Measuring University Quality

Christopher Claassen*

January 8, 2015

*Department of Government, University of Essex, cclaas@essex.ac.uk. Many thanks to Isidro F. Aguillo and Robert Morse for kindly supplying the Webometrics and *US News* National Universities ratings data respectively. Lutz Bornmann provided helpful comments on a earlier version of this paper. Replication code and an online appendix with supplementary tables are available at http://www.chrisclaassen.com.

Measuring University Quality

This paper uses a Bayesian hierarchical latent trait model, and data from eight different university ranking systems, to measure university quality. There are five contributions. First, I find that ratings tap a unidimensional, underlying trait of university quality. Second, by combining information from different systems, I obtain more accurate ratings than are currently available from any single source. And rather than dropping institutions that receive only a few ratings, the model simply uses whatever information is available. Third, while most ratings focus on point estimates and their attendant ranks, I focus on the uncertainty in quality estimates, showing that the difference between universities ranked 50th and 100th, and 100th and 250th, is insignificant. Finally, by measuring the accuracy of each ranking system, as well as the degree of bias toward universities in particular countries, I am able to rank the rankings.

Keywords: Latent trait models, Bayesian models, University rankings

RANKINGS have become central to the workings of universities across the world. They are now used as explicit targets by university administrators (Hazelkorn, 2007), in funding decisions by government authorities (Rauhvargers, 2013; Salmi and Saroyan, 2007), and as important factors in students' application and enrolment decisions (Bowman and Bastedo, 2009; Monks and Ehrenberg, 1999).

Despite their already-weighty and steadily-growing importance, a number of criticisms have been levelled at these ranking systems.[1] Some scholars contend that university quality is not an unidimensional construct (van Vught and Ziegele, 2012). Others regard the methods of these ranking systems with skepticism, noting that indicators are included without any real justification or may be prone to manipulation, while indicator weights are selected in an entirely arbitrary fashion (Lee, 2009; Waltman et al., 2012). In addition, while all rankings rely on overall ratings of university quality, these quantities are downplayed in favour of the less informative rank-order data. Perhaps even more troubling, the uncertainty of these estimates is ignored, despite a suspicion that little separates a university ranked one hundredth in the world from one ranked two hundredth (Goldstein and Spiegelhalter, 1996). Being primarily an Anglo-American product, rankings are also seen by some as biased toward universities in English-speaking countries (Leeuwen et al., 2001; Waltman et al., 2012). Finally, the public users of these rankings, who might be less troubled by these technical issues, are now confronted by a bewildering array of rankings: at least half a dozen global rankings as well as numerous national and disciplinary rankings. How are these users to go about extracting useful information from this mountain of often contradictory data?

The purpose of this paper is to address these concerns. My goal is to provide transparent information for university administrators, academics, students and higher education authorities by systematically building a model of university quality that in-

[1] A note on terminology. Different organisations, research groups, or individuals produce different university rankings: I refer to these as different *systems*. Each system uses specific *indicators* of research quality, such as citation counts, to produce an overall scale or *rating* of university quality. These ratings are then use to rank universities from best to worst, and it is this information on university *rankings* that is the most widely released and consumed metric of these systems.

corporates as much information from as many rankings as possible. To accomplish this goal, I model university ratings as observations of an unobserved quality variable using a bespoke Bayesian hierarchical latent trait model and data from eight different university ranking systems—two national and six international. This model is designed to be flexible, and so can be easily extended to incorporate additional rankings or predictors at the levels of discipline, university, rating system or country.

This model, and the accompanying estimates of university quality, offer five specific benefits. First, I test whether university quality is, in fact, a single dimension of variation as all ranking systems suppose. Second, more accurate measures of university quality can be obtained by aggregating the estimates of individual ranking systems, all of which include some unique source of data. Third, the model produces uncertainty estimates, which helps to show the meaning, or lack thereof, in any given shift in one, ten or one hundred ranking places. Finally, I rank the rankings by measuring the accuracy of each ranking system and the degree of bias toward the universities of particular countries.

Existing Research on University Ranking Systems

A large number of writers have offered criticisms of university rankings, their constituent indicators and weighting schemes, and their effects on students decisions and the universities themselves (Altbach, 2010; Bowman and Bastedo, 2011; Enserink, 2007; International Ranking Expert Group, 2011; Hallinger, 2014; Hazelkorn, 2007; Leeuwen et al., 2001; Monks and Ehrenberg, 1999; Rauhvargers, 2013; Salmi and Saroyan, 2007; Waltman et al., 2012). Surprisingly, given the quantitative nature of university rankings, qualitative analysis and commentary predominates.

Turning to the relatively few quantitative studies of university rankings, Bowman and Bastedo (2009) and Monks and Ehrenberg (1999) examine the effects of changes in the *US News & World Report's* National Universities rankings, finding that universities that climb the ladder receive more applications and enroll a more highly qualified pool of students. Grewala et al. (2008) focus instead on the factors that result in changes in *US News & World Report* rankings: while persistence is the norm, graduation and retention

rates emerge as the most important causes. Bowman and Bastedo (2011) examine the *Times Higher Education* World University Rankings, finding that the initial published rankings had an anchoring effect on the survey measure of university reputation used in subsequent editions, while Soh (2014) shows that the indicators used in these rankings are highly collinear such that all but two could be dropped without much loss in information. Soh (2011) examines indicators from the THE rankings, Shanghai Academic Rankings of World Universities and the Quacquarelli Symonds World University Rankings, finding that indicators such as industry income and internationalisation show weak associations with composite rating scores. Finally, Bornmann et al. (2013) model paper citation counts as a function of university- and country-level effects and covariates, with the finding that national systems account for a far greater proportion of the variance in highly cited academic papers than do universities.

While few quantitative studies of rankings exist, even fewer compare ranking systems in a quantitative fashion. The two exceptions are Usher and Savino (2006), who compare the most highly ranked universities in several countries across 19 global and national rating systems, noting the high degree of correspondence, and Aguillo et al. (2010), who estimate similarity measures for four major global rankings systems. No researchers have yet turned their attention to the topic at hand and attempted to combine information from different rankings to estimate university quality.

Data and Methods

Data

Data from six global and two national ranking systems were collected in October and November 2014 (see Table 1).[2] The criteria for including a global ranking were that

[2]Quacquarelli Symonds provide downloadable ratings data on their website. Isidro Aguillo kindly shared with me the Webometrics ratings data for the top 1000 universities, while Robert Morse graciously supplied me with the latest National University Ratings data from *US News & World Report*. For all other rating systems, data were obtained by scraping public websites.

a rating was provided as well as a ranking, and that this rating was calculated from multiple indicators.[3] The global rankings included in this analysis are: (1) the "World University Rankings" from *Times Higher Education*, (2) the "Academic Ranking of World Universities" from Shanghai Jiao Tong University, (3) an unnamed ranking compiled by the Center for World University Rankings in Jeddah, Saudi Arabia, (4) the "QS World University Rankings" provided by Quacquarelli Symonds, (5) the "Best Global Universities" from *US News & World Report*, and (6) the "Webometrics Ranking of World Universities" from the Cybermetrics Lab of the Spanish National Research Council.

Two national ranking systems are also included: one for the USA and the other for the UK. The *US News & World Report's* "National University Rankings"—the original ranking system—is used to include its additional information on research universities in the USA in the university quality model. Although several British rankings are available, all combine essentially the same set of indicators. The most thorough of these UK rankings, the Complete University Guide's "University League Table", 2015 edition, is used in this analysis.

Model

Ratings are assumed to be observed measures of the unobserved but underlying variable of university quality. A latent trait measurement model is used to obtain estimates of this latent variable. While the factor analysis model (FAM) is appropriate for such interval-level observed data, its use of an $N \times J$ matrix of observed indicators for N individuals (universities in this case) and J variables (rating systems) renders it less than optimal for the data at hand. The FAM would require listwise deletion of a university with data missing for even one rating system, which precludes the possibility of utilising data from different national ratings systems.[4]

[3] A notable new system, the Leiden ranking, is thus excluded, because it offers various rankings, each based on single citation count indicator.

[4] An alternative to listwise deletion would be to compute a correlation matrix using pairwise deletions (as I do later in this paper). However, this technique may result in a non-positive definite correlation

Instead, the matrix of observed ratings is stacked as a vector, in "long" format. This vector, y_i, for $i = 1, ..., I$ ratings is then modelled as a linear function of the university latent quality parameters (θ_j for $j = 1, ..., N$ universities) with the link between θ_j and y_i adjusted by intercepts ν_p and slopes λ_p for $p = 1, ..., P$ rating systems. The model is

$$y_i = \nu_{p[i]} + \lambda_{p[i]}\theta_{j[i]} + \epsilon_i.$$

The error term, ϵ_i, is distributed $\mathcal{N}(0, \psi_{p[i]})$, taking one of P different variance estimates depending on which rating system is associated with any particular rating. This model is thus closely related to a FAM but allows for data where $I < P \times N$, which is the case here, because not all universities are rated by every rating system. As is typical with factor analyses, the latent quality estimates, θ_j, are given a $\mathcal{N}(0, 1)$ distribution to identify the location and scale of the parameters.

This model is also closely related to a hierarchical linear regression model (HLM) with varying intercepts and slopes by rating system, although here also with a vector of unobserved latent university quality scores, θ, as a predictor of observed ratings data (as well as a vector of P error variances). To account for the correlation that may be present between hierarchical intercepts and slopes, the ν and λ parameters are drawn from a bivariate normal distribution with variances and covariances, including a correlation parameter, ρ, to be estimated

$$\begin{pmatrix} \nu_p \\ \lambda_p \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\nu^2 & \rho\sigma_\nu^2\sigma_\lambda^2 \\ \rho\sigma_\nu^2\sigma_\lambda^2 & \sigma_\lambda^2 \end{pmatrix} \right).$$

The use of a vector of stacked observed measures also renders this model analogous to item-response theory (IRT) latent trait models, which are typically used to model dichotomous or polytomous indicators.[5] In IRT terminology, the ν intercepts are difficulty

matrix, which is not amenable to factor analysis. Nor does it solve the problem of two variables / ratings without any overlapping observations whatsoever.

[5]Glockner-Rist and Hoijtink (2009) describes the parallels between IRT and FA models.

parameters and the λ factor loadings are discrimination parameters.

Another benefit of this hierarchical parameterization of the FAM is its extensibility. In addition to the levels of rankings (I), universities (J), and rating systems (P), we could model the effects of national systems (Bornmann et al., 2013), include measures of university quality by discipline alongside the full institutional ratings, or introduce covariates (Bafumi et al., 2005).

Estimation

A fully Bayesian, Markov Chain Monte Carlo, approach to estimation is used. MCMC estimation allows for more accurate modelling of the hierarchical variance and covariance parameters. It also readily permits the uncertainty of all parameters to be estimated, which is particularly useful for the university quality measures, θ.

The eight sets of ratings data were preprocessed by standardising to mean zero and unit variance, and, if necessary, log transformation to remove skewness. This permitted the modelling of the vector of observed ratings as normally distributed, albeit with rater-specific variances, which is more efficient and convenient than specifying a log-normal or related distribution. The latent quality parameters were also modelled using a normal distribution.

Non-informative but proper uniform (0, 100) priors are used for the error variances, ψ. Given that the number of rating systems included is not large ($P = 8$), I use more informative half-Cauchy $(0, 2.5)$ prior distributions for the intercept and slope variance parameters, σ^2 (Gelman, 2006). The correlation parameter, ρ, is then modeled using an LKJ correlation distribution (Lewandowski et al., 2009; Stan Development Team, 2014).

Restrictions on the direction, location and scale of the parameters are required to identify such latent trait models (Bafumi et al., 2005). Location and scale are identified with a $\mathcal{N}(0, 1)$ prior on the θ parameters. The direction, or sign, of the varying slopes, λ is identified by using positive initial values for each chain. This is more convenient than using parameter constraints given the matrix manipulations needed to model the variance covariance matrix of hierarchical parameters.

Estimation is conducted using the `Stan` programme, which implements Hamiltonian Monte Carlo sampling (Stan Development Team, 2014). Four parallel chains were run for 2000 iterations each, with the first half of the samples in each chain treated as warm ups and discarded, and the remaining 4000 samples saved and analysed further. This number of iterations proved to be more than sufficient for convergence, as evidenced by the Gelman-Rubin diagnostic (Gelman and Rubin, 1992) having a value close to one for all parameters.

Findings

Dimensionality of University Ratings

Critics have questioned whether university quality is a unidimensional construct, and even whether existing ratings are themselves unidimensional. Latent trait models also have the assumption of local independence, which requires that the dimensionality of the underlying trait be specified correctly. I thus conduct a simple test of the dimensionality of the sets of ratings, by computing the eigenvalues of the correlation matrix of the ratings data. The two national ratings were dropped as they do not share any observations in common. Plotting the six eigenvalues from the remaining global ratings correlation matrix gives Cattell's familiar scree plot (Figure 1), which shows the eigenvalues ranked in order of size. These are contrasted with a scree plot for a set of six random variables, indicated here using a dashed line. This method, know as parallel analysis, calculates the approximate eigenvalues that would be obtained through chance alone, rather than through any structure in the data. As we can see, the observed data is strongly unidimensional: the first eigenvalue is substantially larger than the others and substantially larger than the randomly generated eigenvalue.

It is also worth examining the correlations themselves. Table 2 shows the full 8×8 correlation matrix. The pairwise correlations are all positive, and generally strongly so, which reflects the unidimensionality of the underlying trait. The pairwise sample sizes, however, vary considerably, which shows the difficulty in modelling this latent trait using

factor analysis or other covariance structure models. In particular, there is no overlap between the UK and US national rankings.

Estimates of University Quality

Having established unidimensionality, a single-dimensional latent trait model can be fit. The estimates of greatest interest are the university quality parameters, θ, which are displayed graphically in Figures 2 and 3. Estimates of the other parameters are provided in Table 3 in the Appendix.

The means of the marginal posterior densities of the latent trait—our point estimates of university quality—are indicated with solid points. The variation around these point estimates is indicated with horizontal bars, which show the central 95% of the marginal posterior quality density for each university. The degree of uncertainty is strongly related to the number of ratings that were available in the data: universities with six or seven ratings have more precise estimates, while those with only one or two, such as Science Po Paris and Pepperdine, are less precise. One of the advantages of the model is it uses whatever information is at hand, rather than eliminating or institutions with sparse data or imputing missing values.

Being a linear combination of eight existing rankings, there are few surprises at the top of the table. Prominent American universities dominate, as they do in all major ranking systems. Harvard is at the top, slightly more than three standard deviations away from the mean university.[6] The advantage that these point estimates have over those of any particular ranking system, is that they aggregate information in the individual rankings, making them more accurate.

[6]Two points are worth noting regarding the distribution of these quality ratings. First, a normal distribution is assumed. There is some evidence (see, for example, the distributions of the raw ratings in the Appendix) that university ratings follow a skewed distribution, such as the lognormal. The latent variable of university quality might also be assumed to distributed non-normally. Second, this sample contains 1373 of the highest quality universities in the world, drawn from a population of over 12000 universities. Regardless of the true distribution of this population, the distribution of the top 10% is likely to differ. Further research on this topic would be of interest.

Perhaps of even more interest, however, are the estimates of uncertainty in university quality. A rough idea of whether one university can meaningfully be said to be more highly rated than another is to verify if the error bars overlap.[7] Using this method, one can see that it becomes increasingly more difficult to differentiate universities as we move down the rankings. Harvard can be distinguished from institutions ranked outside the top 15 and the top 10 can be said to be of significantly higher quality than those ranked in the 30s and below. But these institutions in the 30s cannot be differentiated from others that are still in the top 100, but lower in ranking. And, moreover, these universities ranked 100-110 can only be said to be of significantly higher quality than universities ranked far inferior, in the high 200s and below.

Another way to visualise this uncertainty is to rank the universities over each of the the 4000 MCMC simulations and plot the proportion of the simulations in which each university makes it into the top 100, a benchmark frequently used by the ranking systems. These estimates are displayed in Figure 4, and can be interpreted as the probability, given the uncertainty in the data, that each university is a top-100 university.

The plot shows that the leading 50 universities are virtually certain to be top-100, after which the probability decreases. In a neat correspondence, the 100th ranked university (FU Berlin) is top-100 on the balance of probabilities because its probability estimate is almost exactly 50%. Those ranked lower would then be outside the top 100 on this balance-of-probabilities criterion. Another way to interpret this plot is as a direct visualisation of the p-value of being in the top 100. P-values are typically used in conjunction with some critical value, usually $p = 0.05$, to test hypotheses. The plot shows that the "null hypothesis" that university j is in the top 100 cannot be reliably rejected until we reach the 150s (the $p = 5\%$ "rejection region" is indicated on the graph in darker grey).

[7]Strictly speaking, this is not an accurate method. A better test of whether two universities' quality estimates are significantly different would be to compute the 95% confidence interval of the difference and verify that it does not include 0.

Bias and Accuracy of the Ranking Systems

The model also allows us to estimate the accuracy of the various rating systems as well as their degree of bias toward universities from particular countries. I turn first to the issue of accuracy, two measures of which are plotted in Figure 5. The left panel displays the square root of the ψ parameters, which gives us the rater-specific residual standard errors. The right panel shows the adjusted R^2 of the effect of the latent variable of university quality on each set of ratings. This latter quantity is derived from former but has a readier interpretation.

The USN-GU, Jeddah, and Shanghai rating systems are the most accurate, with R^2 statistics in excess of 0.80. The THE and QS rankings are somewhat noisier, probably because they devote part of their score to indicators like internationalisation that have been shown to have little to no correlation with university quality (Soh, 2011). The two national ranking systems, the USN-NU and CUG, perform worse than the major global rankings. This comes as a surprise because country-specific ranking systems should avoid the errors that accrue from comparing universities across different national systems.[8] The Webometrics rankings, finally, is one of the least accurate. This lack of precision, however, is offset by two attractive features of this ranking system. First, is the vast scope of the Webometrics project: with 12000 universities ranked, it is an order of magnitude larger than the next largest. Second, it is the only ranking system to use entirely unique indicators—relating to web presence and openness of information—which means it avoids rehashing the same citations, reputation and spending indicators used by all the other systems.

In addition to the overall error rate it is also of interest to examine how the error of the rankings varies across national systems. This will allow me to test whether any individual ranking favours the universities in certain countries and discriminates against those in other countries. In particular, we might imagine that rankings are likely to

[8]In other research, currently in progress, I examine UK indicators of university quality and find that the CUG ranking also includes indicators, such as spending on facilities, that are weakly related to the underlying construct.

favour home universities, those from the country in which the ranking system is based.

To test the proposition that rankings are biased toward home country universities, I examine the model residuals within system and country. The mean residual for the home country universities of each global rating system are plotted in Figure 6. Our MCMC estimation strategy produces 4000 samples of each residual, so it is straightforward to also estimate the 95% credible intervals of these estimates, which are displayed on the figure as horizontal bars. Three of the ranking systems indeed show positive bias toward the universities in their home country. UK universities are rated more highly, on average, by the QS and THE ranking systems than they are by the latent trait model,[9] and the Webometrics ranking also shows a positive bias toward Spanish universities. These three systems favour their home country universities, on average, by at least .25 points on the latent variable scale: enough for a university outside the top 200 to increase its ranking by 20 to 50 places.

Neither the USN-GU nor the Jeddah rankings show an discernible favouritism toward their home country universities. The Jeddah ratings of Saudi Arabian universities are even slightly more negative than the overall estimates from the latent trait model although the small number of home country universities produces a high degree of uncertainty. The USN-GU ratings of home country universities appear to be scrupulously neutral, despite the large number of US universities in their sample and the resulting precision of the bias estimate.

Although it is preferable to rely on composite university ratings, such as those I provide in this paper, the findings outlined in this section allow me to rank the rankers. Overall, on the two dimensions of accuracy and bias, a clear winner emerges. The "Best Global Universities" rankings from the *US News & World Report* shows both the least amount of error overall as well as no evidence of home-country bias. The Jeddah Centre for World University rankings are similarly unbiased and almost as accurate. They are also one of the most ambitious ranking systems, with ratings for the top 1000 universities

[9]Both of these systems also show bias toward universities in Australia and the Netherlands. See Table 4 in the appendix for further results.

calculated and made publicly available.

Conclusions

This paper describes an attempt to improve existing estimates of university quality by building a Bayesian hierarchical latent trait model and inputting data from eight rankings. There are five main findings. First, despite their different sources of information, ranging from objective indicators, such as citation counts, to subjective reputation surveys, existing rating systems clearly tap a unidimensional latent variable of university quality. Second, the model combines information from multiple rankings, producing estimates of quality that offer more accurate ratings than can be obtained from any single ranking system. Universities that are not rated by one or more rating systems present no problem for the model: they simply receive more uncertain estimates of quality. Third, I find considerable error in measurement: the ratings of universities ranked around 100th position are difficult to distinguish from those ranked close to 30th; similarly for those ranked at 100th and those at 250th. Fourth, each rating system performs at least adequately in measuring university quality. Surprisingly, the national ranking systems are the least accurate, which may be due to their usage of numerous indicators, some extraneous. Finally, three of the six international ranking systems show bias toward the universities in their home country. The two unbiased global rankings, from the Center for World University Rankings in Jeddah, and *US News & World Report* are also the two most accurate.

In future research, this hierarchical latent variable model of university quality could be extended in several ways. Following the recommendations of Bornmann et al. (2013) one could estimate the effects of national systems, including how these interact with the rating systems to produce differential weights for the latent quality estimates. Additional data from national ratings systems could be included. Data from disciplinary ratings might also be introduced, nested within the overall university quality estimates. Finally, these quality estimates could themselves be modeled using covariates. Including a regression equation in the latent variable model would allow one to retain the uncertainty

of measurement rather than saving and using point estimates.

References

Aguillo, I. F., Bar-Ilan, J., Levene, M., and Ortega, J. L. (2010). Comparing university rankings. *Scientometrics*, 85(1):243–56.

Altbach, P. G. (2010). The state of the rankings. *Inside Higher Ed*, (November 11).

Bafumi, J., Gelman, A., Park, D. K., and Kaplan, N. (2005). Practical issues in implementing and understanding bayesian ideal point estimation. *Political Analysis*, 13(2):171–87.

Bornmann, L., Mutz, R., and Daniel, H.-D. (2013). Multilevel-statistical reformulation of citation-based university rankings: The leiden ranking 2011/2012. *Journal of the American Society for Information Science and Technology*, 64(8):1649–58.

Bowman, N. A. and Bastedo, M. N. (2009). Getting on the front page: Organizational reputation, status signals, and the impact of *U.S. News and World Report* on student decisions. *Research in Higher Education*, 50(5):415–36.

Bowman, N. A. and Bastedo, M. N. (2011). Anchoring effects in world university rankings: Exploring biases in reputation scores. *Higher Education*, 61(4):431–44.

Enserink, M. (2007). Who ranks the university rankers? *Science*, 317(5841):1026–28.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–33.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–72.

Glockner-Rist, A. and Hoijtink, H. (2009). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, 10(4):544–65.

Goldstein, H. and Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(3):385–443.

Grewala, R., Deardena, J. A., and Llilіena, G. L. (2008). The university rankings game: Modeling the competition among universities for ranking. *The American Statistician*, 62(3):232–7.

Hallinger, P. (2014). Riding the tiger of world university rankings in east asia: Where are we heading? *International Journal of Educational Management*, 28(2):230–45.

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

International Ranking Expert Group (2011). Ireg ranking audit manual. Technical report, IREG Observatory on Academic Ranking and Excellence. www.ireg-observatory.org.

Lee, S. (2009). Reputation without rigor. *Inside Higher Ed*, (August 19).

Leeuwen, T. N. V., Moed, H. F., Tijssen, R. J. W., Visser, M. S., and Raan, A. F. J. V. (2001). Language biases in the coverage of the science citation index and its consequences for international comparisons of national research performance. *Scientometrics*, 51(1):335–46.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001.

Monks, J. and Ehrenberg, R. G. (1999). The impact of us news and world report college rankings on admission outcomes and pricing decisions at selective private institutions. *NBER Working Paper*, (7227).

Rauhvargers, A. (2013). Global university rankings and their impact: Report ii. Technical report, European University Association, Brussels.

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

Soh, K. (2011). Don't read university rankings like reading football league tables: Taking a close look at the indicators. *Higher Education Review*, 44(1):15–29.

Soh, K. (2014). Multicolinearity and indicator redundancy problem in world university rankings: An example using times higher education world university ranking 2013–2014 data. *Higher Education Quarterly*, forthcoming.

Stan Development Team (2014). *Stan Modeling Language: User's Guide and Reference Manual*. Stan Development Team.

Usher, A. and Savino, M. (2006). A world of difference: A global survey of university league tables. Technical report, Educational Policy Institute, Toronto, ON.

van Vught, F. A. and Ziegele, F., editors (2012). *Multidimensional Ranking: The Design and Development of U-Multirank*. Springer.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C., Tijssen, R. J., van Eck, N. J., van Leeuwen, T. N., van Raan, A. F., Visser, M. S., and Wouters, P. (2012). The leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, 63(12):2419–32.

Table 1. Rating Systems Used in the Analysis

Name & Institution	Abbrev.	Edition	Sample	Scope
World University Rankings *Times Higher Education*	THE	2014–15	400	Global
Academic Ranking of World Universities Shanghai Jiao Tong University	Shanghai	2014	100	Global
Center for World University Rankings King Abdulaziz University in Jeddah	Jeddah	2014	1000	Global
World University Rankings Quacquarelli Symonds	QS	2014	500	Global
Best Global Universities *US News & World Report*	USN-GU	1st	500	Global
Webometrics Ranking of World Universities Cybermetrics Lab, Spanish National Research Council	Webometrics	2nd 2014	1000*	Global
University League Table Complete University Guide	CUG	2015	123	UK
National University Rankings *US News & World Report*	USN-NU	2015	202	USA

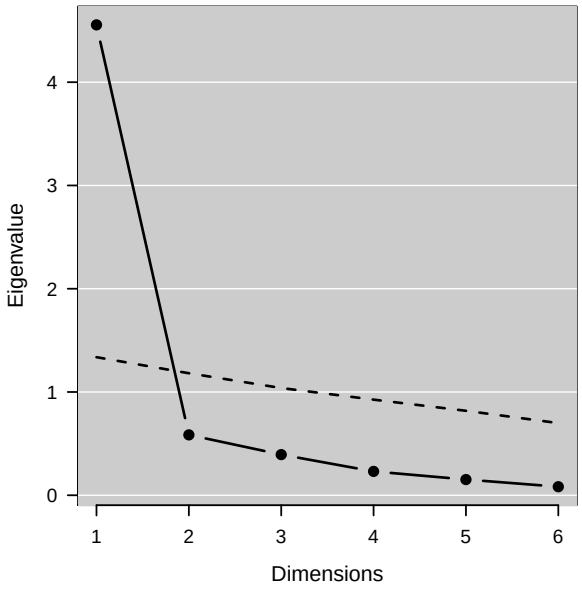
* Data for over 12000 universities available but only the top 1000 were included in this analysis.

Table 2. Inter-Rating System Correlations

	1	2	3	4	5	6	7	8
1. Shanghai	1.00							
	(100)							
2. THE	.76	1.00						
	(97)	(394)						
3. Jeddah	.84	.75	1.00					
	(100)	(385)	(1000)					
4. USN-GU	.81	.82	.86	1.00				
	(100)	(332)	(487)	(500)				
5. QS	.63	.77	.74	.75	1.00			
	(96)	(334)	(447)	(384)	(501)			
6. Webometrics	.46	.57	.69	.61	.52	1.00		
	(99)	(365)	(764)	(467)	(441)	(1000)		
7. CUG	.91	.72	.64	.36	.60	.77	1.00	
	(8)	(44)	(58)	(37)	(47)	(59)	(123)	
8. USN-NU	.67	.73	.78	.68	.75	.58	–	1.00
	(49)	(99)	(160)	(114)	(95)	(159)	(0)	(202)

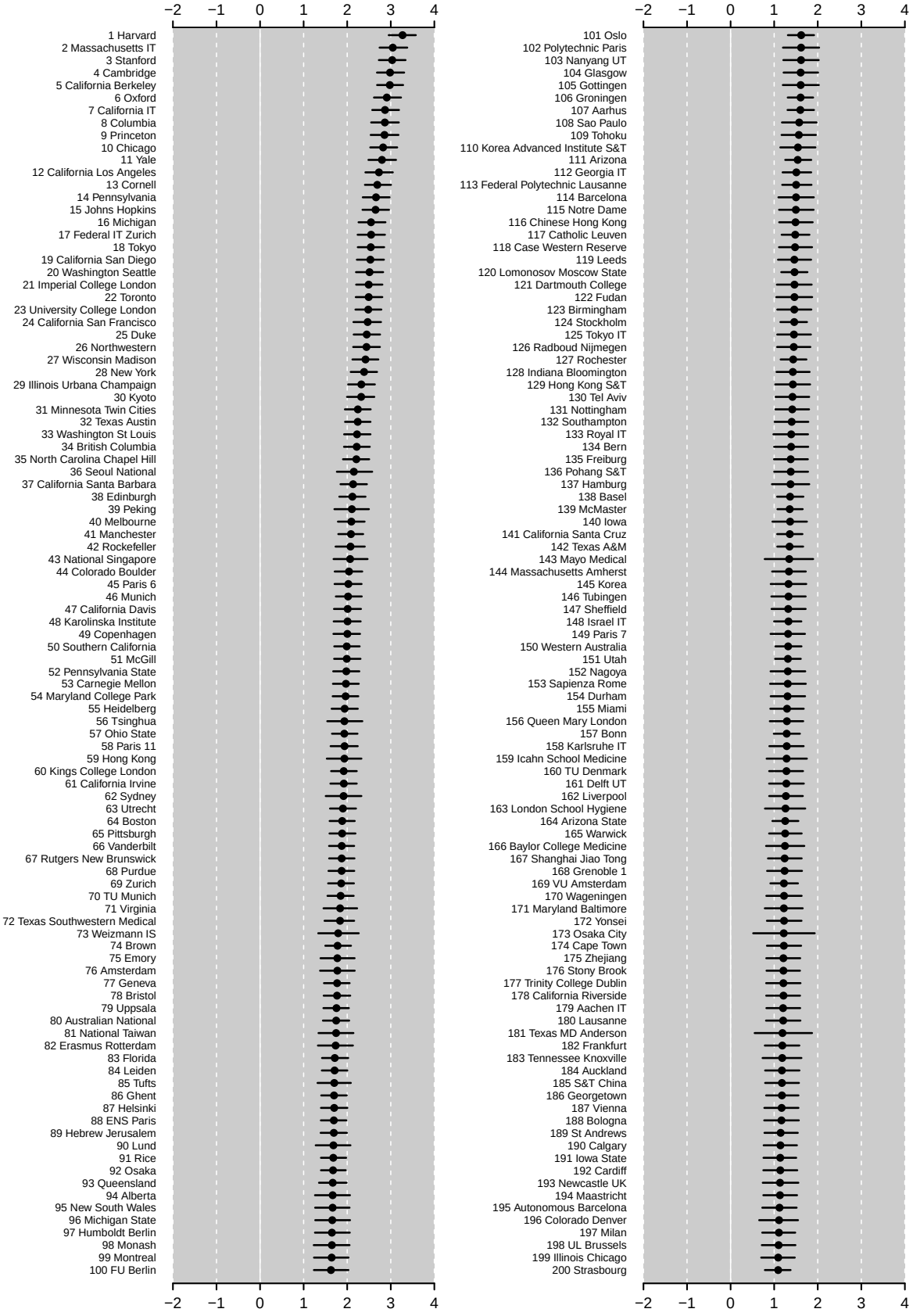
Pairwise Pearson's correlations with pairwise sample size in parentheses.

Figure 1. Scree Plot for Global Ratings Data



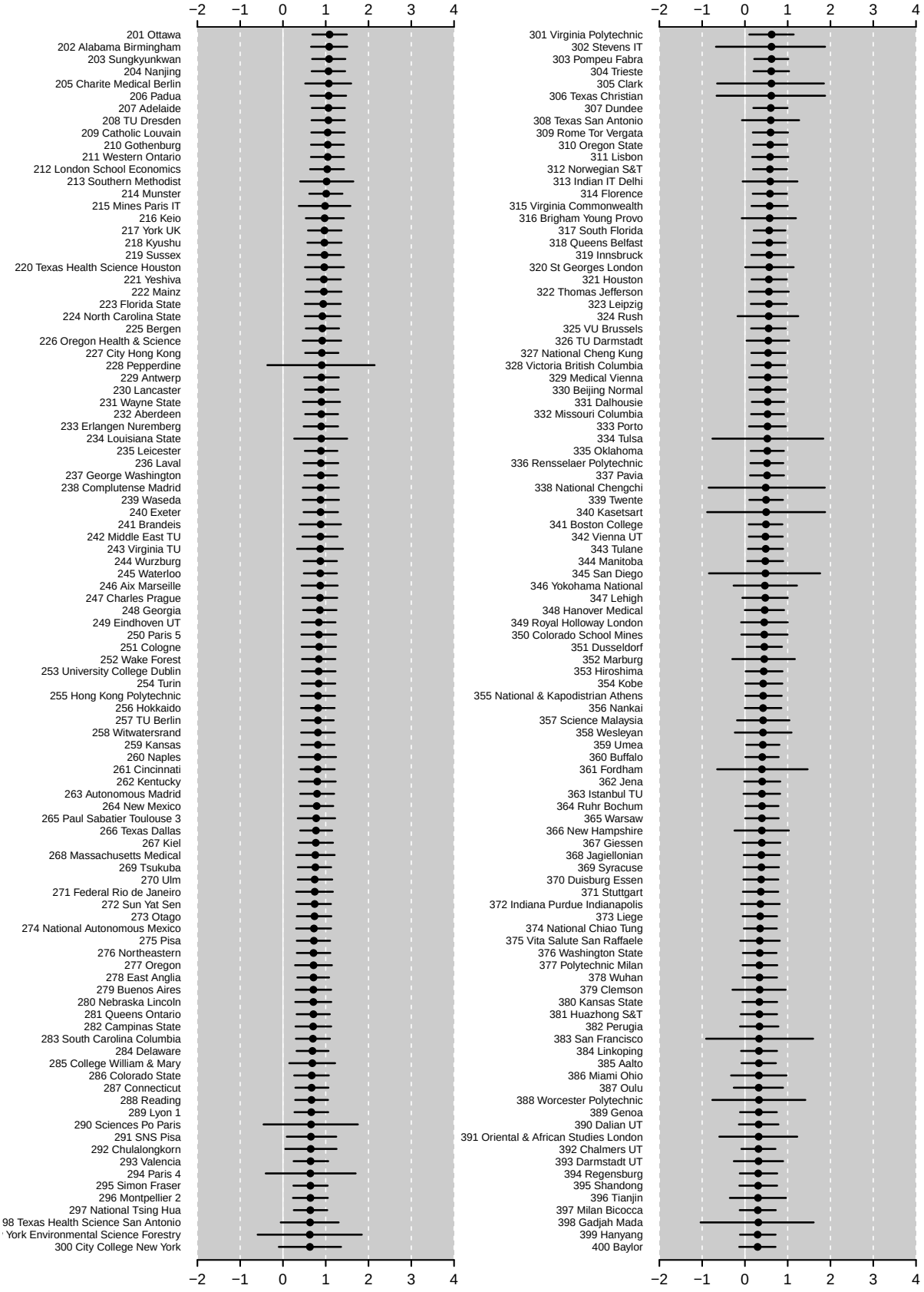
Points show the eigenvalues of the correlation matrix of the six global ratings measures. The dashed line shows the eigenvalues of a correlation matrix of six random normal variables.

Figure 2. Quality Estimates for Universities Ranked 1-200



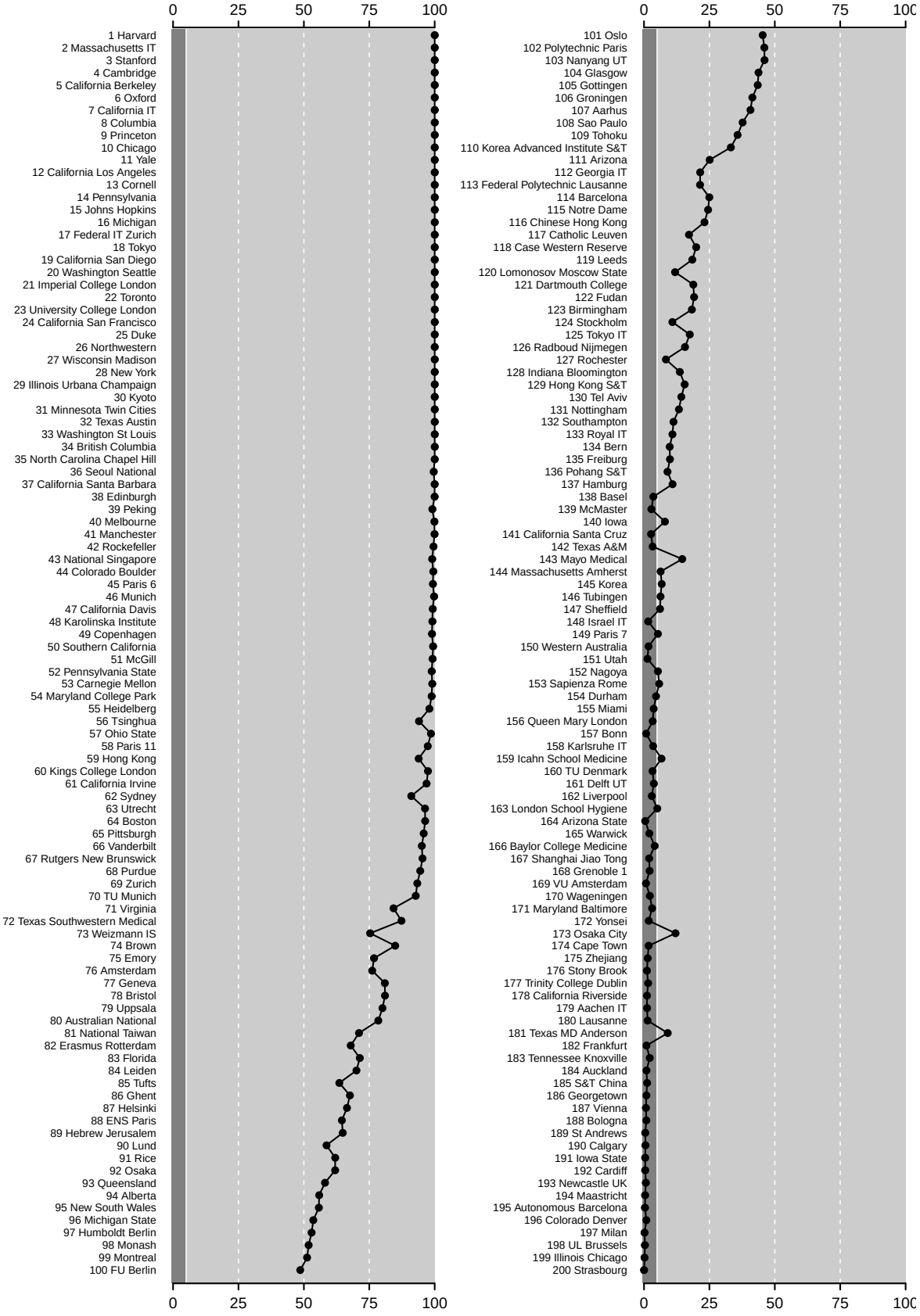
Points show the estimated quality for each university; bars are the 95% credible intervals and thus show the uncertainty of the estimates.

Figure 3. Quality Estimates for Universities Ranked 201-400



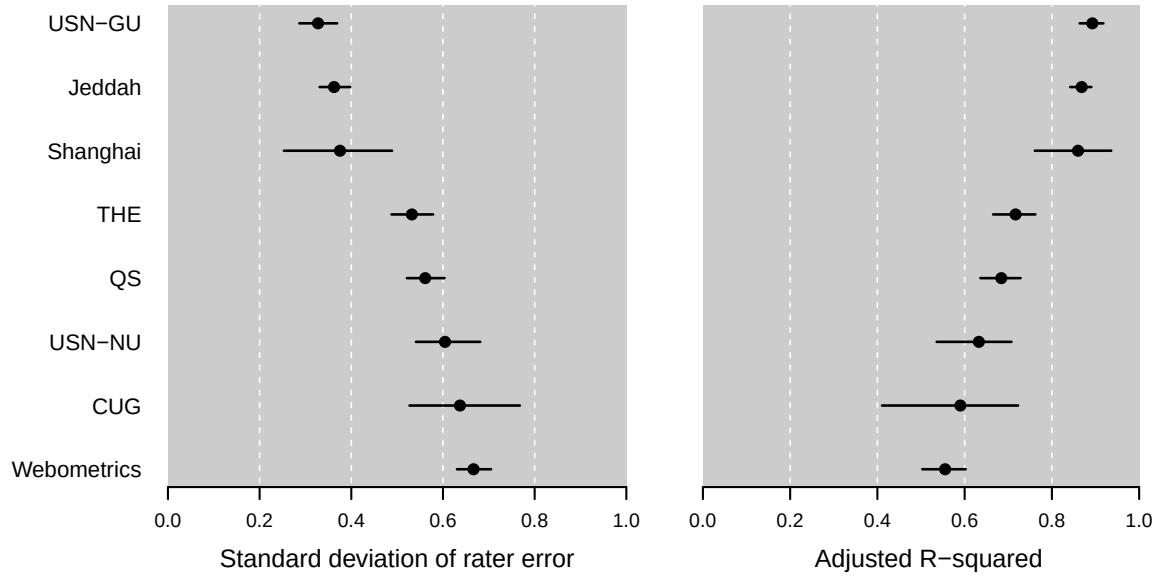
Points show the estimated quality for each university; bars are the 95% credible intervals and thus show the uncertainty of the estimates.

Figure 4. Probability that Universities Ranked 1-200 are in top 100



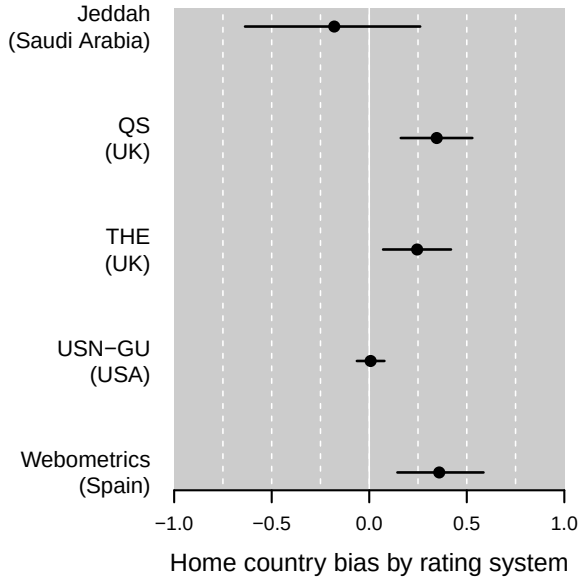
Points show the percentage of the 4000 MCMC samples where each university was one of the top-100 universities.

Figure 5. The Accuracy of the Rating Systems



Left plot: points show the standard deviation of the rater error terms (square root of ψ parameters). Right plot: points show the adjusted R-squared of the effects of latent university quality on each set of ratings. Bars are 95% credible intervals.

Figure 6. Home Country Bias by Rating System



Points show the mean residual within each global rating system and country (ϵ parameters). Bars are the 95% credible intervals. No Chinese university featured in the top 100 of the Shanghai rankings, so no home country bias estimate is possible for this system.

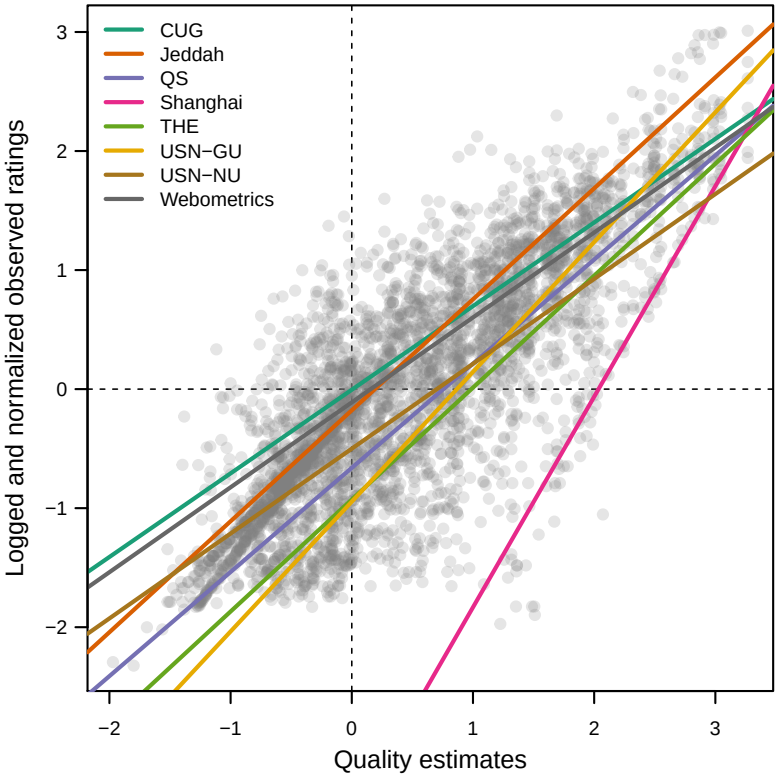
Appendix

Table 3. Parameter Estimates

Parameter	Mean	Std. Dev.
ν_{CUG}	−.01	.07
ν_{Jed}	−.18	.03
ν_{QS}	−.66	.04
ν_{Shan}	−3.60	.24
ν_{THE}	−.93	.05
ν_{USNGU}	−.95	.04
ν_{USNNU}	−.50	.06
ν_{Webm}	−.11	.03
λ_{CUG}	.70	.07
λ_{Jed}	.93	.02
λ_{QS}	.87	.04
λ_{Shan}	1.77	.12
λ_{THE}	.94	.04
λ_{USNGU}	1.09	.03
λ_{USNNU}	.71	.05
λ_{Webm}	.71	.03
ψ_{CUG}	.64	.06
ψ_{Jed}	.36	.02
ψ_{QS}	.56	.02
ψ_{Shan}	.37	.06
ψ_{THE}	.53	.02
ψ_{USNGU}	.33	.02
ψ_{USNNU}	.61	.04
ψ_{Webm}	.67	.02
ρ	−.62	.21
σ^2_ν	2.40	1.47
σ^2_λ	1.31	.82

Columns reports the means and standard deviations of the posterior distribution of MCMC 4000 samples for each parameter.

Figure 7. Rater Intercepts and Slopes with Fitted Data



Points show the university quality estimates (θ parameters), repeated for each rating system, and plotted against the observed ratings. Lines are the fitted intercepts (ν) and slopes (λ)for each rater.

Table 4. Country Bias by Global Rating System

	Jeddah	QS	Shanghai	THE	USN-GU	Webometrics
Australia	-.22	.30	-.19	.19	.11	.05
Brazil	-.24	-.29	–	-.43	.19	.24
Canada	.02	-.02	-.15	.05	-.03	.31
China	-.12	-.18	–	-.41	.05	.12
France	.05	-.08	.26	-.15	.05	-.70
Germany	.03	-.19	.02	-.11	.03	.28
Italy	.05	-.25	–	-.30	.08	-.19
Japan	.12	.21	.01	-.48	-.20	-.57
Netherlands	-.16	.35	-.15	.31	.09	.00
South Korea	.11	.25	–	.04	-.29	-.21
Spain	-.07	-.12	–	-.51	-.02	.36
Sweden	.05	.26	.20	-.07	-.09	.12
Taiwan	-.14	.21	–	-.26	-.04	.07
UK	-.09	.35	.05	.25	-.05	-.29
USA	.12	-.43	.00	.04	.01	.06

Each cell entry is the average residual within the corresponding rating system and country. The 15 countries with the most universities are included.

Figure 8. Distributions of the Raw Ratings

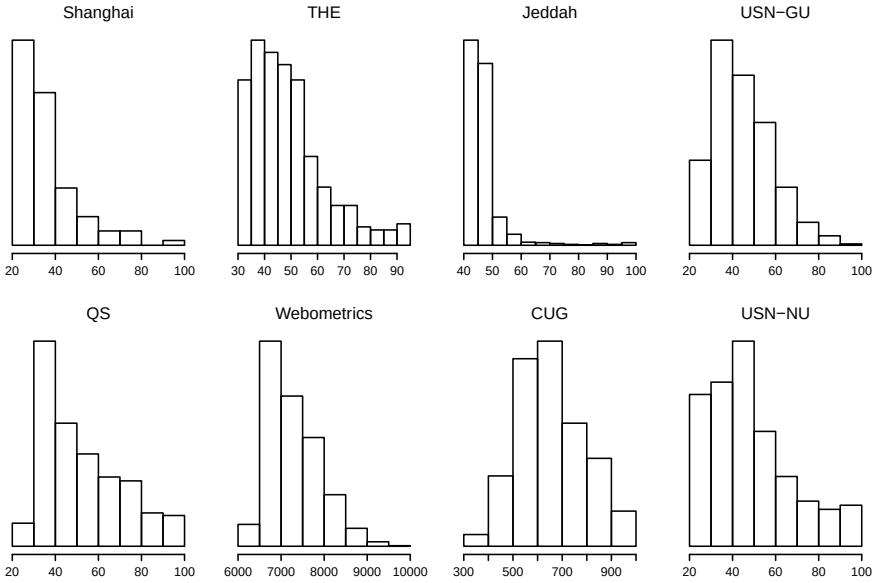


Figure 9. Distributions of the Transformed Ratings

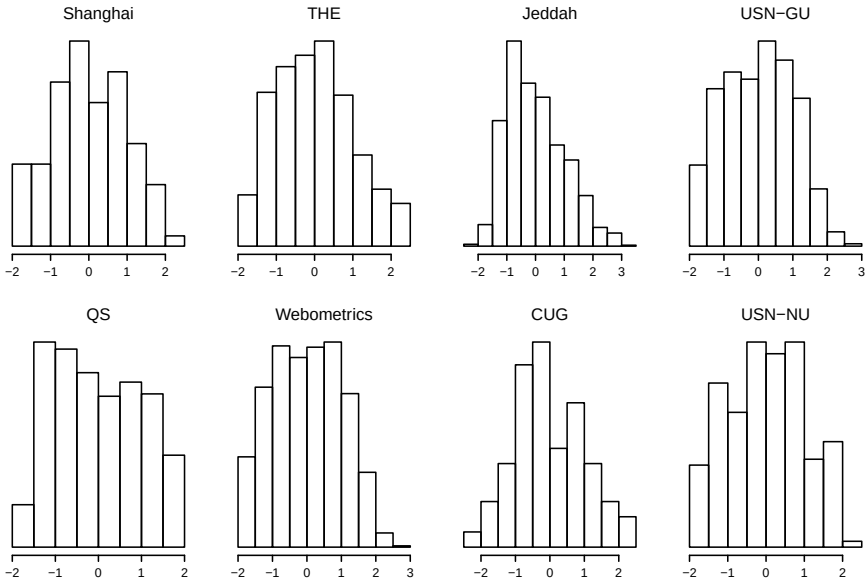
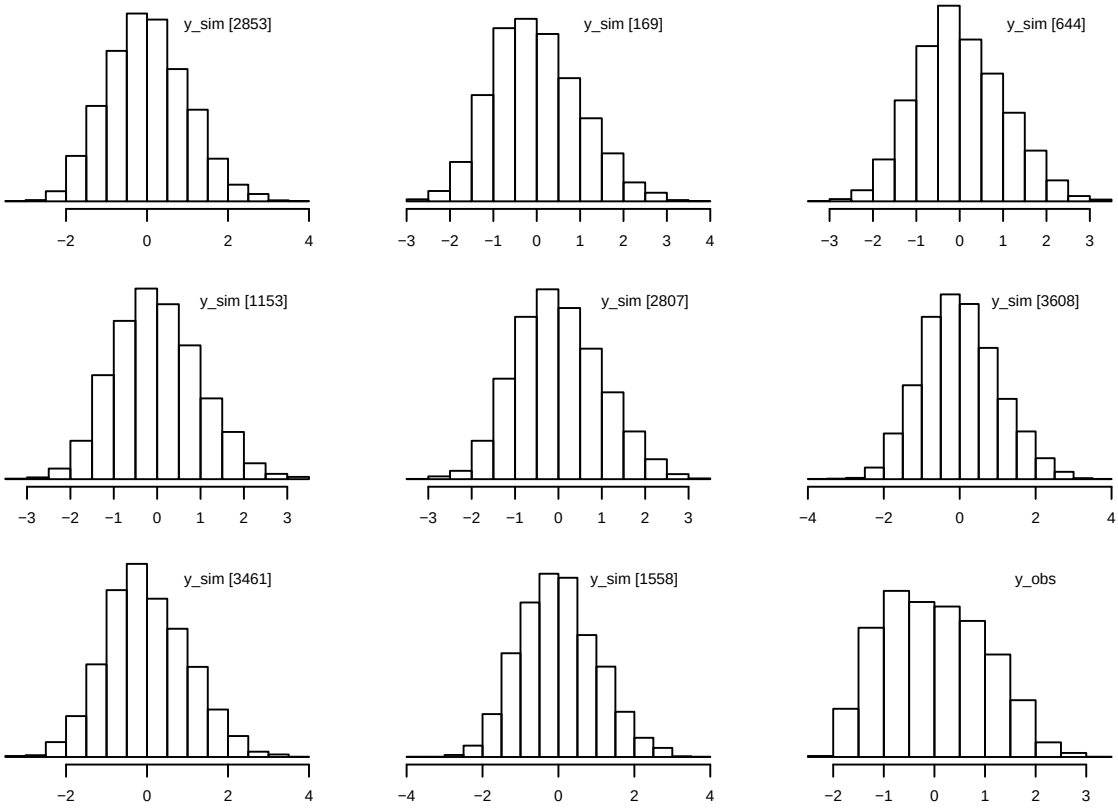


Figure 10. Posterior Predictive Checks



The first 8 plots show histograms of 8 independent draws of values from the posterior distribution of the outcome variable y_i. The final plot shows the histogram of the observed vector of data.