University of Hertfordshire

# Add or multiply? A tutorial on ranking and choosing with multiple criteria

**Chris Tofallis**

Hertfordshire Business School Working Paper (2014)

The Working Paper Series is intended for rapid dissemination of research results, work-in-progress, and innovative teaching methods, at the pre-publication stage. Comments are welcomed and should be addressed to the individual author(s). It should be noted that papers in this series are often provisional and comments and/or citations should take account of this.

Hertfordshire Business School Working Papers are freely downloadable from https://uhra.herts.ac.uk/dspace/handle/2299/5549 and also from the British Library: www.mbsportal.bl.uk

*Copyright and all rights therein are retained by the authors. All persons copying this information are expected to adhere to the terms and conditions invoked by each author's copyright. These works may not be re-posted without the explicit permission of the copyright holders.*

Hertfordshire Business School employs approximately 200 academic staff in a state-of-the-art environment located in Hatfield Business Park. It offers 17 undergraduate degree programmes and 21 postgraduate programmes; there are about 75 research students working at doctoral level. The University of Hertfordshire is the UK's leading business-facing university and an exemplar in the sector. It is one of the region's largest employers with over 2,650 staff and a turnover of almost £233 million. It ranks in the top 4% of all universities in the world according to the Times Higher Education World Rankings and is also one of the top 100 universities in the world under 50 years old. In the last UK Research Assessment Exercise it was given the highest rank for research quality among the post-1992 universities.

# Add or multiply? A tutorial on ranking and choosing with multiple criteria

*Accepted for publication in INFORMS Transactions on Education.*

**Chris Tofallis**
**The Business School**
**University of Hertfordshire**
**College Lane**
**Hatfield**
**Herts, AL10 9AB**
**UK**

**c.tofallis@herts.ac.uk**

## ABSTRACT

Simple additive weighting is a well-known method for scoring and ranking alternative options based on multiple attributes. However the pitfalls associated with this approach are not widely appreciated. For example, the apparently innocuous step of normalizing the various attribute data in order to obtain comparable figures leads to markedly different rankings depending on which normalization is chosen. When the criteria are aggregated using multiplication such difficulties are avoided because normalization is no longer required. This removes an important source of subjectivity in the analysis because the analyst no longer has to make a choice of normalization type. Moreover, it also permits the modelling of more realistic preference behaviour, such as diminishing marginal utility, which simple additive weighting does not provide. The multiplicative approach also has advantages when aggregating the ratings of panel members. This method is not new but has been ignored for too long by both practitioners and teachers. We aim to present it in a non-technical way and illustrate its use with data on business schools.

***KEYWORDS**:*

*multi-attribute decision making, ranking, scoring, aggregation, selection, teaching management science, teaching decision analysis, multiple criteria*

# 1. Introduction

Most real-life decisions involve *multiple criteria,* yet many business and management university courses, MBA included, do not deal explicitly with this topic within any quantitative module. It is a remarkable fact that a number of current OR/MS textbooks do not cover the topic of multiple criteria decision analysis e.g. Hillier and Lieberman (2010), Powell and Baker (2011). Decision makers are sometimes faced with choices between a number of alternative options where multiple factors have to be taken into consideration. We consider the situation where the selection is to be based on quantitative data. A common approach is to attach weights to the factors (criteria or attribute values, which are first 'normalized' in some way) and then produce an overall score by adding the weighted scores. We shall refer to this as 'simple additive weighting' and highlight some misconceptions which surround it. We then look at an alternative approach based on multiplication rather than addition.

The paper is suitable for undergraduate students and practitioners in business, management, or other fields where decisions under multiple criteria appear. The technical level has been kept low: apart from arithmetic, there is very little mathematics involved and this has been placed in the appendices.

# 2. Common normalizations

A standard step in processing the data prior to aggregation is the normalization step. The aim is to make the magnitude of the numbers 'comparable' across criteria, and to avoid adding together quantities measured in different units.

A number of normalization methods are available and we now describe the most commonly used ones. Each criterion is treated separately in what follows.

1. Divide by the largest value. This causes the largest value on each criterion to equal unity and all others represent a fraction of the largest value. Equivalently, one can express each value as a percentage of the largest observed score on that criterion.

2. Range normalization. As before, the largest value is converted to 1 or 100% as above, but furthermore the lowest observed value is converted to zero. Thus the range has an actual observation at each end, and all criteria now have an equal range. The formula for converting the data is:

$$\text{Range normalized score} = (x - x_{min})/(x_{max}\text{-}x_{min})$$

where $x_{max}$ and $x_{min}$ are the largest and smallest observed values respectively. (For those criteria where lower scores are preferred the numerator is replaced by $x_{max}$-$x$).

Unfortunately, this type of normalization destroys proportionality: If, say, one raw score is double that of another, then this proportionality will not be preserved for the range normalized scores. This can easily be overlooked by the non-technical user. Consider, for example, the case of choosing a school where one of the criteria is class size. The range of class size may be small, say from 28 to 33, in which case a small difference between schools may appear very large when normalized according to this approach.

3. z-scores (statistical standardization). In statistics standardized scores (z-scores) are obtained by subtracting the mean and then dividing by the standard deviation. So, for example, a value of $z$ = -1 indicates a value that is one standard deviation below the mean. Despite being widely used, z-scores do have disadvantages which are not always appreciated. As with range normalization, proportionality is lost: doubling the x-value corresponding to $z$ = -2, does not correspond to $z$ = -4. Secondly, z-scores are sensitive to outliers: one or more extreme observations can make the mean value extremely unrepresentative. This affects the z-score calculation in two ways: in the mean subtraction step, as well as in the calculation of the standard deviation.

4. Dividing by the total. For each criterion we add together all the values, and divide each one of them by this total. This allows scores to be viewed as proportions of some whole, which may or not be a meaningful interpretation. Here proportionality of scores is retained.

We now illustrate the above using data and attribute weightings from Yoon and Hwang (1995, p.41) as shown in Table 1.

Table 1. Attributes of six candidates with weightings. (GRE is Graduate Record Examinations)

| Candidate | GRE | GPA | College Rating | Recommendation Rating | Interview Rating |
|---|---|---|---|---|---|
| A | 690 | 3.1 | 9 | 7 | 4 |
| B | 590 | 3.9 | 7 | 6 | 10 |
| C | 600 | 3.6 | 8 | 8 | 7 |
| D | 620 | 3.8 | 7 | 10 | 6 |
| E | 700 | 2.8 | 10 | 4 | 6 |
| F | 650 | 4 | 6 | 9 | 8 |
| *Weight* | *30%* | *20%* | *20%* | *15%* | *15%* |

If we apply the above normalizations and weightings to this data we can produce the corresponding scores and hence ranks. The resulting rankings are shown in Table 2.

Table 2. Rankings of six candidates according to additive scoring based on four different normalizations, and according to the multiplicative approach (discussed later in the paper).

| Candidate | Divide by max | Range norm. | z-score | Divide by sum | Multiplicative |
|---|---|---|---|---|---|
| A | 5 | 3 | 3 | 5 | 5 |
| B | 3 | 6 | 5 | 3 | 4 |
| C | 4 | 5 | 6 | 4 | 3 |
| D | 2 | 4 | 4 | 2 | 2 |
| E | 6 | 2 | 2 | 6 | 6 |
| F | 1 | 1 | 1 | 1 | 1 |

The key point is that the four normalizations do not agree with each other. Candidate B's rank ranges from third down to last. Whilst candidate E's rank ranges from last up to second!

Further examples can be found in Pavlicic (2000), including one case where a candidate comes first for one normalization but comes last according to two others!

There is something clearly unsatisfactory about a method when one of its key steps can be arbitrarily chosen and yet leads to different outcomes. As Pomerol and Barba-Romero (2000, p. 80) starkly put it:

"Why then is a method which is so demanding in theoretical assumptions, and so prone to influence by arbitrary choices at the time of application, so widely (and sometimes badly) used?"

In the next section we look at the issues arising in more detail.

# 3. Three surprising effects of simple additive weighting

## 3.1 Rank reversal due to a change in normalization

Those who are well-versed in multiple criteria analysis will be able to see right through the following example; but for those with less experience there is a valuable lesson here: when an analyst asks you for a set of weights and you don't know exactly how they will be used, then the recommended option will depend crucially on how knowledgeable your analyst is. Consider the

simple example in Table 3 involving just two candidates (from Pomerol and Barba-Romero [2000, p.78]).

Table 3. Two candidates with two criteria.

| Candidate | Attribute 1 | Attribute 2 |
|---|---|---|
| **A** | **9** | **3** |
| **B** | **1** | **8** |

Suppose the decision-maker chooses a weight of 2/5 for the first attribute and 3/5 for the second. If we use the first type of normalization – dividing by the maximum - then B achieves the highest score as shown in Table 4.

Table 4 Weighted average score using type 1 normalization.

| Candidate | Attribute 1 (weight 2/5) | Attribute 2 (weight 3/5) | Weighted sum |
|---|---|---|---|
| **A** | **9/9** | **3/8** | **0.625** |
| **B** | **1/9** | **8/8** | **0.644** |

If instead, we apply the fourth type of normalization - dividing by the total – we then discover that the highest score now goes to candidate A (Table 5). This unexpected reversal has nothing to do with a loss of proportionality since we have used normalizations which maintain proportionality.

Table 5 Weighted average score using type 4 normalization.

| Candidate | Attribute X (weight 2/5) | Attribute Y (weight 3/5) | Weighted sum |
|---|---|---|---|
| **A** | **9/10** | **3/11** | **0.524** |
| **B** | **1/10** | **8/11** | **0.476** |

To those who are not trained in decision theory this may appear to be a surprising or paradoxical result. The apparently innocuous step of normalizing the data for the sake of convenience should not affect the final decision, but it does. Let us see why this happens. Suppose that the raw data has X measured in $k and Y is years of experience.

If we normalize using 'divide by the maximum' and use the given weights of 2/5 and 3/5 we obtain: Additive score = 2/5 ($X$/9) + 3/5 ($Y$/8) = 0.044$X$ + 0.075$Y$

which implies that 1 year of experience is worth about $1.7k, (since 0.075/0.044 = 1.7).

Using the 'divide by total' normalization we get:

Additive score = 2/5 ($X$/10) + 3/5 ($Y$/11) = 0.04$X$ + 0.0545$Y$

which means that 1 year of experience is now worth $0.0545/0.04 = $1.36k !

There is a difference and hence a lack of agreement.

It is likely that compilers of rankings in popular publications are not aware of the important point that different normalizations require different weights. It is not sufficient to ask people for weights in isolation; the elicitation procedure needs to take account of how they will subsequently be applied.

### 3.2 Rank reversal due to removal of unwanted candidates: making a shortlist can make an unexpected difference!

Let us take a look at another example (Table 6), which will highlight a different problem, as described in Pavlicic (2000). It does not even involve switching from one method of normalization to another as above. This time we shall stick to just one form of normalization, but we stress that the same difficulty can arise with all four of the above approaches.

Table 6 Five candidates with four criteria.

| Candidate | Attribute 1 | Attribute 2 | Attribute 3 | Attribute 4 | Score |
|---|---|---|---|---|---|
| **A** | **2** | **7** | **180** | **100** | **3.00** |
| **B** | **3** | **8** | **140** | **98** | **2.98** |
| **C** | **5** | **5** | **155** | **100** | **2.97** |
| **D** | **1** | **9** | **155** | **95** | **2.92** |
| **E** | **9** | **1** | **160** | **90** | **2.90** |
| ***Maximum*** | ***9*** | ***9*** | ***180*** | ***100*** | |

For simplicity let us use equal weights. If we divide this raw data by the maximum in each column and then add up the values to obtain total scores we find that all the ranks are in the same order as shown in the table i.e. with A highest and E lowest. Now suppose that in producing the short-list, candidates D and E were rejected because of their very poor scores on attributes 1 and 2 respectively. If the same normalization (divide by the maximum) is then carried out on the reduced table (Table 7) we find that the ranking is completely *reversed*, with C coming top, B second and A last!

Table 7 A shortlist produced by removing rejected candidates from Table 4.

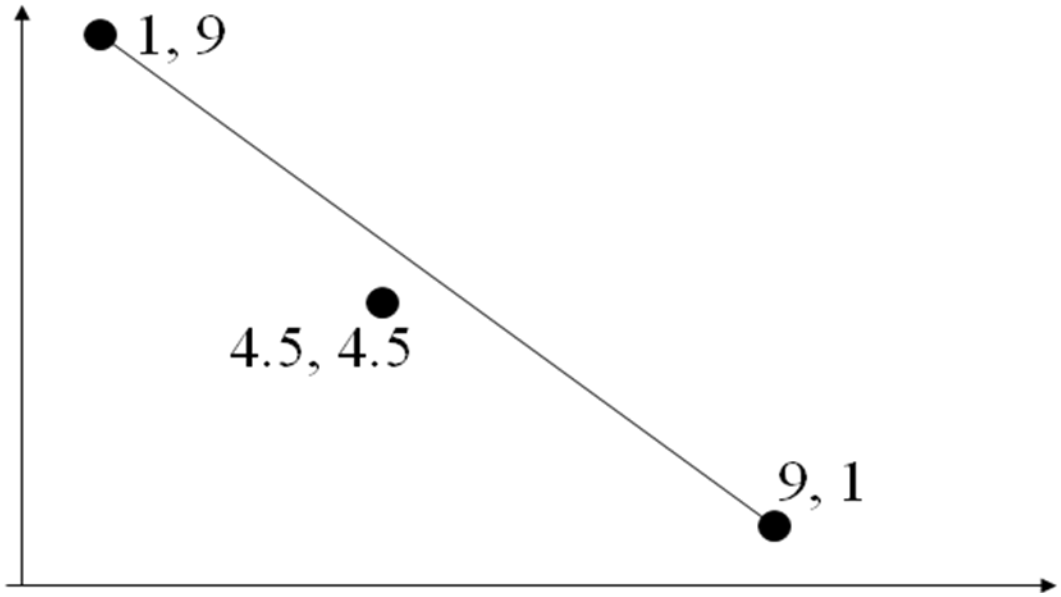
| Candidate | Attribute 1 | Attribute 2 | Attribute 3 | Attribute 4 | Score |
|---|---|---|---|---|---|
| **A** | **2** | **7** | **180** | **100** | **3.275** |
| **B** | **3** | **8** | **140** | **98** | **3.358** |
| **C** | **5** | **5** | **155** | **100** | **3.486** |
| ***Maximum*** | ***5*** | ***8*** | ***180*** | ***100*** | |

The fact that a short-list was produced to aid the selection should not affect the final decision, and yet it has. In this case this effect is explained by noticing that the form of the normalization is data-dependent. Each of the four normalizations described above involves dividing the data by a number which itself is derived from the data. Hence when some of the data are removed this number is altered and leads to different results. Conversely, rank reversal can also arise when a short list has new candidates added to it.

### 3.3 All-rounders overlooked in favour of unbalanced candidates

The next difficulty associated with additive scoring is present even when normalization is not carried out. Consider the three candidates in Figure 1, think of a manager who is choosing a personal assistant based on the criteria of skill and personality as plotted on the two axes. Two of the candidates combine a very high score on one attribute with a very low score on the other.

Figure 1 Three candidates with scores on two attributes. The operation of a weighted sum scoring system can be viewed as moving a straight line towards the origin until it just touches at least one candidate point; points on the line will have the highest score.




We can gain some visual insight into how additive weighting works if we consider that the score formula, being linear, implies that contours for different scores will appear as parallel straight lines with the higher score contours located further from the origin. The top candidate(s) will always appear on a straight line which can be viewed as a 'frontier', with inferior candidates lying behind this frontier. Figure 1 shows a frontier corresponding to equal weights, which is why (1, 9) and (9, 1) are equally valued and appear on the frontier. If more weight were attached to the attribute on the vertical axis then the frontier line would be more horizontal and would only make contact with the upper point. Conversely if the attribute on the horizontal axis had the greater weight, then the frontier line would be closer to vertical and the lowest point in the figure would be selected. Now the important point is that whatever weights are attached the intermediate point

will never be chosen! Forcing the line to pass through that point would mean that at least one of the extreme points would lie above the line (which therefore does not qualify as a frontier), indicating that the extreme point had a higher score. Technically, the intermediate point is said to be 'convex-dominated' by the other two, but is not strictly dominated by either one of them. What this teaches us is that is that a simple additive value function can dismiss candidates which may be described as 'reasonable all-rounders', even if these may be what the decision maker would have preferred. Zeleny (1982) calls this the 'linearity trap'. Of course, in this illustrative example involving just two criteria and three candidates, the decision maker would have no problem making the choice without the assistance of any method, but the point is that such a method would be used as an aid for larger filtering or selection problems, where one cannot hold all the relevant information in one's head to give a gestalt or holistic decision. Wierzbicki (2006) refers to this effect as the Korhonen paradox, due to an example given by Pekka Korhonen involving the (hypothetical) choice of a life-partner. He uses this to show that a weighted sum 'tends to promote decisions with unbalanced criteria; in order to obtain a balanced solution, we have either to use additional constraints or a nonlinear aggregation scheme'. He argues that 'human preferences have essentially nonlinear character, including a preference for balanced solutions, and that any linear approximation of preferences e.g. by a weighted sum, distorts them...This is in opposition to the methods taught in most management schools'. Wierzbicki (2006) goes on to describe the real life introduction of a public tender law in Poland requiring greater transparency in regard to the weights used in aggregating criteria: 'Organizers of the tenders soon discovered that they were forced to select the offer that is cheapest and worst in quality, or the best in quality but most expensive'.

## 4. A more flexible score function

Simple additive weighting implies that there is a *fixed* trade-off rate between each pair of criteria. Moreover this trade-off, or 'exchange rate', is assumed to remain the same irrespective of the level of the attributes. For example, if you are willing to work for an employer for $50 per hour then you would be willing to do this whether you worked eight hours per day or 20 hours per day. Of course this is not realistic, the employee would expect a greater hourly return for working 20 hours per day. This illustrates the fact that human preference (score) functions cannot be assumed

to be linear. Another clear example is to consider the marginal utility of repeatedly giving someone $100 bills. The first such bill will be considered more personally valuable than the thousandth one. This is an instance of the diminishing value of marginal returns – something which simple additive weighting does not take into account. (An informal illustration of this was given by the actor and Governor of California, Arnold Schwarzenegger: "I have 50 million dollars, but I was just as happy when I had 48 million". At the other end of the scale, when a king has zero horses and needs one to escape he may be heard to exclaim 'A horse! A horse! My kingdom for a horse'.)

Since human preferences are not always well described by a simple additive value function, we move to explore a nonlinear form. In an effort to keep things simple, let us consider the multiplicative formula, so that for two attributes the score would be given by the product XY.

Figure 2 A contour or indifference curve for a multiplicative score function (XY = constant). Each point on the curve represents combinations of attribute or criteria values which lead to the same overall score.

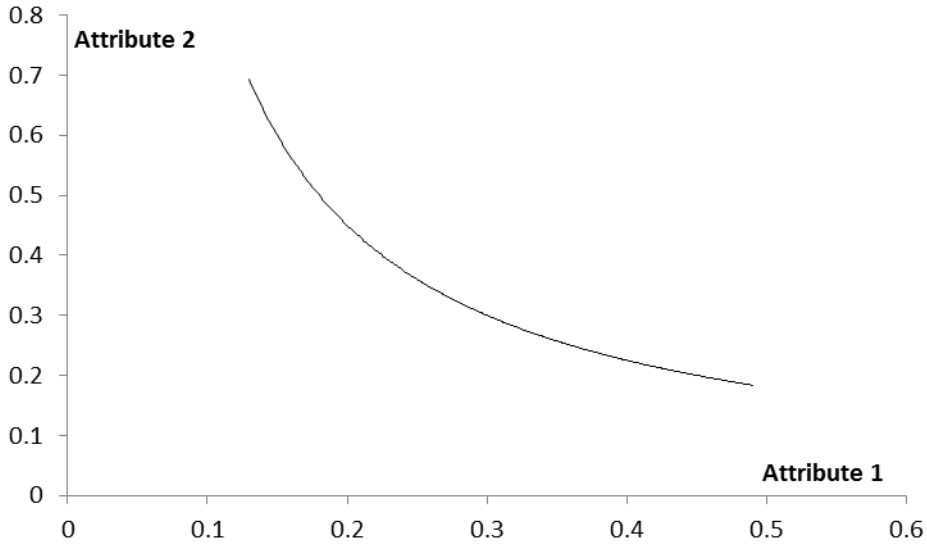



Figure 2 shows what a contour or frontier would look for the multiplicative score function. Each point on this curve has equal value or utility to the decision maker and so we can refer to it as an indifference curve since the decision maker is indifferent between candidates represented by points on this curve. The fact that it is not straight implies that we longer have a constant trade-off or exchange rate between the two attributes. Instead we have an exchange rate that depends on the levels of the attributes. Thus as we achieve high values on one attribute (say job salary), one further unit of that attribute is worth fewer units of the other attribute (holiday days off per

year). Notice that superimposing such a curve on Figure 1 could permit selection of the reasonable all-rounder candidate, whereas this was ruled out by the linear form. This is not say that the multiplicative form necessarily excludes selection of one of the candidates at the extremes. For if the decision maker places a high weight on a particular attribute, this too can be achieved. Note that weights now appear as powers, so for example to emphasise criterion Y, the score function $X\ Y^3$ would lead to the selection of the point at the upper left of Figure 1. Thus we see that the multiplicative score function is flexible enough to represent a wider range of views without excluding those which might have been selected previously using simple additive weighting.

## 5. Measurement Theory

The field of *measurement theory* is one which is not well known but does provide useful insights to the type of question we have been considering. One of its concerns is whether the truth of a statement remains unchanged when the measurement units are changed. For each measurement scale there are unit conversions which are said to be *admissible transformations*. For example, ratio scale measurements can be defined to be those which allow transformations consisting of multiplication by a positive constant (e.g. converting minutes to hours, or cents to dollars). Ratio scales (e.g. mass, length, time, money) require an absolute zero which signifies the absence of the quantity being measured; this allows one to sensibly make comparisons based on ratios of measurements (e.g. "this costs twice as much as that"). By contrast, interval scales can be defined as those which allow positive linear transformations (i.e. multiplication by a positive constant followed by addition of a constant), for example Celsius to Fahrenheit conversions. In measurement theory a statement is said to be *meaningful* if its truth is not affected when an admissible transformation has been carried out. When working with measurements on an interval scale it is meaningful to compare *differences* (but not ratios) between measurements, for example we can say twenty degrees Celsius is ten degrees more than 10 degrees Celsius, but we cannot sensibly say that one temperature is twice as hot as the other (the falsity of which becomes apparent when one compares the equivalent Fahrenheit temperatures).

A splendid introduction to measurement theory is provided by Roberts (2009). He begins by stating the problem very clearly:

> ‘In almost every practical use of operations research, something is measured. Yet in many cases little attention is paid to the limitations that the scales of measurement being used might place on the conclusions which can be drawn from them. Scales may to some degree be arbitrary, involving choices about zero points or units or the like. It would be unwise to make decisions which could turn out differently if the arbitrary choice or zero point or unit is changed’.

He mentions the common situation where a group of experts or judges are scoring a number of alternatives or candidates. The usual approach is to add together their scores to find a winner. However, it is likely that some judges are stricter than others – for example we all remember teachers who did not believe in giving 100% scores, whereas others did. They are essentially using different measurement units, and so adding them together is not appropriate. However multiplying the experts’ scores is a safer approach. (In an ideal world all the experts would agree on the relative rankings and so there would be no issue to deal with, but more commonly there are disagreements and so the way the scores are combined does matter.) Of course one could be more sophisticated and find a way to make the experts’ scores comparable before adding them together, but the approach of simple multiplication avoids this complication and is more straightforward.

Further cases where additive aggregation leads to meaningless results are given by Fleming and Wallace (1986). They provide stark examples involving the performance of computers on a variety of benchmarks where additive approaches give results that do not make sense. They demonstrate that the correct approach is to use the geometric mean of these performance measures i.e. a multiplicative approach.

For the purposes of our discussion a key contribution of measurement theory is in the permitted transformations on data in a given measurement scale. Starting with data on a ratio scale i.e. a scale with a non-arbitrary absolute zero (e.g. years of experience), if we apply a normalization of type 2 or 3 (range normalization or standard deviations) we shall alter the zero point. In range normalization the zero corresponds to the lowest observed value in the sample, whilst in standard deviation units the zero corresponds to the mean. This shift in the zero point results in a loss of the ratio scale properties; the resulting data have been converted to an interval scale. More

generally, all four types of normalization involve division by a number derived from the data. This causes the measurement units to be altered.

Since all four normalizations are data-dependent it follows that removal or insertion of data arising from the rejection or inclusion of particular candidates can affect the results. In conclusion we see that employing normalizations has potential pitfalls for the unwary.

## 6. Choosing between an additive and a multiplicative score function

It is worth noting that a useful way of choosing between an additive score function and a multiplicative one, is to consider the consequences of employing the straight indifference line of Figure 1 versus the curve of Figure 2. The straight line will, if extended, intersect the axes. This means that one is willing to keep giving up or trading off units of one attribute in exchange for some units of the other attribute, at a given *fixed* exchange rate, even to the point where one has zero of the first attribute. If this is not acceptable to the decision maker then the additive score function is not appropriate. Likewise, if one does not feel the exchange rate should stay the same however high or low the attributes levels, then again an additive score function is not suitable. However, one can argue that such linear functions might be useable as approximations when the attribute ranges involved are narrow. Let us therefore proceed with some practical guidance. Since this paper is aimed at those who are primarily concerned with using a method based on a simple formula, we restrict attention to the three classes: weighted average, weighted multiplicative, and 'other'.

Start by selecting the criterion with which you are most comfortable or familiar, for example cost in dollars, and note the lowest value and highest value of this criterion among the available candidates. Observe the lowest and highest available value of one of the other criteria. Then ask yourself this question: How many units of the first criterion am I willing to give up in order to improve the second criterion by one unit? Next ask yourself: 'Am I willing to apply this exchange rate in both directions (worsening and improving) until I reach either end of the observed range of these attributes?' If the answer is YES, and this positive response also applies when the second criterion is replaced by the third, the fourth, etc. then the additive approach is

appropriate because there are fixed exchange rates between every pair of variables across the whole range of the observed values of the criteria. Note that it's not necessary to consider every possible pairing of variables; if every attribute has its own fixed exchange rate with the first attribute, then logically it follows that there is a fixed exchange rate between every pair.

Now suppose instead that the above procedure leads to the conclusion that simple additive weighting is not appropriate. Since we will now be dealing with percentage changes we require a reference point: use the middle point of the range for each criterion and calculate the percentage change required to reach each end of the observed attribute range. Select once more the criterion with which you are most comfortable or familiar. Then ask yourself: what percentage of the first criterion am I willing to give up in order to improve the second criterion by one percent? Next ask yourself: 'Am I willing to apply this percentage substitution rate in both directions (worsening and improving) until I reach either end of the observed range of these attributes?' If the answer is YES, and this positive response also applies when the second criterion is replaced by the third, the fourth, etc. then the multiplicative approach is appropriate.

In either of the above cases the decision maker may be fortunate enough to have an analyst or facilitator to assist them; alternatively they may be using a decision support system. In such cases there will be checks on consistency regarding the weights that have been selected. The analyst will be able to deduce the implied exchange rate or substitution rate between pairs of criteria that have not been directly compared by the decision maker. If the decision maker is not comfortable with these implied rates then adjustments will have to be made. The elicitation of weights is often an iterative process in which the decision maker learns about the implications of their choices. For example it may be clear in the decision maker's mind that candidate 1 is definitely superior to candidate 2 yet the weights they have supplied lead to the opposite ranking. The analyst could assist in uncovering the source of the inconsistency. At this stage it may even become apparent that certain relevant criteria have been overlooked and accidentally excluded from the analysis.

It may be that neither the additive nor the multiplicative procedure above leads to a positive response. This then implies that these simple approaches are not strictly applicable. It is beyond the scope of this paper to deal with these more complicated scenarios.

# 7. Properties of the multiplicative score function

Let us consider a scoring function where all criteria of the 'more is better' type are multiplied together, and those where 'more is worse' are divided into the former.

A simple example would be the function $X_1\ X_2\ /X_3$ where $X_3$ is a 'more is worse' attribute such as cost. Here, each weight (exponent or power) is equal to unity, but this need not be the case. More generally we would have a multiplicative score function with unequal weights:

Multiplicative score function: $[X_1^{W1}\ X_2^{W2} \ldots] / [X_3^{W3} \ldots]$

### Re-scaling has no effect on the outcome

If we change the units of measurement such that an attribute's values are all multiplied or divided by some positive constant, then the rankings are unchanged. This is because this conversion is equivalent to multiplying the overall score of all the candidates by the same factor. Moreover this remains true even when that attribute has a weight attached to it.

The normalizations of type 1 or type 4 correspond to a re-scaling (i.e. a change of measurement units) and so would also have no effect on the result.

### The weights are interpreted in percentage terms and can allow for diminishing returns

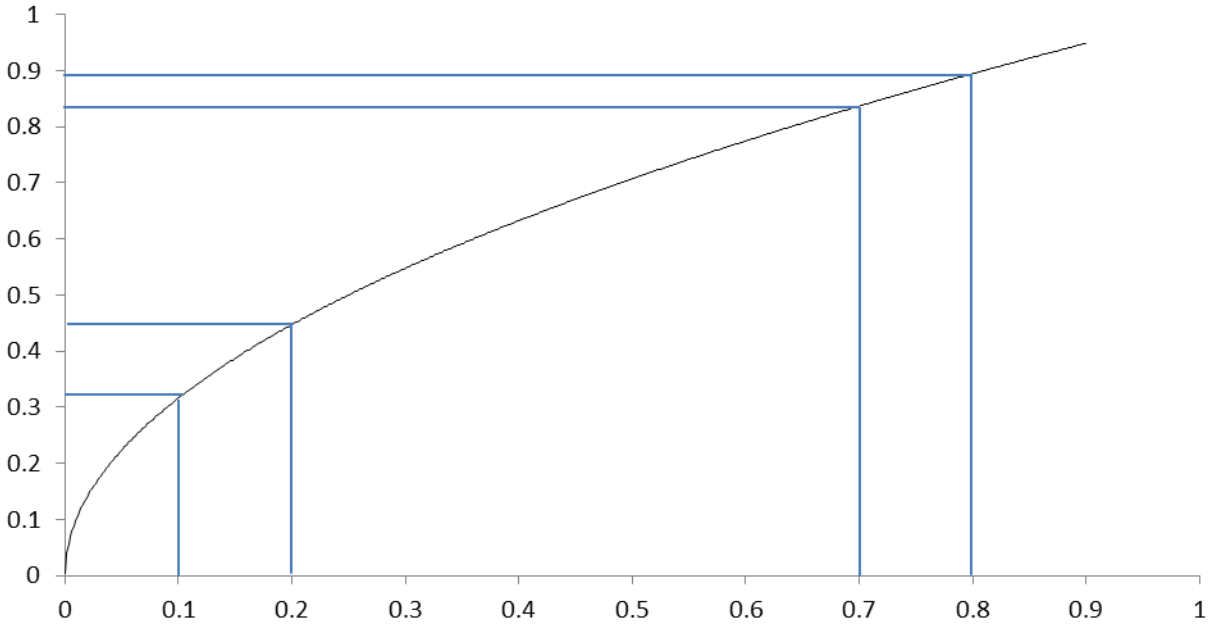
If the score on an attribute is given a weight (exponent) of *w* this means that a 1% change in the attribute gives a w% change to the overall score. See Appendix 1 for a derivation. It follows that trade-off rates between criteria will be in terms of percentages. Thus, in eliciting such weights from a decision maker the question put to them might be: 'what percentage improvement in criterion 1 would you require to compensate for a 1% worsening in criterion 2?'.

Alternatively, one can elicit weights for each attribute separately by referring to the overall score: 'if this attribute is improved by 1%, by what percentage should the overall score rise?' If we assume decreasing marginal returns (see below) the answer to this question must be less than 1%. Things can be made easier by referring to basis points, as used in finance: 100 basis points =1%. If the sum of weights is to equal unity then the decision maker can be asked to split up or share out the 100 basis points among the attributes, thus providing the required individual weights. It is

recommended to carry out a check by inspecting these weights to see what the associated trade-off rates would be between pairs of attributes, to ensure these are agreeable to the decision maker. For desirable attributes there is usually a gradual satiation effect as we acquire more: an extra dollar added to a million is felt to be less significant to the owner than when that dollar is added to ten dollars. This is a case of diminishing marginal returns. This psychological effect can easily be represented in our system by requiring the weight *w* to be numerically less than one. This provides a concave score function, see Figure 3.

If the decision maker is uncomfortable in choosing weights in the above way, another approach is to take two or more observed values of a particular attribute and ask the decision maker what the perceived value or benefit is to them for each of these. For example the question might be phrased as 'If your perceived benefit or score for the best value is 100, what score would you give to these other values?' The responses can be used to deduce what the corresponding weight (power) is, see Appendix 2.

Figure 3. Diminishing marginal returns: the perceived benefit (vertical axis) of equal increases in an attribute (x) decline the more one has of that attribute. This effect can be represented by using the power function $y = x^w$ with the power *w* between 0 and 1.



Lootsma (1996) notes that psycho-physical research shows that humans are sensitive to relative change rather than absolute change: 'human beings generally perceive relative gains and losses, that is, gains and losses in relation to the levels from which the move starts'. Suppose a marginal change in a stimulus leads to a just-noticeable change in response. If the latter is proportional to the response level itself then integration leads to a response that is a power function of the stimulus. In our terms, this corresponds to a score function that is a power function of the attribute level. Our multiplicative score function is precisely composed of such power functions.

### The weights do not depend on the units of measurement of the criteria

This follows from the fact that weights are interpreted in percentage terms. This is an important difference between additive and multiplicative scoring. One implication is that one can now speak of equally weighted criteria without having to state additional details such as measurement units and normalization scheme. This is a significant advantage because the non-technical decision maker may use the term 'equally weighted' without realising that under additive scoring this is not so well-defined.

### Requirement of a ratio scale

For re-scaling to be permissible requires that the attributes are measured on a ratio scale. Note that if a candidate has a zero score on any criterion then it automatically has a zero score on the multiplicative score function. In practice the presence of a zero score would likely lead to that candidate not being considered in any case, so this in itself is not a great limitation. (The distinguishing feature of a ratio scale is the possession of a non-arbitrary zero value.) In the context of aggregating scores by members of a panel, a zero score by one member would act as a veto. Thus, if such vetoes are not to be permitted then zero scoring would have to be ruled out.

Sometimes an attribute is not a quantity which can be objectively measured, but instead is a subjective expression of preference. One way of eliciting values in such a case is to ask the decision maker to select the candidate which is the most preferred on the given attribute and to

attach a score of 100 to the attribute in question. One then explains that a score of 50 would be considered half as good, and that a score of zero implies that the attribute is not present at all. The decision maker is then asked to rate the remaining candidates for that attribute. This is called *magnitude estimation*. In such a scenario it would be sensible to assume a ratio scale is appropriate for that attribute.

## 8. Aggregating the ratings of a panel

The discussion in section 3.3 also applies to the situation where a panel is selecting from a number of candidates. In this case the weights would usually be equal and so the situation could be represented as in Figure 1 where the line is at 45 degrees. The coordinates for a given point (candidate) correspond to the ratings of two judges of that person. Figure 1 would then show that the candidate on the left is rated extremely highly by one judge and very poorly by the other, whereas the candidate/point on the right has received the opposite ratings from the two judges. Under an additive scheme where the judges' ratings are added together the person in the middle would not be chosen, but that person would be selected if their ratings were multiplied. If there were three members on the panel the chart would have to be imagined in three dimensions, where the three coordinates of a point now refer to the ratings of the three panel members etc.

There is another issue arising with the addition of scores of panel members – they may have different standards. In other words they may, in their minds, be using different "units of goodness". Even if they are told to provide a score out of ten, each decision maker will hold a different image of what a 10-score candidate will be like – based on their personal expectations and past experience. The multiplication-based geometric mean overcomes this problem of internalised personal units and standards. Consider the case of a judge who never gives a score exceeding 9 ('because that would imply perfection'), whilst the other judges do give perfect-10 scores. He or she is effectively using a different unit of measurement and, as we have previously observed, the unit of measurement makes no difference to the outcome under multiplicative aggregation. If that person's score is divided by 9, then the units would become the same as for the other judges; but this is equivalent to dividing all the aggregate (multiplicative) scores by 9 and so it makes no difference to the final ranking, and hence this rescaling is unnecessary. Thus

we say that the multiplicative result is invariant to the choice of units, whereas the additive result is affected.

The disadvantages of the arithmetic mean for aggregation can thus be overcome by using the geometric mean i.e. multiplying. We shall not go into the mathematical foundations which have been laid down in measurement theory, but will simply draw the attention of the reader to the important, but little-known, work that has been carried out, by quoting (Roberts, 2009):

> "Suppose different experts are asked to give their judgements of the relative value or importance or potential of different alternatives, e.g. alternative strategies, alternative new technologies, alternative diagnoses, etc. Suppose the choice will depend on the 'average' of the experts' judgements … If all the experts use a ratio scale, i.e. if only their units change, then which alternative has the highest *arithmetic mean* depends on the choice of these units (if they can be chosen independently). However, the alternative with the highest *geometric mean* is invariant. Thus it is 'safer' to use the geometric mean ... Aczel and Roberts (1989) have put this on a firm mathematical foundation by showing that under certain reasonable assumptions about the averaging or merging procedure, it is not only the case that the arithmetic mean is unacceptable and that the geometric mean is acceptable, but that the geometric mean is the *only* averaging procedure which is acceptable."

A detailed axiomatic justification demonstrating that this is the correct way to aggregate ratio-scaled evaluations from multiple decision makers is provided by Aczel and Saaty (1983).

## 9. Illustration using Business School Rankings

We have used data on British Business Schools compiled by Mayfield University Consultants for 'The Complete University Guide 2013' (CUG) and which is freely available online (Mayfield Consultants, 2012). The data consists of four criteria: Student satisfaction, as measured by the National Student Survey; Entry standards, measured by the number of A-Level points (national qualifications obtained at the end of secondary school); research assessment, measured by the

most recent UK Research Assessment Exercise; and Graduate prospects, measured by the percentage of graduates who take up employment or further study within six months of graduating.

The CUG league table uses z-scores (statistical standardization) to normalise the four component criteria. It then applies equal weighting to the z-scores.

On attempting to reproduce the CUG ranking using their stated methodology we found the ordering was different. For example the London School of Economics was ranked in second place in CUG, whereas our calculations placed it seventh. CUG ranked St. Andrews fifth, whereas we placed it in second position. We therefore contacted the compilers of the table for an explanation. They explained that the range of the student satisfaction scores was so compressed that very small changes in that variable would have large changes in the impact on the table, they therefore decided to reduce the weight on this variable. Student satisfaction was measured on a five-point scale, but the observed range was much narrower: from 3.1 to 4.3. To make matters worse, the lowest observed value of 3.1 was four standard deviations below the mean, and so was an outlier. The remaining observations were between 3.4 and 4.3.

Histograms showed that the student satisfaction variable was the only one that looked normally distributed; the other three variables appeared to be bimodal, with one being left-skewed and another right-skewed.

A separate but unrelated issue: there was missing data for some criteria for some institutions; CUG dealt with this by averaging whichever components were available. In what follows we shall not adopt this practice; instead, we shall omit institutions with missing data from our discussion, so that institutions below these are moved further up the table. This left 83 business schools for consideration, shown in Table 8.

Table 8 Comparison of rankings for the subset of business schools with no missing data: multiplicative aggregation and Complete University Guide (CUG).

| Institution | Multiplicative Rank | CUG Rank | Difference in Rank |
|---|---|---|---|
| London School of Economics | 1 | 1 | 0 |
| Bath | 2 | 2 | 0 |
| Warwick | 3 | 3 | 0 |
| St Andrews | 4 | 4 | 0 |
| Exeter | 5 | 5 | 0 |
| Loughborough | 6 | 6 | 0 |
| King's College London | 7 | 7 | 0 |
| Lancaster | 8 | 8 | 0 |
| Leeds | 10 | 9 | 1 |
| Cardiff | 9 | 10 | 1 |
| Nottingham | 11 | 11 | 0 |
| Strathclyde | 15 | 12 | 3 |
| Durham | 14 | 13 | 1 |
| City | 13 | 14 | 1 |
| Aston | 12 | 15 | 3 |
| Sheffield | 18 | 16 | 2 |
| Manchester | 16 | 17 | 1 |
| Newcastle | 19 | 18 | 1 |
| Reading | 21 | 19 | 2 |
| Birmingham | 17 | 20 | 3 |
| Southampton | 20 | 21 | 1 |
| Glasgow | 22 | 22 | 0 |
| Surrey | 23 | 23 | 0 |
| Edinburgh | 24 | 24 | 0 |
| York | 25 | 25 | 0 |
| Heriot-Watt | 26 | 26 | 0 |
| Leicester | 27 | 27 | 0 |
| Liverpool | 28 | 28 | 0 |
| Sussex | 29 | 29 | 0 |
| Brunel | 33 | 30 | 3 |

We applied the equal weights multiplicative approach to these schools and compared the results with CUG's published ranking of the same subset. For reasons of space we do not show the full list but results are displayed in Table 8 for the top 30 according to the CUG approach, together with results from the multiplicative method in column 1. One observes that many institutions have identical rank: 17 out of 30. The mean absolute difference in rank across the 83 institutions was less than one position!

In summary, the results from the two methods are very similar. However, in one approach (CUG) one had to calculate means and standard deviations for all measures, then normalize each of the measures prior to adding them together, then make an adjustment to one of the weights because one of the variables was too influential. Compare this with the other approach where one simply multiplies the raw data together! Many readers of league tables have no knowledge of standard deviation or normalization. Such manipulations or transformations lead to a lack of transparency, and as we have seen, can easily be avoided.

Thus, simply multiplying the components has multiple advantages: it is simple to perform, easy to understand by the public, does not involve arbitrary steps, and is objective. More details on using the multiplicative approach for university ranking can be found in Tofallis (2012).

## 10. Conclusion

Real selection, ranking, and short-listing problems usually involve the consideration of multiple criteria. Given a situation where a choice (or ranking) has to be made from a number of options or candidates, a widely used approach is simple additive weighting. It may come as a surprise to practitioners who use the apparently transparent method of simple additive weighting that it is often not used in a reliable way. The pitfall lies in the normalization step: because there is no universal agreement on how this step should be carried out, a change in the way the data is normalized can lead to different rankings and hence decision outcomes.

A second unexpected aspect of the common scoring approach arises when clearly inadequate candidates are removed from a long list to produce a shortlist. If the same scoring technique is re-applied to the shortlist, a reversal of ranks can occur among the remaining candidates, leading to a change in the final decision. In other words, the 'best in the shortlist may not be the best in the

whole list' when simple additive weighting is used. This problem arises because the normalizations are dependent on the data.

Multiplicative scoring avoids these problems. It is already being used in the real world: The United Nations Development Programme (UNDP) publishes an annual ranking of nations known as the Human Development Index (http://hdr.undp.org), which is very influential and is used by first world nations to guide their aid allocations. It is also used by pharmaceutical companies to decide which countries should receive discounted prices. This index is an aggregate of three criteria: life expectancy, education, and gross national income per capita. For many years the aggregation was carried out using additive weighting. This was repeatedly criticised (e.g. by Sagar and Najan, 1998) because this assumed that the criteria were perfectly substitutable i.e. constant trade-off rates. Consequently the UNDP chose to change their methodology (UNDP, 2010), and the index is now calculated using a multiplicative scheme. The adoption of such a scheme by an internationally recognised organization will hopefully help to publicise an alternative to additive scoring. The choice of weights to be attached to criteria is always a difficult issue; Tofallis (2013) presents an 'automatic-democratic' method for generating weights in a non-subjective way.

Kondraske's General Systems Performance Theory also calls for the adoption of multiplicative performance measures. Kondraske (2011) includes among its examples the problem of estimating the value of a diamond, which is generally agreed to depend on the four Cs: carat (weight), cut, colour, and clarity. He took data on 257 diamond prices and their attributes and compared the predictive performance of additive aggregation of the four Cs, and a multiplicative one. The goodness of fit (R-squared) for the multiplicative measure was 0.85, whereas for the additive model it was only 0.28. The multiplicative performance measure "reflects that price will be low if any one of the four factors is poor and that a highly priced (high quality) diamond requires high ratings for all four factors". Kondraske and Stewart (2006) show that when it comes to measuring disease severity, composites based on multiplication are again superior.

This paper was motivated by the fact that certain issues associated with additive aggregation can be overcome by using a multiplicative approach. This approach also brings the benefit of

modelling human preferences more realistically than simple additive weighting because it allows diminishing returns to be represented. We hope that we have shown that by using multiplication we can overcome problems arising from making multi-criteria short-listing or selection decisions based on addition.

# APPENDIX 1

## Deducing the effect of weights in the multiplicative score function

We shall show how to calculate the effect of a change in the value of an attribute on the multiplicative score function. They key difference to using an additive score function for this analysis is that it's more convenient to think in terms of percentage changes. Consider the multiplicative score function $S = X_1^{W1} X_2^{W2} X_3^{W3} ...$

Suppose attribute 1 is increased by 1%, then the new value of the score function is

$S_1 = (1.01\ X_1)^{W1} (X_2)^{W2} (X_3)^{W3} ..... = 1.01^{W1} S$

So the effect is that the score is simply multiplied by $1.01^{W1}$.

This is approximately a change of $W_1$%. To see this, consider the power expansion:

$(1 + 0.01)^{W1} = 1 + 0.01W_1 + 0.00005\ W_1 (W_1\text{-}1) + .....$

The third term is negligible, and so the effect is that the score is increased by $0.01W_1$ , in other words by $W_1$%.

To see the effect of large percentage changes it is best not to rely on the approximation, especially as the exact calculation is so easy to accomplish: A p% change in an attribute leads to a new score of

$S_P = [(1 + P/100)\ X_1]^{W1} (X_2)^{W2} (X_3)^{W3} ..... = (1 + P/100)^{W1} S$

[The power expansion of this is:

$S_P = 1 + W_1 P/100 + 0.00005\ W_1 (W_1\text{-}1) P^2 + .....$

Which shows that for larger percentage changes $P$, and weights $W$, the third term may make a non-negligible contribution.]

# APPENDIX 2

## Deducing the appropriate weight in a power function

This appendix deals with estimating the value function or relationship between the perceived benefit ($Y$) and the values of an attribute ($X$). Let us begin with the simple case where the decision maker compares two actual values of an attribute and allocates a score of $Y_{max} = 100$ for the best value $X_{max}$ and a score of $Y_1$ for another value $X_1$.

We are assuming that the value function (Y) is proportional to the attribute value raised to a power ($X^w$). Hence $Y = cX^w$ and we want to deduce the value of the weight $w$. Taking ratios of the two observations cancels out the constant:

$Y_1/100 = X_1^w / (X_{max})^w$

and taking logs:

$\log (Y_1/100) = w \log[X_1 / X_{max}]$

hence $w = \log (Y_1/100) / \log[X_1 / X_{max}]$

If we are presented with 3 or more ratings by the decision maker then the weight can be estimated using regression. This can be done very quickly in Microsoft Excel: plot the points on a scatter graph, right-click any point and select "Add Trendline". Choose the power function and check the option to display the equation.

# REFERENCES


Aczel, J and Roberts, FS (1989). On the possible merging functions. *Mathematical Social Sci*, 17, 205-243.

Aczel, J and Saaty, TL (1983). Procedures for synthesizing ratio judgements. *J Mathematical Psychology*, 27(1), 93-102.

Fleming, PJ and Wallace, JJ (1986). How not to lie with statistics: The correct way to summarize benchmark results. *Comms. ACM* , 29, 218-221.

Hillier, FS and Lieberman, GJ (2010). *Introduction to operations research*, 9th edition. McGraw Hill, Boston.

Kondraske, GV (2011). General Systems Performance Theory and its application to understanding complex system performance, *Information Knowledge Systems Management***,** 10 (1-4): 235-259.

Kondraske, GV. and Stewart, R.M. (2006). Quantitative characterization of disease severity in diseases with complex symptom profiles. Proceedings of the 28th Int. Conf. of the IEEE Engineering in Medicine and Biology Society, New York.

Lootsma, FA (1996). A model for the relative importance of the criteria in the multiplicative AHP and SMART. *EJOR*, 94, 467-476.

Mayfield University Consultants (2012). Complete University Guide – Business http://www.thecompleteuniversityguide.co.uk/league-tables/rankings?s=Business+Studies

Pavlicic, DM. (2000). Normalization of attribute values in MADM violates the conditions of consistent choice IV, and DI and α. *Yugoslav Journal of Operations Research*, 10(1), 109-122.

Pomerol, J-C and Barba-Romero, S. (2000). *Multicriterion Decision in Management-Principles and Practice*. Kluwer Academic.

Powell, SG, and Baker, KR. (2011). *Management Science: The Art of Modeling with Spreadsheets,* 3rd Edition. Wiley, NY.

Roberts, FS (2009). Limitations on conclusions using scales of measurement. In *Operations Research and the Public Sector*, Handbooks in Operations Research and Management Science, volume 6, Eds. SM Pollock et al. Elsevier Science.

Sagar, AD and Najam, A. (1998). The human development index: a critical review. *Ecological Economics*, 25, 249-264.

Tofallis, C. (2012). A Different Approach to University Rankings. *Higher Education*, 2012, 63(1), 1-18.

Tofallis, C (2013). An automatic-democratic approach to weight setting for the new human development index. *J Population Economics*, 26(4), 1325-1345.

UNDP (United Nations Development Programme). (2010). *Human Development Report 2010: The Real Wealth of Nations.* UN Development Programme, New York: Palgrave Macmillan.

UNDP (United Nations Development Programme). (2011). *Human Development Report 2011*. UN Development Programme.

Wierzbicki, A P (2006). Reference point approaches and objective ranking. Dagstuhl Seminar Proceedings. Downloadable from http://drops.dagstuhl.de/opus/frontdoor.php?source_opus=1121

Yoon, KP and Hwang, C-L (1995). *Multiple attribute decision making*. Sage Publications, London.

Zeleny, M (1982). *Multiple Criteria Decision Making*. McGraw-Hill, New York.