Value measurement theory and league tables

AUTHOR:
Theodor J. Stewart[1]

AFFILIATION:
[1]Emeritus Professor, Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Theodor Stewart

EMAIL:
theodor.stewart@uct.ac.za

POSTAL ADDRESS:
Department of Statistical Sciences, University of Cape Town, Rondebosch 7701, Cape Town, South Africa

KEYWORDS:
multicriteria decision analysis; university rankings; scoring systems

HOW TO CITE:
Stewart TJ. Value measurement theory and league tables. S Afr J Sci. 2014;110(3/4), Art. #a0057, 2 pages. http://dx.doi.org/10.1590/sajs.2014/a0057

This communication was stimulated by the editorial in the Sep/Oct 2013 issue of the *South African Journal of Science* concerning university rankings, which rang a strong sympathetic chord with me. I have for a long time felt a professional responsibility as a decision scientist or decision analyst (specialising in multicriteria decision analysis) to respond to the irresponsible use of scoring systems, league tables, etc., which so many people accept uncritically.

From a decision analytic point of view, most ranking systems are based on a simple multicriteria value function, typically additive in structure. In other words, for each alternative to be ranked (largely universities in this discussion), a score is calculated of the form $V_a = \sum_{i=1}^{m} W_i V_{ai}$, where W_i is the importance weight associated with criterion i (out of m), and V_{ai} is the evaluation (or partial score) achieved by alternative a with respect to criterion i.

There exists a substantial theory on the validity and properties of such value functions (which are reviewed for example in Belton and Stewart[1]), but the developers of ranking systems show little evidence of awareness of this theory. In particular, there are two assumptions which are critical to the validity of additive value functions:

- The criteria on which they are based must be *preferentially independent*. This is a frequently misunderstood condition and has nothing to do with statistical independence; it is the ability to express trade-offs between two criteria, on the assumption that performances of all other criteria are the same, independent of the levels of these constant performances. It is not evident that preferential independence assumptions have received much, if any, attention by developers of university ranking schemes.
- The scores on each criterion must satisfy the properties of an *interval scale of preference*, so that the differences in scores have the same importance irrespective of the baseline level. Consider for example two alternatives (universities), say a and b. These would achieve equal ranks if and only if $\sum_{i=1}^{m} W_i[V_{ai} - V_{bi}] = 0$ which depends entirely on the differences $V_{ai} - V_{bi}$ and not on their absolute values. Natural measurements do not naturally satisfy this interval preference scale property – a change in publication count, for example, from 100 to 150 pa cannot in general be associated with the same increase in preference as a change from 1000 to 1050. The definition and mode of assessment of the partial scores V_{ai} can thus be critical to the validity of an additive model (which is discussed below).

It is self-evident (I hope) that no ranking system is an objective measure of performance. It includes a variety of value judgements that should be assessed separately by any users, without uncritically accepting the values used by publishers of rankings. This subjectivity is perhaps quite widely recognised (if not acted upon) in regard to the importance weightings W_i. But sensitivity to the definition of the partial scores V_{ai} is not widely recognised, and yet has been shown in a number of studies (e.g. Stewart[2,3]) to have potentially even larger impacts on output rankings generated by the additive model than that of variations in the importance weights.

Chapter 5 of Belton and Stewart[1] details the processes that need to be undertaken in order to establish defensible assessments of the partial scores. These processes are quite demanding, and developers of league tables resort to simpler approaches. Two common approaches are to select some easily available attribute related to the criterion of interest (e.g. numbers of publications as a surrogate for research output) and to set V_{ai} equal to a linearly scaled value of the attribute or to simply rank order the universities according to the criterion, and then equate the partial scores to quantile values for an assumed population distribution (e.g. the so-called Z-scores assuming a Gaussian distribution). There are no credible grounds for assuming that scores obtained in either of the above manners correspond to the required interval scale for preferences, so validity of any derived rankings must be questioned.

As illustration of the last claim, the use of Z-scores (quantiles of a Gaussian distribution) for defining partial scores is examined. This approach would be valid for a large number of alternatives if and only if the distribution of values to decision-makers was normal for the population being assessed. No argument from a central limit theorem makes sense here – when looking at the specific population being ranked. On prima facie grounds I find the following implications of Z-scores from a normal distribution to be highly implausible as rules to be generally applied in all cases:

- The gain in moving from the alternative appearing as the 25th quantile in the population to the median alternative is deemed to be equivalent to the gain in moving from the median to the 75th quantile
- The gain in moving from the alternative appearing as the 90th percentile to the 95th is deemed to be less than half the gain in moving from the 95th to the 99th

In my experience of fitting value functions by the approaches described in Chapter 5 of Belton and Stewart[1], non-symmetrical value functions are typical (invalidating the first implied property), while the top few alternatives are often not strongly distinguished (invalidating the second). The onus must be on the developers of the ranking system to provide valid evidence for the preference gaps being assumed.

About all that can be inferred with complete justification from the data shown in most league tables is that if one alternative (university) dominates another, in the sense of being better on all criteria, then it is the better of the two (although even here we must assume that no important criteria have been omitted). But, in general, such dominance properties will not provide particularly complete rankings – comparisons between most pairs of alternatives would be indeterminate. My conclusion is that extreme caution and scepticism needs to be applied to league tables, such as university rankings, before they are used for any significant decision-making. This caveat applies equally well to students (and their families) choosing universities, and to university administrators either in boasting about improved league positions or in allocating resources to improving league position.

References

1. Belton V, Stewart TJ. Multiple criteria decision analysis: An integrated approach. Boston, MA: Kluwer Academic Publishers; 2002. http://dx.doi.org/10.1007/978-1-4615-1495-4

2. Stewart TJ. Use of piecewise linear value functions in interactive multicriteria decision support: A Monte Carlo study. Manage Sci. 1993;39:1369–1381. http://dx.doi.org/10.1287/mnsc.39.11.1369

3. Stewart TJ. Robustness of additive value function methods in MCDM. J Multi-Crit Decis Anal. 1996;5:301–309. http://dx.doi.org/10.1002/(SICI)1099-1360(199612)5:4<301::AID-MCDA120>3.0.CO;2-Q