

Too much noise in the Times Higher Education rankings

Fred L. Bookstein · Horst Seidler · Martin Fieder · Georg Winckler

Received: 23 November 2009 / Published online: 10 February 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract Several individual indicators from the Times Higher Education Survey (THES) data base—the overall score, the reported staff-to-student ratio, and the peer ratings—demonstrate unacceptably high fluctuation from year to year. The inappropriateness of the summary tabulations for assessing the majority of the “top 200” universities would be apparent purely for reason of this obvious statistical instability regardless of other grounds of criticism. There are far too many anomalies in the change scores of the various indices for them to be of use in the course of university management.

Keywords Times Higher Education ranking · Rankings · Statistical noise

The 2009 Times Higher Education Survey (THES) ratings appeared on October 8 of that year to considerable interest and some chagrin worldwide. We understand that one primary context for these tabulations is a social one, the assortment of students to “the world’s leading universities” across the six continents and their regions. There is, however, another common use of these rankings: to inform individual universities or faculties about how the institutions themselves are faring from year to year. We have carried out a small statistical analysis of some of the publicly available data from the THES websites, and draw some conclusions about the validity of this second use.

F. L. Bookstein (✉) · H. Seidler · M. Fieder
Department for Anthropology, University of Vienna, Vienna, Austria
e-mail: fred.bookstein@univie.ac.at

F. L. Bookstein
Department of Statistics, University of Washington, Seattle, WA, USA

G. Winckler
Rectorate of the University of Vienna, Vienna, Austria

Overall scores

The overall score of the top 200 universities has ranged from 100.0 (Harvard) down to about 54 points in the last 3 years. Most schools (about 95%) remain in the top 200 from year to year, but their positions are often unstable. There is a statistical association between the overall score and the variation of its change. Universities in approximately the top one-third of the 200-university list (down through about score 73) show a year-to-year standard deviation of only about 1.5 score points, but the schools below this cutoff systematically show a year-to-year standard deviation of 3.25 points, which is more than a doubling of the year-to-year instability (four times the variance). See the first accompanying figure (Fig. 1).

Three components that together hold the majority of the weight in the overall THES score—peer review, employer review, and citations per faculty member—originate in what are called Pareto distributions, also known as “power law” distributions similar to those for sizes of towns, personal wealth, or fame. When a power-law distribution changes under some year-to-year fluctuation having a standard deviation of a few percent, the top few scores will remain the top few, while those in the mass can alter their relative positions by a considerable distance. Because the raw measurements contributing to the THES are converted to a 0-to-100 scale first, this fallacy creeps into every single component, and thereby into their weighted average, the overall score, adding this additional variance disproportionately in the lower range. To be considered as providing valid information on ranks beyond 73, the distribution function needs to have a stable year-to-year variance independent of score.

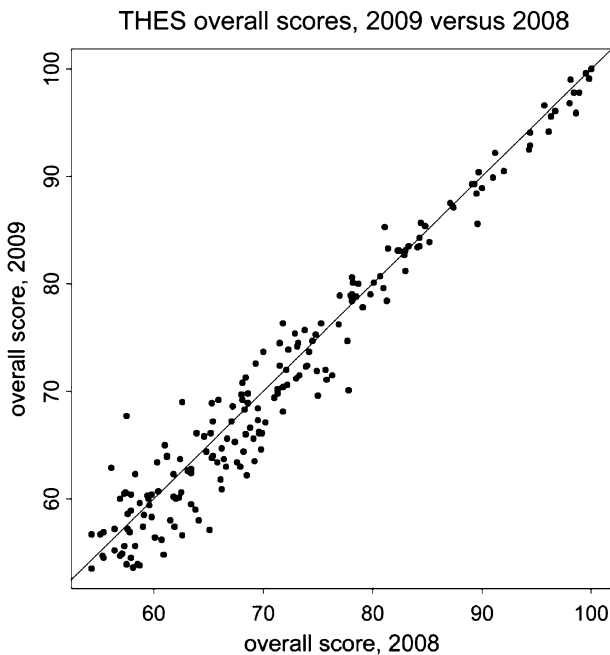


Fig. 1 THES overall scores, 2009 versus 2008. Each point is one university that ranked in the top 200 universities worldwide in both years. The diagonal is the line of no change in scores from year to year

Year-to-year changes in the THES data base can be examined in at least three categories: changes in “overall score” as we have just been examining, changes in the objective components of that score as submitted by the surveyed universities themselves, and changes in the so-called “peer ratings,” which are mentions of the university faculty in question by others regionally or worldwide. Like the overall score, each of these other two components shows serious statistical problems when year-to-year shifts are examined in detail.

One component score supplied by the universities themselves

One “objective” component of the THES scores is staff/student ratio. The correlation between staff/student ratio 2007 and staff/student ratio 2009 is about 0.84. But this relation appears to be a mixture of two quite different subgroups: a set of schools whose scores change hardly at all between successive years, but also a set of about 40 schools whose scores change radically in one or the other direction. This is unreasonable. Staff/student ratio must be a stable quantity in any academic institution above a certain size; major year-to-year changes must be expressing changes in definition, not changes in organizational membership. See the year-by-year scatterplots in the second accompanying figure.

Because they constitute a full 20% of the overall score, these incomprehensibly large year-to-year shifts in staff-student ratio provide a new insight into the fluctuation of the net rating. For whatever reason, the quantity THES assigned to the University of Copenhagen staff-student ratio went from 51 (the sample median) in 2007 to 100 (a score attained by only 12 other schools in the top 200) in 2008, and remained at 100 in 2009. The effect was to raise Copenhagen’s overall score in both 2008 and 2009 by one-fifth of 49, a full 9.8 points. Without this boost, Copenhagen’s 2009 score would have been 69.0 instead of 78.8 and the corresponding ranking would have been 94 instead of 51. Another school with a 100 student-staff rating in 2009, Ecole Normale Supérieure, Paris, rose from the value of 68 just a year earlier, thereby adding 6.4 points to its total score and thus earning a ranking of 28 instead of 48. Changes this large in this ratio must be due to changes in definition of one or the other terms of the ratio, rather than real changes in the institutions (Fig. 2).

Some year-on-year changes in the “peer rankings”

Many year-on-year changes in the “peer rankings” of the five faculty subdivisions show unacceptable instability of variance from year to year. In the most glaring example, the

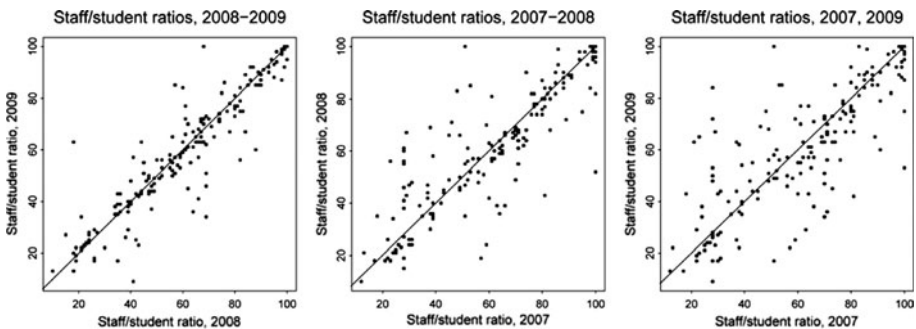


Fig. 2 Correlation of staff/student ratio between the years 2008/2009, the years 2007/2008 and the years 2007/2009. The scores are those supplied by the universities themselves, converted to ranks

variance of the log (base e) of the peer life sciences/biomedicine ranking is .048 for the change 2007 to 2008 but a full .104 for the change 2008 to 2009, as shown in the third figure. We know of nothing in the actual setting of academic life science or biomedicine that could possibly account for this enormous shift of year-on-year variance between two consecutive years; it must have something to do with changes in THES procedure.

A page of the QS website for this study (www.topuniversities.com/university-rankings/world-university-rankings/methodology/simple-overview) notes that the number of “peer raters” jumped by about 50% from 2008 to 2009 (from 6354 to 9386). We have no idea how all these new raters may have systematically differed from the raters of the earlier cohort. (Those six thousand for 2008 are, apparently, a considerable increase from the 1,600 raters for the exercise of just 2 years earlier, 2006: Ioannidis et al. 2007). Furthermore, raters seem to have been offered a range of rewards for submitting a rating, including a “1 month free trial subscription to Times Higher Education” or “the chance to win a free Blackberry personal organizer.” What does it mean about the ratings that nonsubscribers to THES are thereby more likely induced to rank their peers? (Fig. 3).

The instructions to the peer raters for 2008 (http://research.qsnetwork.com/qs_surveysystem/index.php?survey=52&viewonly) regarding the ratings of universities are inappropriately vague. They read, in full: We ask you only to use your best knowledge of the broad subject area you have selected. Please select up to 30 universities from our international list that you regard as producing the best research in the... subject area. (Or the same for the “domestic” list, which is specific to smaller regions.) The “subject area” was one of Arts and Humanities, Life Sciences and Biomedicine, Natural Sciences, Social Sciences, and Technology.

It is unclear by what algorithm a variable number of mentions in an unordered list is converted into a summary “peer ranking,” nor what is a rational behavior for the reader who might wish to know when to stop listing universities. We cannot separate the factors of fame, notoriety, haphazard name recognition, or incidental knowledge from the underlying dimension of excellence at which the survey instrument is presumably aimed. The more than doubling of change variance in the life sciences score from 2007–2008 to 2008–2009 is surely due in great part to the complete absence of any explanation to the raters about just how the information sought is to be analyzed, in addition to the enormous change in the panel of raters already noted.

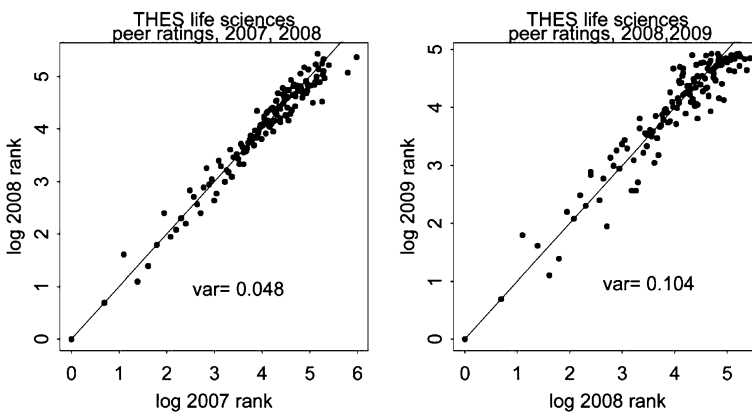


Fig. 3 Some year-on-year changes in the “peer rankings” of the five faculty subdivisions show unacceptable instability of variance from year to year

Conclusion

Most of the THES rating system appears to show unacceptable year-to-year variances. They are not likely to be a useful component of any rational approach to academic management. This instability can only strengthen the existing critique of the overall ranking system by earlier evaluators, such as van Raan (2005), who highlighted the invalidity of yet another component of these totals, the bibliometric component (the citation-based scores).

Note added in revision

On December 16, 2009, Times Higher Education magazine announced that it is replacing the peer rating component of the THES by a restructured opinion survey involving at least 25,000 responses “from a carefully targeted and properly representative sample.” “We want to aspire to the standards that university social scientists expect to see,” said Jonathan Adams, director of research evaluation at Thomson Reuters. The replacement of the old survey methodology by a new one will of course render changes from 2009 overall scores to 2010 overall scores uninterpretable, but may stabilize the year-on-year variances of this component later in the decade.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Ioannidis, J. P. A., Patsopoulos, N. A., Kavvoura, F. K., Tatsioni, A., Evangelou, E., Kouri, I., et al. (2007). International ranking system for universities and institutions: A critical appraisal. *BMC Medicine*, 5, 30. doi:[10.1186/1741-7015-5-30](https://doi.org/10.1186/1741-7015-5-30).
- Raan, F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62(1), 133–143.