***	0.51				20	0.24	0.18				68	0.79***	0.54			
QS	29	0.7***	0.57	35	0.80***	0.55	16	0.34	0.09	17	0.34	0.13	61	0.70***	0.57	76	0.72***	0.58
2016																		
THE	37	0.83***	0.54				21	−0.35	0.12				74	0.71***	0.57			
QS	31	0.67***	0.58	38	0.77***	0.56	17	−0.21	0.12	17	0.2	0.13	66	0.63***	0.58	76	0.75***	0.59
2017																		
THE	36	0.81***	0.55				20	−0.1	0.09				73	0.75***	0.57			
QS	30	0.66***	0.57	37	0.73***	0.55	13	0.43	0.05	13	−0.2	0.07	63	0.64***	0.58	72	0.75***	0.58
2018																		
THE	36	0.83***	0.56				22	−0.3	0.11				75	0.73***	0.59			
QS	31	0.68***	0.56	40	0.78***	0.56	14	0.14	0.08	17	0.3	0.12	64	0.64***	0.57	74	0.8***	0.6

Note: * $p < .05$, ** $p < .01$, *** $p < .001$.
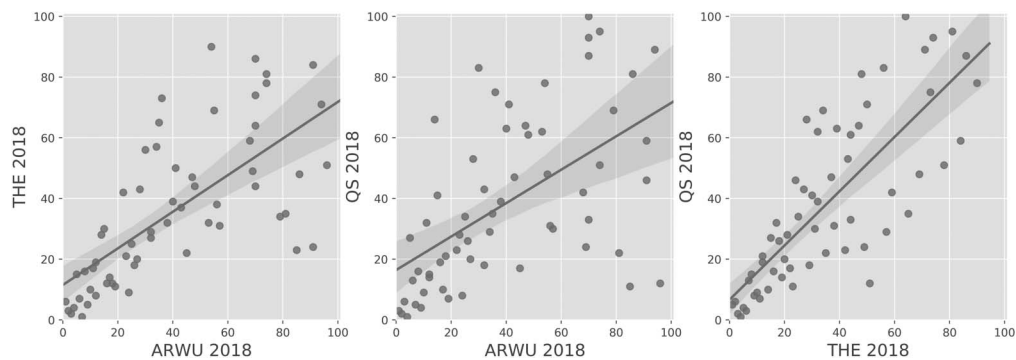
Figure 2. Similarity between different rankings.

quite strong similarity between rankings. At the same time, however, a rather large number of nonoverlapping institutions is present, resulting in medium M-scores.

5. FACTOR EXTRACTION

The three rankings use overlapping concepts (teaching quality, research quality) but diverse input variables to evaluate these concepts; see Table 1. The above findings show that the rankings do not vary much over time but that the similarity between rankings is less and differs according to the ranking position analyzed. We now take a more in-depth look at the input measures of the rankings.

Previous research suggests that there are two latent factors underlying the ARWU ranking of 2008 and three underlying the THE ranking in 2013 (Dehon et al., 2009; Safón, 2013). To further examine the similarities and differences between rankings, we analyze whether these factors are stable in the rankings over time. This was done using two techniques: PCA, which has been employed by Dehon et al. (2009) and Docampo (2011), and EFA as used by Safón (2013) and Soh (2015). The studies of Dehon et al. (2009) and Safón (2013) only reviewed a subset of the ranking data by studying the top 150 or a group of overlapping universities. We are interested in comparing the overall structure of the rankings over multiple years. Therefore, all universities present in the rankings are analyzed. Only universities for which the rankings do not provide information on all input measures are removed, because PCA and EFA cannot be applied to missing values. The number of universities that were analyzed each year can thus be seen in the lambda columns in Table 2. All input measures analyzed were scaled to have unit variance.

Although PCA and EFA are related and often produce the same results, the application of both techniques has two advantages. First, the university ranking data show multivariate outliers. Results from both the PCA and EFA will be influenced by this. Therefore, for both analyses robust techniques are implemented. By applying two methods we can have more confidence that the extracted factors are genuine. Furthermore, PCA and EFA measure different relationships. PCA describes how much of the total variance present in the data can be explained by the extracted components. EFA tries to explain the correlations between variables and only considers shared variance (Osborne, Costello, & Kellow, 2008). Therefore, when observing correlated variables using EFA that together explain a substantial part of the variance as indicated by PCA, there is a strong indications that the input measures are related to a latent concept.

5.1. Principal Component Analysis

PCA is implemented using a robust method as formulated by Hubert, Rousseeuw, and Vanden Branden (2005) using the rrcov package for R (Todorov, 2012). All R and Python scripts used to perform the analyses in this study can be found at Selten (2020). This method is robust against the influence that outliers will have on the PCA results. The loadings of the PCA were oblique rotated, because the analyzed variables are expected to be correlated and to make the results easier to interpret. To confirm that this method produces sensible results, the robust PCA method was tested on the top 150 universities from the ARWU ranking of 2008 in an effort to reproduce the results from the analysis of Dehon et al. (2009). Using this PCA method a comparable loading structure to that of Dehon et al. was found, namely that the ARWU consists of two components where the first component is loaded by the Alumni and Award variables and the second component by the other three variables (NS, HiCi and Award). This confirms that the method we use in this study is comparable to that used by Dehon et al. (2009).

The same analysis was carried out for all years from 2012 to 2018 and for all three rankings. For each ranking the results of the analysis on the 2018 rankings are described in depth. Following this, the changes in structure in the other years with respect to the structure observed for 2018 are discussed (the loading structures for these years can be found in Section C of the supplementary material; see Selten, 2020). Because the components can change order we do not refer to them numerically but name them component *A*, *B*, and *C*. Furthermore, for the ARWU ranking we decided to remove the PCP variable from the analysis because of how this variable is constructed: It does not measure a separate concept but is a composite of the other ARWU variables and applies a correction for the size of a university. However, this correction is only performed for universities in specific countries (see Academic Ranking of World Universities (2018) for a list of these countries). For universities outside these countries, the PCP variable measures a weighted score of the other five variables. Removing this variable for these reasons is common because interpretation of the variable is not feasible; it does not measure a single concept and has different meanings for universities in different countries (Dehon et al., 2009; Docampo, 2011).

A rule often used in PCA is to keep only components with eigenvalues exceeding one, because this indicates that the component explains more variation than a single variable (Kaiser, 1960). Extracting eigenvalues for the ARWU ranking proved that this was only true for the first principal component. However, this rule does not always provide a good estimate for the number of components to retain (Cliff, 1988; Velicer & Jackson, 1990). Inspection of the scree plots, prior research, and assessment of the results when keeping one and two components justified extracting the first two components from this ranking (Dehon et al., 2009). For the THE ranking, the first two components had an eigenvalue higher than one, and for the QS ranking the first three components had an eigenvalue exceeding one. Scree plots and analysis of the results confirmed that extracting two and three principal components respectively was justified. The results of this analysis for the 2018 ranking data can be seen in Table 5.

These results show a clear structure in the ARWU ranking. The Alumni and Award variables represent component B and the HiCi and PUB variables component A. The NS variable loads on both. This structure is also observed in the years 2016 and 2017. In the years 2012, 2013, 2014, and 2015 the Alumni, Award, NS, and HiCi variables load on one component, while only the Pub variable loads on the other component.

In the THE ranking we also observe two components. One input variable (Research) loads on both components, while the other four variables load distinctively on one of the two components. The Research variable loads components A and B. Component A is also influenced by the Citations and International Outlook variables. Component B gets additionally influenced by the Teaching and Industry Income variables. This structure is also observed in the years 2016 and 2017. Before 2016 there is variability in the loading structure. In the years 2012, 2013, and 2014 the Teaching and Research variables load strongly together on component A and the International Outlook and Industry Income variables load on component B. Citations load on both components. The year 2015 is divergent from the other years: In this year the Citation and International Outlook variables influence component B, and Industry Income explains a big proportion of the variance in the other component. Teaching and Research in that year load on both components.

For the QS ranking, a clearer distinction between components can be observed. The Academic and Employer Reputation variables represent component A. International Faculty and Students represent component B. Finally, the Faculty Student and Citations variables form component C. The QS ranking also showed the most stability over time. The first components

Table 5. Rotated PCA Loadings on Components 2018

Measure	PC-A	PC-B	PC-C
ARWU			
1. Alumni	0.04	**0.66**	
2. Award	−0.03	**0.48**	
3. HiCi	**−0.66**	−0.02	
4. NS	−0.36	0.35	
5. PUB	**−0.84**	−0.01	
THE			
1. Teaching	−0.28	**−0.45**	
2. Research	**−0.42**	**−0.45**	
3. Citations	**−0.89**	−0.12	
4. Industry Income	0.07	**−0.88**	
5. International Outlook	**−0.96**	0.12	
QS			
1. Academic reputation	**−0.99**	−0.05	0.03
2. Employer reputation	**−0.92**	0.04	0.10
3. Faculty Student	−0.15	0.04	**0.92**
4. International Faculty	0.01	**0.94**	−0.08
5. International Student	0.02	**0.95**	0.10
6. Citations	**−0.55**	0.14	**−0.58**

Note: Loadings larger than .40 are in bold.

A and B are the same in all years analyzed. However, the Faculty student variable in 2016 also loads on component A. The Citation variable is most volatile and loads differently across years.

For each of the three rankings, the robust PCA showed that it is possible to reveal structure in the data. Some variables are stable and load on the same component in all years. However, there are also variables that show more variation.

5.2. Exploratory Factor Analysis

To explore the factorial structure of the data further, an EFA using oblique rotations was performed. First, for all three rankings in all years the Kaiser-Meyer-Olkin measure (KMO) must be verified to test sampling adequacy, and Bartlett's test of sphericity (χ^2) needs to be performed to analyze whether the correlation structure of the data is adequate for factor analyses.

The tests indicate that all years of all rankings are adequate for factor analysis. For ARWU in all years KMO > 0.80 and Barlett's χ^2 test is significant ($p < 0.001$). For THE in all years KMO > 0.55 and Barlett's χ^2 test is significant ($p < 0.001$). For all years of the QS KMO > 0.52 and χ^2

test is significant ($p < 0.001$). The KMO values for the THE and QS ranking are quite low. This indicates the existence of relatively high partial correlations between the variables in these two rankings (Field, 2013; Pett, Lackey, & Sullivan, 2003). This shows that there is less unique variance in the THE and QS rankings compared with the ARWU. The existence of high partial correlation in URs is to be expected. The ranking variables attempt to measure university performance, so it is therefore not surprising that the ranking variables, at least partly, account for common variance. Higher KMO values for the ARWU indicate that fewer partial correlations exist in this ranking. The variables in the ARWU, thus, capture more unique variance. Here, it should be noted that KMO assumes normally distributed data and the ranking data diverts from this. It is useful to test the KMO statistic, but one should not place too much emphasis on this test. That being said, given that this research is of an exploratory nature and in all years the KMO values exceed a minimum value of 0.50, it is possible to perform factor analysis on the data for all years of all three of the rankings (Field, 2013; Hair, Black, Babin, & Anderson, 2014; Kaiser, 1974).

The principal axis factors (PAF) extraction method was used because the data deviate from multivariate normality. PAF is the preferred extraction method in this situation (Osborne et al., 2008). The noniterated version was used because the iterated solution yielded Heywood cases, a common problem when using the iterated version of this method (Habing, 2003). The same number of factors were extracted as the number of extracted components in the PCA. Scree tests are also a viable strategy for determining the number of factors to retain in factor analysis and a parallel analysis supported the amount of factors to extract. The results of this analysis for the 2018 ranking data can be found in Table 6; for the other years see Section D in the supplementary material (Selten, 2020).

These results generally follow those obtained with PCA, with the structure being more clear. The ARWU consist of two distinct factors: Factor A is loaded by the Alumni and Award variables and factor B strongly by the HiCi, NS, and PUB variables. This structure is visible in all years. The THE ranking is also made up of two factors. Factor A is loaded by the Teaching, Research, and Industry Income variables, whereas factor B is constructed of the Citations and International Outlook variables. This structure is visible in 2015, 2016, and 2017. In 2012, 2013, and 2014 factor A is not loaded by the Industry Income variable. Factor B in those years is only loaded on by the Citations variable and not by International Outlook. The QS ranking is made up of three factors. Factor A is loaded by the Academic and Employer Reputation variables, factor B by the International Faculty and Students variables, and factor C by the Faculty Student and Citations variables. The QS ranking shows more volatility than the other two rankings. In all years analyzed, Factors A and B are respectively loaded on by the reputation variables and the two variables measuring internationality, but there is variation in how the Faculty Student and Citations variables load. In the years 2012, 2013, and 2014, factor C was loaded on only by the Citations variable and the Faculty Student variable did not load on any of the factors. In 2015 and 2016, both the Faculty Student and Citations variables loaded on factor A together with the Academic and Employer Reputation variables. In 2017, both Citations and Faculty Student did not load higher than .4 on either of the three factors.

5.3.   Explaining the Factors

The structure in the three rankings was evaluated using two different methods: robust PCA and EFA. The first method is robust against the presence of outliers in the data, while the second is resistant against the data being nonnormally distributed. We now examine whether the factors

Table 6. NIPA loadings on factors in 2018

Measure	PA-A	PA-B	PA-C
ARWU			
1. Alumni	**0.84**	0.00	
2. Award	**0.86**	0.01	
3. HiCi	0.01	**0.79**	
4. NS	**0.41**	**0.58**	
5. PUB	−0.09	**0.75**	
THE			
1. Teaching	**0.92**	−0.02	
2. Research	**0.86**	0.15	
3. Citations	0.16	**0.66**	
4. Industry Income	**0.63**	−0.20	
5. International Outlook	−0.04	**0.73**	
QS			
1. Academic reputation	**0.88**	−0.06	0.07
2. Employer reputation	**0.83**	0.08	−0.08
3. Faculty Student	0.24	−0.01	**−0.40**
4. International Faculty	−0.02	**0.75**	0.07
5. International Students	0.02	**0.76**	−0.06
6. Citations	0.26	0.13	**0.44**

Note: Loadings larger than .40 are in bold.

that were empirically found by these two analyses, are also theoretically explainable and what underlying concepts these factors measure.

In the ARWU ranking, two distinct factors can be observed in the EFA, whereas the PCA shows more volatility. Generally, however, it can be stated that the HiCi, PUB, and N&S variables appear to form a factor together and the Alumni and Award variables form a second factor. This structure was also found in the research of Soh (2015) and Dehon et al. (2009). The first factor measures the number of citations and publications and together is weighted 60% on the ranking. The variables that form the second factor, Alumni and Award, measure the number of Nobel Prizes and Fields Medals won by a university's employees or alumni and is weighted 30% in the ARWU ranking. Safón (2013) came to a different conclusion, showing that all ARWU variables load on the same factor. This study, however, used a specific subset of the data, which had a significant effect on the extracted structure.

In the THE ranking, two distinct factors also are extracted in both the PCA and EFA. The first factor is composed of the Teaching and Research variables. These two variables are measured by multiple subvariables, as described in Section 2. Only in the years 2016 to 2018 do we see

this reflected in the results of the robust PCA. In these years the research variable loads on both components. This may be a reflection that this variable, when correcting for the influence of outliers, is derived from two quite different notional indicators. However, when assessing all years and the EFA results, we, in accordance with the interpretation of Moed (2017), expect that the Teaching and Research variables loading together is caused by the influence of the surveys, and the other variables used to construct these variables have little impact because of the low weights assigned to them. This component is therefore mainly a representation of a university's reputation and accounts for 60% of the ranking. The second component, when considering all years, is influenced mainly by the Citations variable, which provides 30% of the final ranking. There is quite some variation in how the Industry Income and International Outlook variables load. These are not clearly related to a single factor, and both weigh only 5% on the ranking. The research of Safón (2013) and Soh (2015) shows comparable results. However, in these studies the Citations variable loaded with the Research and Teaching variables. Our results suggest that, when taking the whole ranking into account over multiple years, the Citation measure is a separate factor in the THE ranking.

The QS ranking is the only ranking for which the extraction of three factors proved useful according to scree plots and parallel analysis. However, when considering multiple years, only two are consistent. The Academic and Employer Reputation variables load together in both PCA and EFA. This suggests, as in the THE ranking that they are a measure of the general reputation of a university. This factor provides 50% of the ranking. Also, the International Faculty and International Students variables form a construct together. This factor accounts for 15% of the weight in the ranking. The last extracted factor in the QS ranking was not consistent. Both Citations and Student to Staff ratio thus appear to be separate components in this ranking when analyzing multiple years of the QS ranking. They both provide 20%. These results differ quite a bit from those obtained by Soh (2015), which might be caused by the fact that that study only extracted two factors.

Reviewing these results and assessing what the variables that form the concepts measure shows that in all three rankings in all years there are two overlapping underlying concepts that contribute substantially to the rankings: (a) reputation and (b) research performance.

In the ARWU ranking, we observed that the N&S, HiCi, and PUB variables often load together. These variables are all proxies for the research performance of a university. The second component is composed of the Alumni and Award variables. Both these variables measure the same achievements but in different groups, and can be seen as a proxy for or influencer of, as indicated by the work of Altbach (2012), a university's reputation. In the THE ranking, reputation is measured by the Teaching and Research variables, while the Citations variable is measuring research performance. In the QS ranking, Academic and Employer Reputation comprise the reputation factor, whereas research performance is measured by the Citations variable.

Also, some nonoverlapping concepts were found. PCA and EFA showed that internationality is a separate concept in both the THE and QS rankings, and in the ARWU this concept is not represented. Also, in the QS ranking the student-to-staff ratio plays quite an important role. In the other two rankings, this concept is not assigned much importance.

When taking the weights assigned to the variables into account, 90% of the ARWU ranking, 85% of the THE ranking and 70% of the QS ranking are accounted for by the two concepts. Reputation and research performance, are thus very influential in all three rankings. A final difference that can be observed in the rankings is that in the ARWU ranking indicators of research performance are more important, while in the THE and QS rankings the indicators associated with reputation are the most influential.

Table 7. Spearman-Brown scale reliability

	ARWU		THE	QS
Scale	1	2	1	1
2012	0.86	0.88	0.95	0.82
2013	0.87	0.88	0.95	0.83
2014	0.87	0.89	0.96	0.83
2015	0.87	0.88	0.95	0.88
2016	0.87	0.84	0.95	0.84
2017	0.87	0.83	0.95	0.84
2018	0.87	0.83	0.95	0.89

5.4. Reliability of the Concepts

The analysis described above concluded that there are two overlapping concepts: (a) reputation and (b) research performance, that represent most of the weight in each of the rankings. In the ARWU ranking, both concepts are a combination of multiple variables. In the THE and QS rankings, only the reputation measurement is a multi-item concept. To confirm that the variables that measure one concept together form a reliable scale, the internal validity of the scales was verified. The Spearman-Brown split-half reliability test was used for this because some concepts are composed of two variables (Eisinga, Te Grotenhuis, & Pelzer, 2013). Spearman-Brown reliability is calculated by splitting and correlating the items that form the scale. This reliability can be interpreted similarly to Cronbach's alpha: Scores closer to one demonstrate that the scale is internally more reliable (Field, 2013, pp. 1044–1045). The results of these tests can be found in Table 7. They confirm that in all years the scales are internally reliable. This supports the assertion that for all three rankings the factors that consist of multiple variables are reliable scales measuring the same concept across multiple years. Furthermore, for the THE and QS ranking it can be observed that these scales are more internally reliable when compared to internal reliability for the whole ranking tested using Cronbach's alpha (see supplementary material Section E [Selten, 2020]), and in the ARWU the reliability of the scales is comparable to the internal consistency of the complete ranking. This indicates that, while our analysis shows the existence of two internal reliable scales in the ARWU ranking, these concepts are more interrelated than is the case for the THE and QS rankings. This is consistent with the finding that the ARWU ranking is mostly a one-dimensional scale assessing academic performance, while the other two rankings are more multidimensional (Safón, 2013; Soh, 2015).

6. INVESTIGATING THE SCALES

Based on our analysis to this point, we conclude that two concepts underlie all three rankings. To further investigate what these concepts measure, the variables of which they consist were combined. For each ranking, this creates a two-dimensional representation of each ranking describing the reputation and research performance of the universities.

6.1. Testing Scale Relationships

To assess the relationship between these concepts a Spearman correlation test for each year was performed. Results can be found in Section F of the supplementary material (Selten, 2020).

These show that all concepts in all years are significantly correlated with each other. Across years the THE and QS reputation performance measurements seem to be correlated most strongly, but also the THE reputation and ARWU research performance concepts show strong correlation. These differences are, however, only minor—in general the reputation performance concept of each ranking is not evidently correlated more strongly with the reputation performance concept of the other rankings than with the research performance concept of the other rankings, and vice versa.

We identify two potential explanations for the existence of a relation between these different concepts. First, it is important to note that, while the rankings aim to quantify similar aspects of a university's performance, this performance is measured using different methodologies. For example, as noted in Section 2, the THE and QS rankings correct the research performance measurements for the size of universities, while the ARWU ranking does not normalize these measurements for university size. Furthermore, while the THE and QS rankings use direct measures to capture reputation, the reputation concept for the ARWU ranking is more ambiguous. The number of Nobel Prizes and Fields Medals won was in this study interpreted as a measurement of reputation, but these prizes are only an indirect indicator of reputation, as they also demonstrate scientific accomplishments. Similarly, when a university's staff often publish in high-impact journals or are cited a lot this can also be seen as a proxy for the reputation of a university. This leads us to the second explanation: The existence of a circular effect being present in the rankings. Safón (2013) demonstrates this effect by showing the existence of a reputation-survey-reputation relation in the THE and QS rankings as well as in the ARWU ranking, even though it does not include reputation surveys. However, Robinson-Garcia, Torres-Salinas, et al. (2019) in their study reverse this argument. They hypothesize that the answers people give on surveys are influenced by publication and citation data.

Both interpretations can help explain the results found in this study. We identify the existence of two latent concepts in all rankings: reputation and research performance. However, these two latent concepts might be influencing each other. In the next section the relationship between the concepts is further investigated.

6.2. Plotting the Scales

The correlation coefficients themselves do not provide much insight into how the different components and factors (*scales* for the rest of this discussion) relate to each other. However, they are a two-dimensional reflection of the most important concepts in all three rankings. We were interested in whether using these scales as coordinates to map the relationship of research performance and reputations over time for each ranking would provide insight. In particular, we are interested in the question of whether there are differences in the progress made by universities in different regions and by language spoken, whether this could provide evidence for or against claims of bias in the rankings, and if it could provide evidence for or against the circular reinforcement effects discussed above.

To ease interpretation of the plots, they were created using a subset of universities that are present in all rankings. This results in a subset of 87 high-ranked universities. Also, there is a difference in the number of universities that are ranked per region. This would skew the comparison between regions. Because in the rankings the top institutions are most important, we chose to only aggregate the results of the top five institutions in each region. Figures 3–5 show the movement of (aggregates) of universities on the two scales: reputation and research performance. Each arrow indicates the data point for a given year, starting in 2012 and ending in 2018. The arrow direction shows the movement from year to year.
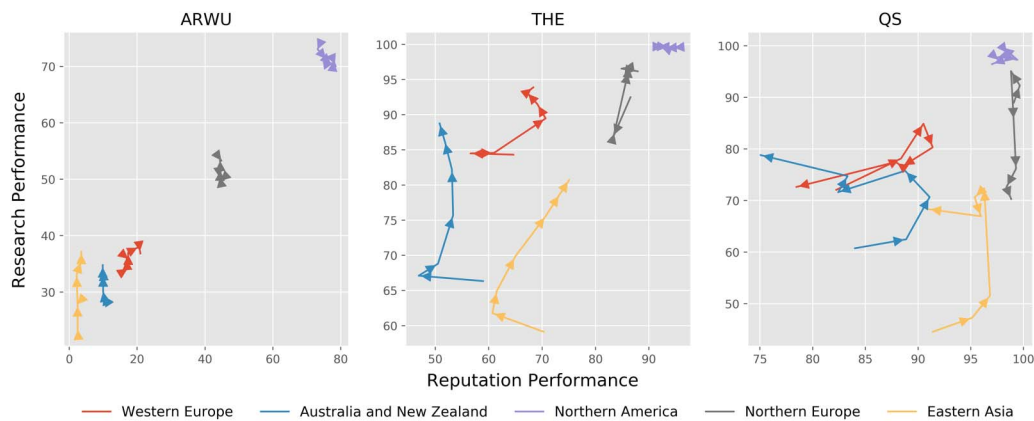
Figure 3. Longitudinal developments per geographical region.

Figure 3 shows how the three rankings behave on a regional level. There are differences in the relative rankings of universities from different regions. We therefore plot the average of the top five institutions from each region. North America, in the ARWU, is far ahead of all other regions on both the reputation and research performance scales. South-Eastern Asia, Eastern Asia, Western Europe, Australia, and New Zealand are all far behind. Northern Europe appears right in the middle. The THE and QS rankings both also show that the top institutions in North America perform best on both scales. However, the advantage with respect to the other regions is much smaller. Northern Europe performs second best on both scales in both rankings, but in the THE and especially QS ranking, Asian universities also perform very well on the reputation measurement. Another interesting observation from this figure is that in all rankings Asian universities are climbing fast on the research performance scale. Finally, universities in Western Europe and Australia and New Zealand in the THE and QS rankings seem to have quite a low reputation score when compared to their score on the research performance scale, whereas in the ARWU ranking this is the case for institutions in Eastern Asia; performance on the research scale for universities in this area rose quickly, but they continue to lag behind on the reputation measurement. The ARWU shows strikingly lower movement on the reputation scale than do the other rankings, indicating the slow accumulation of prizes compared to the volatility or responsiveness of a survey-based measure.
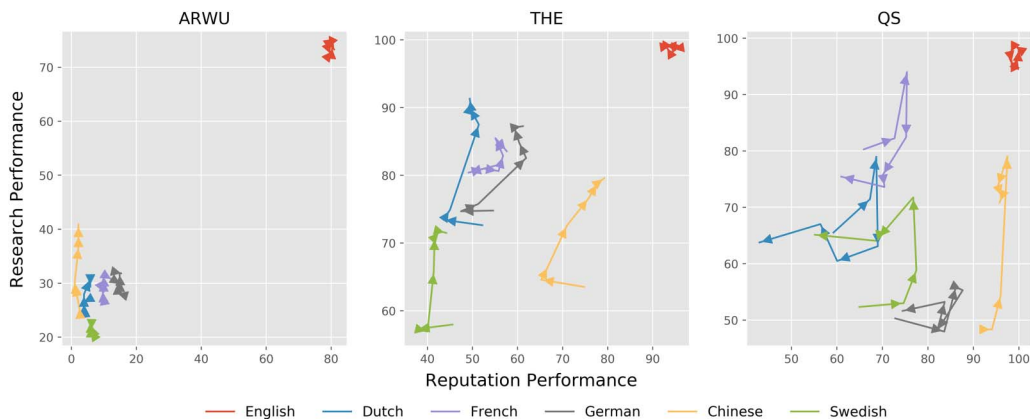

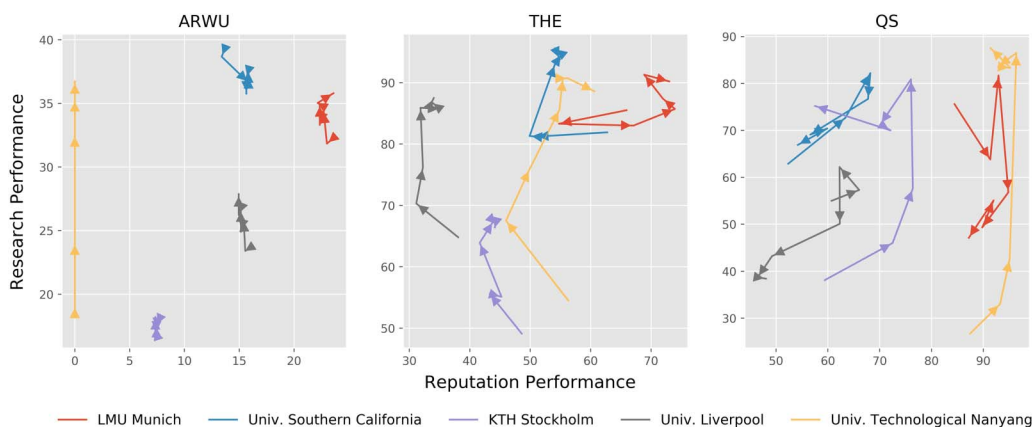**Figure 4.** Longitudinal developments per language region.

Figure 5. Longitudinal developments for a sample of universities.

In the second set of plots (Figure 4), an aggregate of the top five universities within certain language regions is displayed. This shows in all three rankings that universities in English-speaking countries are ahead on both the reputation and research performance scales. Of all three rankings, the ARWU shows the biggest difference between English-speaking countries and the other language regions. In the THE ranking on the research performance scale, universities in Dutch-, French-, and German-speaking countries perform equally and are around 20 points behind English-speaking countries. However, on the reputation scale they are substantially further behind. For universities in China, the opposite is the case. They score well on the reputation scale, but are behind on the research performance scale. The QS ranking shows that institutions in German-speaking countries perform quite well on the reputation scale, whereas Dutch-, French-, and Swedish-speaking countries lag behind on this measurement. Chinese institutions have increased their performance substantially on the research performance scale over the years. There is, however, no effect of this increase visible on the reputation scale, on which they already performed well.

In Figure 5, five universities from diverse countries that are all on average ranked in the 50 to 150 range are compared. When comparing the different plots against each other it can be observed that LMU Munich, the University of Southern California, and KTH Stockholm perform similarly in all rankings. An interesting case is a comparison of the Universities of Nanyang and Liverpool. The first performs very well on the reputation performance scale when this is measured using surveys, as in the THE and QS rankings. In the ARWU ranking Nanyang performs poorly on this scale. This difference might be caused by the fact that this institution was established in 1981, hence it has fewer alumni or university staff that have won a Nobel Prize or Fields Medal. The University of Liverpool, in contrast, scores very well on the reputation scale in the ARWU. However, seven out of the nine Nobel Prizes acquired by the University of Liverpool were won before Nanyang University was founded. This shows how the use of Nobel Prizes and Fields Medals by the ARWU ranking to measure reputation can favor older institutions. Also, the behavior of Nanyang University on the research performance scale is noteworthy. In all rankings in 2012 this institution is ranked as one of the lowest on this scale when compared to the other four universities in this plot, but in seven years it climbs to be among the top performers. In the QS ranking, where the reputation score of this university is also very good, Nanyang University climbs from being ranked position 47 to 12 in this period. This shows that, while the results in Section 4 indicate that the rankings are stable over the years, there are specific universities that manage to climb rapidly to the top of the rankings.

The plots show that the ARWU ranking assigns high scores on both the research performance and reputation scales to institutions in English-speaking countries and particularly in the United States and United Kingdom. Asian universities in the ARWU ranking perform worst on both scales. This in contrast with the other two rankings, in which English institutions are also ranked highest but Asian universities are often are among the best performers on the reputation scale. Finally, the figures show that on the research performance scale the rankings have more in common than on the reputation scale—there is more variation visible between the plots when comparing the aggregates or universities on the reputation scale. Furthermore, we see little correlation overall between reputation and performance scales for any of the groups in any of the rankings. Substantial changes in the performance scale (both positive and negative) are generally not correlated with similar movements in reputation, even with some delay. The exception to this may be East Asian and Chinese-speaking universities for which there is some correlation between increasing research performance and reputation, primarily in the THE rankings. However, this may also be due to an unexpected confounder. Increasing publications and visibility, and in particular the global discussion of the importance of the increase in volume and quality of Chinese research performance, might lead to more researchers from those universities being selected to take part in the survey. This is impossible to assess without detailed longitudinal demographic data on the survey participants.

In general, however, these plots show little evidence of strong relationships between reputation and research performance. This could be consistent with circular reinforcement effects on reputation, where proxy indicators for reputation are largely decoupled from research performance. Overall, examining single universities or groups does not provide evidence for or against circular reinforcement effects. As shown earlier in this paper, there is little change in the rankings. Circular effects are therefore hard to observe, because for most universities performance on the rankings is quite stable.

7. DISCUSSION

Accelerated by the increased demand for accountability and transparency, URs have started to play a major role in the assessment of a university's quality. There has been substantial research criticizing these rankings, but only a few studies have performed a longitudinal data driven comparison of URs. This research set out to take an in-depth look at the data of the ARWU, THE, and QS rankings. Based on this analysis, we draw out five key findings.

7.1. Rankings Primarily Measure Reputation and Research Performance

Dehon et al. (2009), Safón (2013), and Soh (2015) showed that by using PCA and EFA on university ranking data it is possible to reveal structures that underlie these rankings. In this research, these techniques are applied to multiple years of the ARWU, THE, and QS rankings. The results of these analyses provide empirical evidence that all three major URs are predominantly composed of two concepts: reputation and research performance. Research performance is measured by the rankings using the number of citations and, in the ARWU, also the number of publications. Reputation is measured in the ARWU by counting the Nobel Prizes and Field Medals won by affiliated university employees and graduates. The THE and QS rankings mainly measure reputation using surveys. The high weights placed on these two concepts by the rankings are problematic. Surveys are a factor that a university has little to no control over and the measurements used to assess research performance are often claimed to be biased (Vernon et al., 2018).

Moed (2017) shows that individual citation and reputation indicators are strongly correlated. Building upon this, we examined the correlation between the reputation and research performance concepts across the rankings. This showed that all concepts are significantly correlated, but correlations within the "same" concept across rankings are not stronger than with the divergent concept. There are multiple explanations possible for this absence of a strong correlation between notionally overlapping concepts. Previous studies have argued that reputation and research performance might influence each other (Robinson-Garcia et al., 2019; Safón, 2019). A university having a high number of citations can positively affect its reputation, and publications written by scholars working at a prestigious university might get cited more often. This is a plausible assertion and the correlations we identify between nonoverlapping concepts are consistent with this argument. However, when we directly visualized the relationships between research performance and reputation scales for a range of universities and groups we did not see evidence for this, as can be seen in Section 6.2. It is nonetheless worthwhile to further explore this effect in future research to gain more insights into the relation between a university's reputation and research performance.

It is also interesting to explore a different explanation: that the underlying concepts in different rankings are not measuring the same thing. Although these rankings are measuring overlapping concepts, the way they measure these concepts might be more influential in the outcome of the ranking order than the actual concepts that the rankings are attempting to measure. This notion is further elaborated on in Section 7.5 of this discussion. Furthermore, the question is what information do these rankings actually provide if similar concepts between them do not correlate? This uncertainty is problematic considering the influence that URs have on society (Billaut et al., 2009; Marginson, 2014; Saisana et al., 2011). This also leads us into the next point: the complications that arise when measuring reputation.

7.2. Reputation Is Difficult to Measure

Measuring reputation in itself is not unimportant, because graduating from or working at a prestigious university can improve a student's or researcher's job prospects (Taylor & Braddock, 2007), even though the relevance of using surveys to rank universities is debated (Vernon et al., 2018). The rankings should therefore look critically at the methodology used to measure this concept. The THE and QS rankings both use two different surveys to measure reputation. The results from the PCA and EFA showed that in both rankings these surveys are highly related. This suggests that these surveys do not in practice provide information on the (separate) quality of education and research, but actually measure a university's general reputation. This then raises the question of what people base their judgment on regarding the reputation of a university. It is not unlikely that the rankings themselves play an important role in this, reinforcing the idea that rankings become a self-fulfilling prophecy (Espeland & Sauder, 2007; Marginson, 2007). The use of Nobel Prizes and Fields Medals as a substitute might appear more objective. However, we have shown that this leads to favoring older universities because it includes alumni who graduated since 1911 and prizewinners since 1921. This is seen in the example of Nanyang University. Furthermore, these prizes are mostly science oriented. The Nobel Prizes in physics, chemistry, and medicine are focused on the natural and medical sciences. The Nobel Prizes in economics, peace, and literature are not specifically science oriented, but the latter two are only counted in the staff variable (i.e., they are less influential) and are arguably quite decoupled from the quality of the university that the winners attended to begin with. This measure, along with many others, therefore favors science-oriented universities.

Another concern with reputation measurement in the ARWU and THE rankings is that these rankings use the variables measuring this reputation concept as proxies for a university's

education and research quality (Academic Ranking of World Universities, 2018; THE World University Ranking, 2018). The variables loading together and forming a reliable scale reflect the notion that it is doubtful whether these variables are a good representation of these unique qualities. Especially in the THE ranking case, it seems that the reputation surveys have such a big influence on the variables that these are mainly a reputation measurement. For the QS ranking, while the problem is the same, there is at least the merit that it is explicitly noted in the methodology that the ranking is measuring reputation directly (QS World University Ranking, 2018).

7.3. Universities in the United States and United Kingdom Dominate the Rankings

Section 6.2 shows that universities in English-speaking countries are ahead of universities in other regions. This seems to support the critique that the ranking methodologies benefit Western, especially English-speaking, universities (Pusser & Marginson, 2013; Van Raan, 2005; Vernon et al., 2018). For all rankings, we see a substantial advantage for English-speaking universities on the research performance scale, even though more and more universities in non-English-speaking countries publish predominantly in English (Altbach, 2013; Curry & Lillis, 2004). However, despite the fact that both the THE and QS rankings employ methodologies to account for the fact that non-English articles receive fewer citations, universities in English-speaking countries still lead the rankings.

There is also a strong regional effect between the rankings on the reputation component. Eastern Asian, especially Chinese, universities score highly on the reputation measurement in the THE and QS ranking. Non-English-speaking European universities and institutions from Australia and New Zealand perform substantially worse on this scale, even when the research performance component is the same or higher. Reputation measures for Australian and New Zealand universities appear particularly volatile in the QS ranking. This may indicate that the THE and QS rankings reputation measurements favor Asian universities. This could be due to increasing profile and marketing, more effective gaming of the survey by top East Asian and Chinese universities, or some other difference in the methodology. More research is thus needed to draw definitive conclusions on this matter.

7.4. Rankings Are Stable over Time but Differ from Each Other

Our analysis shows that for all three rankings consecutive years of the same ranking are strongly correlated and are very similar according to the M-measure. This is in accordance with results of Aguillo et al. (2010). This means that it is hard to change position within a ranking. This year-to-year similarity can be explained by different choices made by the ranking designers. All three rankings use rolling averages to calculate research performance indicators. Also, 30% of the ARWU ranking is constructed by variables that measure prizes won since 1921 and which are therefore very stable. For the THE and QS stability can be explained by the high weighting assigned to reputation surveys. A university's reputation is not likely to change substantially within one, or even a small number, of years. Generally speaking, all URs employ conservative methodologies, which results in the rankings being very stable. For another perspective on stability, we refer readers to Gingras (2016). Circular effects between ranking years, as described by Safón (2019) and Robinson-Garcia et al. (2019), could also results in the rankings being stable. The plots created in this research did not indicate the existence of such effects, but the correlation between reputation and research performance can be taken as evidence for the claim made by Safón (2019) that research performance is also influenced by prior rankings.

The rankings were also compared to each other. These analyses showed that in the top 50 and top 100 the different rankings are correlated strongly; however, the M-measure indicated only medium similarity, showing substantial variation between the rankings. These results are

in accordance with the findings of Aguillo et al. (2010) and the overlap measurements of Moed (2017). Given this stability, one would expect that the rankings would be more similar; thus, it is surprising that they are not. It is even more noteworthy that there is dramatically more similarity in the top 50 than in positions 51–100. This is most likely caused by the fact that performance differences are only minor between lower ranked universities. Designer choices are, as will be shown next, influential for ranking order and become more influential when the differences in performance are becoming smaller.

7.5. Ranking Designers Influence Ranking Order

The relative absence of similarity between the three rankings is noteworthy, since in this article it was established that they to a large extent measure similar concepts. Several reasons can be used to explain the differences between the rankings. First, the rankings assign different weights to the variables that compose these concepts, which, as been shown in multiple studies, has a large effect on a university's ranking position (Dehon et al., 2009; Marginson, 2014; Saisana et al., 2011). It should also be noted that there are also nonoverlapping measurements that can explain differences between rankings. For example, the QS assigning a substantial weight to the student-staff ratio and the decision of the ARWU not to include internationality are the most important. Second, rankings use different methods to normalize their data. The THE and QS correct their research performance measurement for university size, while in the ARWU raw numbers are used. Choices made by the ranking designers for specific weighting and normalization schemes are thus important determinants of the final ranking order (Moed, 2017). Perhaps most importantly, our paper, in agreement with previous work, shows that the majority of the ranking variables are attempts to quantify two specific concepts of university performance. The differences between the rankings is therefore not what they are trying to measure but how they seek to measure it.

Two limitations of this research should be addressed. First, we concluded that the reputation variables in the THE and QS ranking loading together is caused by the fact that these are both measuring a general reputation concept. However, it is possible that these variables do actually measure distinct reputation properties, but that teaching quality and research quality are extraordinarily highly correlated. While there is a likely connection between teaching and research quality, we are skeptical that (a) this correlation would be so high and (b) that survey respondents are in a position to distinguish between details of education and research provision, especially in a context where they are being asked about both. Attempts to distinguish between teaching quality and research quality, such as in the UK's Teaching and Research Excellence Framework, show low correlation between highly evaluated institutions. It is thus reasonable to expect that their judgment is, at least partially, caused by more general reputation attributes, for example the number of Nobel Prizes and Fields Medals won (Altbach, 2012). More research is needed to identify what influences survey respondents' judgment of a university's reputation and how the selection of respondents and questions might influence that. This could be studied by reviewing the questions used to measure the reputation variables and analysis of the raw data collected from these questionnaires. It may also be interesting to see how external data sources relate to these measurements, for example, by measuring the impact of a university appearing in popular or social media (Priem, Taraborelli, Groth, & Neylon, 2010). Our results might be seen as supportive of the INORMS statement that surveys should not form the basis of rankings (INORMS Research Evaluation Group, 2019). In any case, greater transparency on the sample selection and questions posed (as well as how they may have changed) would be of value in probing this issue.

Second, some qualifications should be made when interpreting the extracted loading structures in the PCA and EFA. In the QS ranking in some years a number of universities had to be

removed from the analysis because of missing data elements. However, since the loadings in the PCA and EFA for the QS were similar across the years, we are quite confident that a genuine structure was extracted from this ranking. Nonetheless, the large number of missing values in the QS makes it unclear how the overall ranking score for a wide range of universities was constructed and makes it hard to study and verify the QS data. We would urge the QS ranking to provide more transparency in this area. Furthermore, there is research that suggests that all ARWU variables measure one concept (Safón, 2013). The results of our PCA also showed most variables loading on one component in some years and the second component's eigenvalue did not exceed one. Of the three rankings, the ARWU therefore appears to be measuring the most singular concept. This is most likely caused by the fact that there are a substantial number of universities that score very low on the Alumni and Award variables, which in turn is a logical result of how these variables are measured (see Section 2). For the institutions that score low on these two variables the ARWU thus only measures academic performance. But, when reviewing the EFA results and previous work by Dehon et al. (2009), we think it is reasonable that these Alumni and Award variables are actually measuring a distinct factor.

This paper provided a longitudinal comparison between the three major URs. It showed that rankings are stable over time but differ significantly from each other. Furthermore, it revealed that the rankings all primarily measure two concepts—reputation and research performance—but it is likely that these concepts are influencing each other. Last, it discussed these findings in light of the critiques that have been raised on URs. This provides insights into what URs do and do not measure. Our results also show that there is uncertainty surrounding what the rankings' variables exactly quantify. One thing is certain, however: It is impossible to measure all aspects of the complex concept that is university performance (Van Parijs, 2009). Despite this, universities are focusing on and restricting their activities to ranking criteria (Marginson, 2014). But because it is unclear what the rankings quantify it is also unclear what exactly the universities are conforming to. Universities aim to perform well on ambiguous and inconsistent ranking criteria, which at the same time can hinder their performance on activities that are not measured by the rankings. We conclude that universities should be extremely cautious in the use of rankings and rankings data for internal assessment of performance and should not rely on rankings as a measure to drive strategy. A ranking is simply a representation of the ranking data. It does not cover all aspects of a university's performance, but it may also be a poor measure of the aspects it is intended to cover.

AUTHOR CONTRIBUTIONS

Friso Selten: Conceptualization, Methodology, Investigation, Software, Formal analysis, Data curation, Visualization, Writing—original draft. Cameron Neylon: Conceptualization, Methodology, Writing—review & editing, Resources. Chun-Kai Huang: Writing—review & editing. Paul Groth: Conceptualization, Methodology, Writing—review & editing, Supervision.

COMPETING INTERESTS

PG is coscientific director of the ICAI Elsevier AI Lab—an artificial intelligence lab cofinanced by Elsevier. Elsevier is a URs data provider.

DATA AVAILABILITY

The data used in this paper were obtained from the ARWU, THE, and QS websites. All scripts used to collect data, perform the analyses, and create visualizations are available at Selten (2020). Due to a lack of clarity on the reuse rights for the data provided on rankings websites presented, it is not clear that we are permitted to publicly redistribute the data that have been scraped from the ranking websites. For all rankings, this data is publicly available at the ranking websites.

REFERENCES

Academic Ranking of World Universities. (2018). *Arwu2018 methodology*. Retrieved April 18, 2019 from http://shanghairanking.com/ARWU-Methodology-2018.html

Aguillo, I., Bar-Ilan, J., Levene, M., & Ortega, J. (2010). Comparing university rankings. *Scientometrics*, *85*(1), 243–256.

Altbach, P. G. (2012). The globalization of college and university rankings. *Change: The Magazine of Higher Learning*, *44*(1), 26–31.

Altbach, P. G. (2013). The imperial tongue: English as the dominating academic language. In *The international imperative in higher education* (pp. 1–6). Leiden: Brill Sense.

Altbach, P. G., & Knight, J. (2007). The internationalization of higher education: Motivations and realities. *Journal of Studies in International Education*, *11*(3–4), 290–305.

Bar-Ilan, J., Levene, M., & Lin, A. (2007). Some measures for comparing citation databases. *Journal of Informetrics*, *1*(1), 26–34.

Billaut, J.-C., Bouyssou, D., & Vincke, P. (2009). Should you believe in the Shanghai ranking? An MCDM view. *Scientometrics*, *84*(1), 237–263.

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin*, *103*(2), 276.

Curry, M. J., & Lillis, T. (2004). Multilingual scholars and the imperative to publish in english: Negotiating interests, demands, and rewards. *TESOL Quarterly*, *38*(4), 663–688.

Dehon, C., McCathie, A., & Verardi, V. (2009). Uncovering excellence in academic rankings: A closer look at the Shanghai ranking. *Scientometrics*, *83*(2), 515–524.

Digital-science. (2019). *Grid release 2019-05-06*. Retrieved from https://digitalscience.figshare.com/articles/GRID_release_2019-05-06/8137970/1

Docampo, D. (2011). On using the Shanghai ranking to assess the research performance of university systems. *Scientometrics*, *86*(1), 77–92.

Eisinga, R., Te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, *58*(4), 637–642.

Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, *113*(1), 1–40.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Gingras, Y. (2016). *Bibliometrics and research evaluation: Uses and abuses*. Cambridge, MA: MIT Press.

Gravetter, F. J., & Wallnau, L. B. (2016). *Statistics for the behavioral sciences*. Independence, KY: Cengage Learning.

Habing, B. (2003). *Exploratory factor analysis*. University of South Carolina, October, 15.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis*. Essex: Pearson Education Limited.

Hazelkorn, E. (2007). The impact of league tables and ranking systems on higher education decision making. *Higher Education Management and Policy*, *19*(2), 1–24.

Huang, M.-H. (2012). Opening the black box of QS World University Rankings. *Research Evaluation*, *21*(1), 71–78.

Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, *47*(1), 64–79.

INORMS Research Evaluation Group. (2019). *What makes a fair and responsible university ranking? Draft criteria for comment* (Technical Report).

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36.

Marginson, S. (2007). Global university rankings: Implications in general and for Australia. *Journal of Higher Education Policy and Management*, *29*(2), 131–142.

Marginson, S. (2014). University rankings and social science. *European Journal of Education*, *49*(1), 45–59.

Moed, H. F. (2017). A critical comparative analysis of five world university rankings. *Scientometrics*, *110*(2), 967–990.

Osborne, J. W., Costello, A. B., & Kellow, J. T. (2008). Best practices in exploratory factor analysis. *Best Practices in Quantitative Methods*, 86–99.

Pett, M., Lackey, N., & Sullivan, J. (2003). *Making sense of factor analysis*. Thousand Oaks, CA: Sage.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *Altmetrics: A manifesto*. Retrieved from http://altmetrics.org/manifesto/

Pusser, B., & Marginson, S. (2013). University rankings in critical perspective. *Journal of Higher Education*, *84*(4), 544–568.

QS World University Ranking. (2018). *World university ranking methodology*. Retrieved April 18, 2019 from https://www.topuniversities.com/qs-world-university-rankings/methodology

Robinson-Garcia, N., Torres-Salinas, D., Herrera-Viedma, E., & Docampo, D. (2019). Mining university rankings: Publication output and citation impact as their basis. *Research Evaluation*, *28*(3), 232–240.

Romzek, B. S. (2000). Dynamics of public sector accountability in an era of reform. *International Review of Administrative Sciences*, *66*(1), 21–44.

Safón, V. (2013). What do global university rankings really measure? The search for the X factor and the X entity. *Scientometrics*, *97*(2), 223–244.

Safón, V. (2019). Inter-ranking reputational effects: an analysis of the Academic Ranking of World Universities (ARWU) and the Times Higher Education world university rankings (THE) reputational relationship. *Scientometrics*, *121*(2), 897–915.

Saisana, M., d'Hombres, B., & Saltelli, A. (2011). Rickety numbers: Volatility of university rankings and policy implications. *Research Policy*, *40*(1), 165–177.

Scott, P. (2013). Ranking higher education institutions: A critical perspective. In *Rankings and Accountability in Higher Education: Uses and Misuses*, (p. 113). Van Haren Publishing.

Selten, F. (2020). *A Longitudinal Analysis of University Rankings*. Zenodo. Retrieved from https://doi.org/10.5281/zenodo.3775251

Soh, K. (2015). What the Overall doesn't tell about world university rankings: Examples from ARWU, QSWUR, and THEWUR in 2013. *Journal of Higher Education Policy and Management*, *37*(3), 295–307.

Stergiou, K. I., & Lessenich, S. (2014). On impact factors and university rankings: From birth to boycott. *Ethics in Science and Environmental Politics*, *13*(2), 101–111.

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.

THE World University Ranking. (2018). *World university rankings 2019: Methodology*. Retrieved April 18, 2019 from https://timeshighereducation.com/world-university-rankings/methodology-world-university-rankings-2019

Todorov, V. (2012). Robust location and scatter estimation and robust multivariate analysis with high breakdown point. http://www.cran.r-project.org/web/packages/rrcov

Usher, A., & Savino, M. (2006). A world of difference: a global survey of university league tables. *Canadian Education Report Series*. Online Submission.

Van Parijs, P. (2009). European higher education under the spell of university rankings. *Ethical Perspectives*, *16*(2), 189–206.

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

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some further observations. *Multivariate Behavioral Research*, *25*(1), 97–114.

Vernon, M. M., Balas, E. A., & Momani, S. (2018). Are university rankings useful to improve research? A systematic review. *PLOS ONE*, *13*(3), e0193762.

Wikidata Contributors. (2019). *Wikidata*. Retrieved March 20, 2019 from https://www.wikidata.org