A Portfolio Model for Performance Assessment: the *Financial Times* MBA ranking.

Alan Jessop
Durham Business School
Mill Hill Lane
Durham
DH1 3LB
UK

tel: (0)191 334 5403

email: a.t.jessop@durham.ac.uk

Abstract

The ranking of MBA programmes by newspapers and magazines is common and usually contoversial. This paper discusses the use of the most popular method of making these rankings via a multicriteria model which uses the weighted sum of a number of performance measures to give an overall score on which selection or ranking may be based. The weights are a quantitative model of the preferences of those making the evaluation. Many methods are available to obtain weights from preference statements so that for any set of preferences a number of different weight sets can be found depending on the method used. Cognitive limits lead to inconsistency in preference judgements so that weights may be subject both to uncertainty and to bias. It is proposed that choosing weights to minimise discrimination between alternatives (not weights) guards against unjustified discrimination between alternatives.

Applying the method to data collected by the *Financial Times* shows the effect of varying the level of discrimination between weights and also the effect of using a reduced data set made necessary by the partial publication of information.

Keywords: multicriteria; quadratic programming; performance; MBA.

A Portfolio Model for Performance Assessment: the *Financial Times* MBA ranking.

Introduction

Measures of the performance of organisations are commonly used either to initiate some improving changes in the organisation or to provide a guide for those wishing to buy the goods or services produced by it. These evaluations are often presented, however unwisely, as rankings or league tables. This is a particular example of the general multicriteria problem of selecting a number of alternatives depending on their attributes. A commonly used method is to find for each alternative the weighted sum of a number of performance measures or attributes:

$$y_i = \sum_j w_j u_j(x_{ij}) \qquad (1)$$

where y_i is the score of alternative i; $i = 1,2\ldots n$

x_{ij} is the value of attribute j for alternative i; $j = 1,2\ldots m$

$u_j(.)$ is a value function chosen so that high values are preferred

$w_j \geq 0$ is the weight reflecting the importance of or preference for attribute j

and, by convention,

$$\sum_j w_j = 1 \qquad (2)$$

where weights are chosen to encode the judgemental preferences of decision makers.

In what follows the function chosen for $u_j(.)$ is the familiar z transformation

$$u_j(x_{ij}) = z_{ij} = (x_{ij} - \mu_j)/\sigma_j \qquad (3)$$

where μ_j and σ_j are the mean and standard deviation of the values of X for attribute j. This form of value function may be preferred to simply rescaling data for each attribute to the range [0,1] because it is less sensitive to changes in the set of alternatives.

In principle neither the data, X, nor the weights, W, are known precisely – the first because of errors of sampling and measurement and the second because of the inescapable imprecision imposed by the cognitive limits of those providing preference information. This paper describes a reaction to the second problem; uncertainty and bias in weight evaluations. It is to be understood that bias is not meant pejoratively but rather to indicate systematic error.

Given the cognitive and other difficulties which necessarily attend the elicitation of preference statements and which may lead to biased estimates (Borcherding, Schmeer and Weber 1995; Tversky and Kahneman, 1974) it is customary to use some form of sensitivity testing to see if plausible variations in weight values might materially alter the result (see, for example, Barron and Schmidt, 1988; Mustajoki *et al*, 2006). Although these analyses provide an effective means of engagement of the user with the problem they do not reduce the desirability of formally incorporating into the structure of the evaluation model measures which reduce the effects of uncertainty and bias in the derivation of weights. The issue is one of justification. In particular, given some preference judgements the method whereby weights are found ought not to introduce any more information than is contained in those judgements. It is proposed that the variance of scores is an appropriate measure of discrimination and that the effect of information is to increase its value.

Most evaluations receive no great publicity but of those that do the rankings which are routinely published of business schools and MBA programmes attract much attention. The data used by one of these, that provided by the *Financial Times*, is used here to demonstrate the derivation of minimally biased evaluations.

The paper is organised as follows: a method for guarding against bias is proposed; following a brief discussion of MBA rankings and some of the issues which surround them the data used in this paper are given; finally, the method is applied and the results discussed.

A portfolio model

In making an evaluation weights are determined based on some preference information. It is usually the case that fitting weights to these judgemental data is a problem with positive degrees of freedom and so there exist no unique weights, only those which are optimal according to some criterion. In multicriteria analyses the motivation is often to maximise discrimination (e.g. Green and Doyle, 1995) by choosing weights which give maximum differences between scores. But is this right?

The all too human desire to differentiate, to be decisive, can result in seeing differences where none are justified. Using methods which guard against this unjustified attribution of differences is therefore desirable. A similar concern arises in deciding probability distributions: that the probabilities are selected, unwittingly or not, in a way which biases the result towards a particular outcome. To guard against the effects of this bias Jaynes (1957) argued that a distribution should be sought which is minimally discriminating between the probabilities of alternative values of a variable. He did this by maximising the entropy of that distribution, $-\sum p_i \ln(p_i)$, to give a distribution which is as flat as is consistent with any stated conditions which are to be respected. There remains the issue of just how such conditions (of mean and variance, for instance) are given: if firmly based on data there are no difficulties but if they are to some degree subjective they may themselves be prone to bias. Jaynes (2003, p373) is quite clear on this; he believes that distributions may be found via the entropy maximising formalism only if conditions are specified as a result of what is known, not what is believed (he was no subjectivist). However, even if conditions are to some degree the result of judgement, the maximum entropy formalism does ensure that, given those conditions, the means whereby probabilities are subsequently derived induces no further bias.

In multicriteria evaluation the task is to find weights rather than probabilities, but the same considerations hold (Jessop, 1999) and maximum entropy methods have been used to find weights by maximising the entropy of the weight distribution (Barron and Schmidt, 1988; Gabbert and Brown, 1989; Soofi, 1990; Soofi and Retzer, 1992) . The conditions imposed may be that it is desirable to have weights which are close to some initial distribution or that weights should stay within some stated bounds or both.

The entropy of weights, $-\sum w_i \ln(w_i)$, is not the only measure of the flatness of a distribution, the variance,

$$\sigma_w^2 = \left(\sum_i w_i^2 - 1/m \right) / m \qquad (4)$$

being an obvious alternative. The sum $\sum w_i^2$, is a related measure used for the analysis of the concentration of market shares (Herfindahl OC, 1950; Hirschman, 1964) and of species diversity (Simpson, 1949). When the mean is unaltered by the distribution of weights, as here because of (3), optimising this index is effectively the same as optimising variance. Theil (1972) discusses some differences between the Herfindahl-Hirschman index and entropy. An example by Jessop (2004) of the application of both to a multicriteria problem shows that the results obtained are very similar (see also Breiman et al, 1998, ch 4). It is therefore reasonable to use variance as a measure of the flatness of a distribution and to minimise variance as a way of ensuring that weights contain no more information than is implicit in whatever constraints are set which encode the judgements of the decision maker. But first it is necessary to decide just what one should be minimally discriminating about. It perhaps seems too obvious to state that we wish to be minimally discriminating between attributes and so to choose a set of weights which reflects this. The flattest (most uniform) set of weights is one for which the variance (4) is minimised subject to (2) and any other constraints which encode statements of preference. With no such extra constraints this gives a uniform distribution of weights ($\sigma_w^2 = 0$).

However, the purpose of the model is to assess the performance of alternatives and not to investigate weight distributions *per se*. Weights are a means to an end: it is the scores, Y, and the discrimination between alternatives which they permit, which are important and so we wish to minimise the effects of unwanted biases in weight assessment inasmuch as they effect these scores. This is done by minimising the variance of the scores. The variance may be expressed to emphasise the underlying structure of the problem as

$$\sigma_y^2 = \sum_i \sum_j w_i w_j \sigma_{ij} \qquad (5)$$

where σ_{ij} is the covariance between the values of $u_j(.)$ for pairs of attributes i and j ($\sigma_{ii} = \sigma_i^2$ is the variance of the attribute i). This is just the model for the estimation of portfolio risk due to Markowitz (1952) in which the weights are proportions invested in different stocks and the covariance matrix describes the relation between the returns from pairs of stocks. In a multicriteria model the covariance describes the degree to which performance against different criteria are correlated. Because of (3) $\sigma_i^2 = 1$ and so (5) may be rewritten as

$$\sigma_y^2 = \sum_i w_i^2 + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_i w_j \sigma_{ij} \qquad (6)$$

It is clear that when minimising σ_y^2 the resulting weights are a compromise between those needed to minimise the variance of the weight distribution and those which best exploit the structure in the data by using the covariances. If there is no structure in the performance matrix, X, the performance measured by one attribute is unrelated to performance measured by any other ($\sigma_{ij} = 0$; $i \neq j$) so that minimising σ_y^2 will be the same as minimising σ_w^2. This is unlikely.

Given the importance which is often attached to published rankings of performance the attribution of differences is a matter of some practical moment, hence the need for well-founded evaluations. The ranking of MBA programmes is a case in point.

The various rankings and their importance both to business schools and their clients are outlined below. While the data gathered in these exercises are of value for the information they provide it is the aggregation of the data into rankings which receives the greater attention, as intended by the newspapers and magazines which publish them. Any such aggregation depends on decisions made about the relative importance of performance measures – weights – so that those schools not ranked as highly as they believe is their due see that the source of the perceived injustice is that the weights describe a set of values not shared by the school. This is inevitable if a single aggregate measure is to be constructed. But it is not the weights themselves but rather the discrimination implied by the rankings which is contentious and this raises two questions: first, why are results published ordinally as ranks (certainly in the example used in this paper) rather than as scores and, second, is the difference in scores between any two schools large enough to justify the conclusion that one is inferior to the other, as different ranks do? About the first question one can only speculate that the motivation lies with the interests of the publisher. A partial response to the second question is that however the results are given it will help in their acceptance by schools and others if the scores are as fair as possible given the nature of the exercise and that part of this fairness is that the method should not unreasonably differentiate between schools. Using the portfolio model to minimise the variance of the scores does this.

MBA rankings

The performance of MBA programmes has been examined from the point of view of their relative strategic positions (Mar-Molinero and Mingers, 2007; Paucar-Caceres and Thorpe, 2005; Segev *et al*, 1999) and their efficiencies as determined by frontier analyses such as DEA (Colbert, Levary and Shaner, 2000). However, it is the published rankings which receive most attention, not least by business schools themselves. While it may be thought that business schools would welcome the publicity the very existence of the rankings induces behaviours which may be unintended and, perhaps, undesirable (Policano, 2005). Although the evaluations are usually aimed just at MBA programmes applicants may use rankings as a proxy for reputation and prestige both of the programme and the school (Engwall, 2007).

Siemens *et al* (2005) cite the findings of Dichev (1999) that there exists some unreliability in rankings which is likely to be due to errors in the data. Peters (2007), in summarising critiques of the ranking methodology, cites response rates from alumni, on which evaluations are based, as low as one percent. Such poor sampling undermines even those claims of data collection advanced in defence of rankings. Fortunately, the data used here are unlikely to be so flawed since the *FT* data are audited by KPMG, the only such data to be audited (Bradshaw, 2007).

These are not trivial matters. For instance, Peters (2007) cites large increases in the number of applicants enjoyed by schools following improvements in their ranking. Other claims for rankings, for instance that they “are another important measure of customer satisfaction” (Thomas, 2007), also rely on defensible data collection and ranking methodology. Recognising the practical importance of published rankings Bickerstaffe and Ridgers (2007) give as one of the benefits that they have resulted in the accumulation of many data on which applicants and recruiters may in part base their decisions. The *FT* website provides the facility for the data to be used in just this sensible way.

These and other current issues concerning business schools and rankings were examined in a recent issue of the Journal of Management Development (volume 26, no. 1, 2007).

It should be borne in mind that any ranking reflects the view of the publisher and that this is articulated mainly in the choice of variables. For example, the MBA ranking provided by *The Economist Intelligence Unit* uses thirteen variables with a much greater emphasis on career support, networking opportunities and the like than the *FT*. The methodology is the same for both but the results quite different. It is not the case that one is superior to the other, just that they reflect different views. Similarly, with university rankings, those given by *The Times*, *The Guardian* and *The Independent* are concerned with UK universities and all include staff-student ratio, entry standards, some measure of expenditure per student and degree classification. They also have one or two measures of student satisfaction. *The Higher*, on the other hand, considers universities worldwide and uses six variables two of which, *peer review* and *employer review*, reflect constituencies of a very different kind than the usual relation between university and student. To the extent that these views may not be well founded the result is unreliable, although the views of peers may be useful to career academics when considering a job move and those of employers of more than passing interest to students. However well or ill founded the views of, say, employers may be they are their views and presumably condition the decisions made on recruitment. Provided that the data collection process is not flawed the user of the data (though not necessarily of the ranking) needs to be clear what variables are used and how important they are. It is not the case that because different rankings do not agree that their results are invalid, rather that one needs to be clear about what they measure.

The data

The ranking of the top 100 full-time MBA programmes worldwide published by the *Financial Times* on 29 January 2007 is used to illustrate the application of the portfolio model. Data were collected from each school and from a sample of its alumni. Performance was assessed using twenty attributes. Details of the measures used and the collection of data are given in the newspaper. Table 1 shows the attributes and, in column *a*, the weights which were used. Aggregation to obtain an overall score uses the method described above in (1) to (3).

It is a peculiarity of the published tables that while data are provided for twelve attributes the remaining eight are given only as ranks. This makes it impossible to replicate the *FT* calculations. To make some use of the data either the ranks may be taken as real values or the eight ranked attributes may be omitted. There seems to be a small advantage in favour of the reduced set in that it is at least possible to have some confidence in the data used rather than having to misrepresent ordinal data. The purpose here is to use an illustration rather than to provide a detailed re-examination of the calculations made by the newspaper. Column *b* of Table 1 shows the weights for the reduced set scaled to sum to 1.

It is worth noting that using (3) means that the worse and best performing alternatives will generally have different scaled values for each attribute. This is shown in Figure 1. Extremely good performance is rewarded not just because it is the best of the group but because it is comparatively rare. For example, the programme with highest alumni salaries is scored more than twice as highly as that with the highest percentage of alumni employed at three months. There is an analogous penalty regime for particularly poor performance.

Results

Judgemental input can take the form of either ranges for weights or a preferred ordering or both. To illustrate the effects of imposing a preference ordering the *FT* scheme

$$w_1,w_2 > w_3 > w_4\,,\, w_5 > w_6 > w_7,\, w_8,w_9\; w_{10},w_{11} > w_{12} \qquad (7)$$

was used as a base. The required differentiation between the six groups of weights was specified by setting the minimum difference between weights as a fraction of the higher, $(w_i\text{-}w_j) > a.w_i$ or

$$(1\text{-}a)w_i - w_j > 0 \qquad (8)$$

where w_i and w_j are in adjacent groups in (7) and $w_i > w_j$. Using these sixteen constraints and (2) and minimising σ_y^2 using (5) and the covariances shown in the upper half of Table 2 gives a quadratic programme which finds a set of weights and thereby scores on which a ranking may be based. Values a=0.2, 0.3 and 0.4 were used. Three other models were calculated for comparison: that minimising σ_y^2 without weight constraints (8); uniform weights minimising σ_w^2; the *FT* weights shown in column *b* of Table 1. The results are shown in Tables 3 and 4 and Figure 2. It is notable that the weakly ordered weights and the *FT* weights (C to F) give highly correlated values of σ_y and σ_w (Figure 2), showing that requiring a greater differentiation between weights necessarily increases the spread of scores and so the differentiation between alternatives. In addition, as *a* increases the weights more closely resemble the *FT* weights (Table 4). The weights E derived when a=0.4 are very similar to those given by the *FT*, though this requirement is much less onerous than having to give point estimates for the weights. The weights within each group in (7) are the same although this was not a requirement; the constraints only specified the ordering with respect to weights in an adjacent group.

Figure 3 shows distributions of scores for those models which minimise σ_y^2 and for the *FT* weights. Although the interquartile range does not change much the scores of the higher scoring schools become more attenuated as the weights are increasingly differentiated. The distributions of attribute values generally exhibit some skew (Table 1, column *c*): attributes 3 and 7 notably negative; 1, 4, 10, 11 and 12 notably positive. It might be expected that the aggregation to form an overall score would mean that score distributions tend to Normality. In general they do in that the skew is much reduced (Table 3) though it becomes increasingly positive as weights are more constrained. Although results C and D are not significantly different from Normal (Kolmogorov-Smirnoff and Shapiro-Wilk tests with 5% significance level) the others are. Figure 4 shows an example. Weights increasingly favour the two variables describing alumni income, both of which have a positive skew, so that the distribution of scores itself shows increasingly positive skew.

The extreme outlier at the low end of the rankings in C to E is much less pronounced with the *FT* weights F. This outlier is the programme with the extremely low *z* value of -6.8 for attribute 3 (Figure 1). Comparing the E and F weights (Table 4) shows similar values for all except w_3, which is only half the value in F that it is in E. This accounts for the improved performance.

Table 5 shows the top ten programmes in all six evaluations. Of those evaluations with weight constraints (C to F) seven programmes appear in all four lists. It would seem that introducing just a little preference information gives a base structure to the evaluation which does not change that much as the constraints become more prescriptive. This is also shown by the correlations in Table 6. The correlation between the equal weights and *FT* weights (A and F) is particularly poor. Policano (2007) found this same result in his analysis of the *US News* ranking. The correlations between the base model, min σ_y^2, with no weight constraints (B) and other models are poor. It is not surprising that results with an increasingly sharp articulation of a preference structure should be dissimilar to that with none. The value of B is that it establishes an unconstrained minimum for σ_y^2 against which others may be compared.

A comment on the 2008 data

The rankings for 2008 were published on January 28th. The resulting correlations are shown in the lower half of Table 2. A full comparison with the 2007 data is not possible because variable 6, *aims achieved*, was given as a rank rather than a percentage and so correlations with this variable cannot be calculated. This treatment is odd; *aims achieved* was also given as a rank in 2003 and 2004 but not before then nor between 2005 and 2007. In 2001 only six of the twenty variables were given indirectly (as scaled indices rather than ranks) whereas now it is nine. It would be interesting to know why this is so.

Although calculations similar to those for 2007 are not possible it can be seen from (6) that if the covariances (and so correlations) are not dissimilar neither will be the resulting value of σ_y^2. Figure 5 shows a graphical comparison from which it can be seen that the structural relations between variables have not changed much and so neither would the result be expected to change much either.

Clusters and gaps

The purpose of an analysis is to determine which programmes have performance levels similar to that of others: scores have value only inasmuch as they distinguish between alternatives. A ranking is just an extreme application of this idea. Making these distinctions is difficult for two reasons. First, because the model (1) reduces a multidimensional description of each alternative to a single figure, the score. Had this not been done some form of cluster analysis would find groups of programmes which had similar form, similar profiles of performance across their vectors of attributes. Second, had the uncertainty due to sampling or due to weight specification or both been described statistically then the resulting probabilistic estimate of each score could be expressed as a confidence interval and statistically significant differences between scores found. Groups of statistically similar alternatives might then be formed, though the result may be just a set of overlapping confidence intervals (The Royal Statistical Society, 2003) rather than clusters. This would still make it possible to say, for instance, that an alternative ranked fifth has a score significantly different from that ranked thirty fifth even though it is not possible to differentiate between alternatives which are adjacent or close. This is an important idea.

Although neither approach is used in published rankings, the inability perfectly to discriminate is recognised. For example, the *FT*, in its ranking, identifies four groups: 16 in the topmost group and then groups of 16, 21 and 47. (It has to be said that these groupings are not apparent in the analyses here and so must depend on the variables which were excluded from this analysis.) Given that the distribution of scores is a positively skewed continuum (hence the increasing size of the groups) it is unsurprising that the differences between programmes in the low density tail tend to be greater than those in the higher density body, but to say that these larger differences are significant or remarkable leads to an inconsistency. For example, in the *FT* rankings it must surely be the case that the gap in scores between, say, the programmes ranked 60 and 80 is at least as big as that between the programmes ranked 53 and 54, yet the latter gap is seen as in some sense significant while programmes 60 and 80 are part of an undifferentiated group. The argument – that if neighbours in a chain cannot be differentiated then no two members of the chain can be differentiated – is reminiscent of an incremental reasoning with roots in antiquity (see, for instance, Barnes, 1982). It is an argument which is hard to support.

MBA rankings will remain. They appear simple and many find them useful. The purpose here was not to examine rankings but only to describe a method which could be of use in finding the weights which they, and others, might use.

Conclusion

This paper is concerned with finding a set of weights which are a justifiable articulation of the judgements implicit in making a ranking. Taking the broadly statistical view of the portfolio model shows clearly the relation between the information contained in the weights and in the scores on which ranks are based. Given the imprecision inherent in any expression of preference it is important to guard against undue discrimination between alternatives. It is this focus on results (scores) rather than on the weights themselves which is the basis of this method.

Weights are seen as expressions of priority so that, at least, a rank ordering of weights is possible. Using this as a constraint in an optimisation which enforces caution reconciles apparently conflicting requirements: to be opinionated and yet to reach minimally biased conclusions.

It is the rank order of weights and the relative differences between them which largely determine the result. As Table 4 shows, the greater this relative difference the more the minimally discriminating weights resemble the *FT* weights. Even so, the results obtained (Table 5) are different from those found by the *FT*. The ranking of the top ten using the reduced data set but *FT* weights (result set *F*) contains eight of the *FT* top ten, but even using a small relative difference between weights (*C*) it contains six. It is the information contained in the unrecoverable eight variables that makes the difference.

It is usual (though not in published rankings) that weights may be altered in a sensitivity analysis to see the effects of imprecision and whether plausible differences in weights have a large effect. It would certainly be helpful if some such results could be published so that real differences in performance could reliably be identified. In recognition of the problem most publications contain some statement to the effect that differences between scores may not be significant. Further analysis would, of course, undermine the existing simple ranking and it is easy to see why newspapers and magazines would not find this helpful.

The success of the multicriteria model is due in part to its modularity. This enables users to consider sub-problems rather than the whole so that the elicitation of weights is a task the results of which are used in computing scores. In this framework weights may assume an unnecessary importance. Finding weights which minimise the variance of scores goes some way to reintegrating the modular framework. The same set of judgemental constraints will give different weights with a different problem. This seems odd only if the weights themselves are invested with a significance other than their role of encoding judgements: it is the judgements themselves and the results which are key. In particular, the problem provides the context in which the judgements are articulated as numbers: changing the context will change the numbers while retaining the judgements.

Evaluations must be believed, at least to the extent that they act as a guide. Using a method which minimises the effects of bias helps to justify the analysis. The consumer of the results ought to be reassured that the method of calculation does not itself contribute to apparently significant (important, at least) results but rather that all the information contained in the scores is a function of preference statements and data and nothing else.

References

Barnes J (1982). Medicine, experience and logic. In: Barnes J, Brunschwig J, Burnyeat M and Schofield M (eds.) *Science and Speculation: Studies in Hellenistic theory and practice*. Cambridge University Press: Cambridge.

Barron H, Schmidt CP (1988). Sensitivity analysis of additive multiattribute value models. *Operations Research* **36**: 122–127.

Borcherding K, Schmeer S and Weber M (1995). Biases in multiattribute weight elicitation. In: Caverni J-P, Bar-Hillel M, Barron FH and Junngermann H (Eds). *Contributions to Decision Making – I*. Elsevier: Amsterdam.

Bickerstaffe G and Ridgers B (2007). Ranking of business schools. *Journal of Management Development* **26**: 61–66.

Bradshaw D (2007). Business school rankings: the love-hate relationship. *Journal of Management Development* **26**: 54–60.

Breiman L, Friedman JH, Olshen RA and Stone CJ (1998). *Classification and Regression Trees*. CRC: Boca Raton.

Colbert A, Levary RR and Shaner MC (2000). Determining the relative efficiency of MBA programs using DEA. *European Journal of Operational Research* **125**: 656–669.

Dichev I (1999). How good are Business School rankings? *Journal of Business* **72**: 201–213.

Engwall L (2007). The anatomy of management education. *Scandinavian Journal of Management* **23**: 4–35.

Gabert P and Brown DE (1989) Knowledge-based computer-aided design of materials handling systems. *IEEE Transactions in Systems, Management and Cybernetics* **SMC-19**: 188–196.

Green RH and Doyle JR (1995). On maximising discrimination in multiple criteria decision making. *Journal of the Operational Research Society* **46**: 192–204.

Herfindahl OC (1950). *Concentration in the U.S. Steel Industry*. Columbia University: New York. (Ph.D. Thesis.)

Hirschman AO (1964). The paternity of an index. *The American Economic Review* **54**:761–762.

Jaynes ET (1957). Information theory and statistical mechanics. *Physics Review* **106**: 620–630 and **108**: 171–190.

Jaynes ET (2003). *Probability Theory: The Logic of Science*. Cambridge University Press: Cambridge.

Jessop A (1999). Entropy in multiattribute problems. *Journal ofMulti-Criteria Decision Analysis* **8**: 61–70.

Jessop A (2004). Sensitivity and robustness in selection problems. *Computers and Operations Research* **31**: 607–622.

Mar-Molinero C and Mingers J (2007). An evaluation of the limitations of, and alternatives to, the Co-Plot methodology. *Journal of the Operational Research Society* **58**: 874–886.

Markowitz H (1952). Portfolio selection. *Journal of Finance* **7**: 77–91.

Mustajoki J, Hämäläinen RP and Lindstedt MRK (2006). Using intervals for global sensitivity and worst-case analyses in multiattribute value trees. *European Journal of Operational Research* **174**: 278–292.

Paucar-Caceres A and Thorpe R (2005). Mapping the structure of MBA programmes: a comparative study of the structure of accredited AMBA programmes in the United Kingdom. *Journal of the Operational Research Society* **56**: 25–38.

Peters K (2007). Business school rankings: content and context. *Journal of Management Development* **26**: 49–53.

Policano AJ (2005). What price rankings? *BizEd* **4**: 26–32.

Policano AJ (2007). The rankings game: and the winner is… *Journal of Management Development* **26**: 43–48.

Theil H (1972). *Statistical decomposition analysis: with applications in the social and administrative sciences*. North-Holland: Amsterdam.

The Royal Statistical Society (2003). *Performance indicators: good, bad and ugly. Report of the Royal Statistical Society working party on performance monitoring in the public services*. London: Royal Statistical Society.

Segev E, Raveh A and Farjoun M (1999). Conceptual maps of the leading MBA programs in the United States: core courses, concentration areas, and the ranking of the School. *Strategic Management Journal* **20**: 549–565.

Siemens JC, Burton S, Jensen T and Mendoza NA (2005). An examination of the relationship between research productivity in prestigious business journals and popular press business school rankings. *Journal of Business Research* **58**: 467–476.

Simpson EH (1949). Measurement of diversity. *Nature* **163**: 688.

Soofi ES (1990). Generalized entropy-based weights for multiattribute value models. *Operations Research* **38**: 362–363.

Soofi ES and Retzer JJ (1992) Adjustments of importance weights in multiattribute value models. *European Journal of Operational Research* **60**: 99–108.

Thomas H (2007). Business school strategy and the metrics for success. *Journal of Management Development* **26**: 33–42.

Tversky A and Kahneman D (1974). Judgement under uncertainty: heuristics and biases. *Science* **185**: 1124–1131.

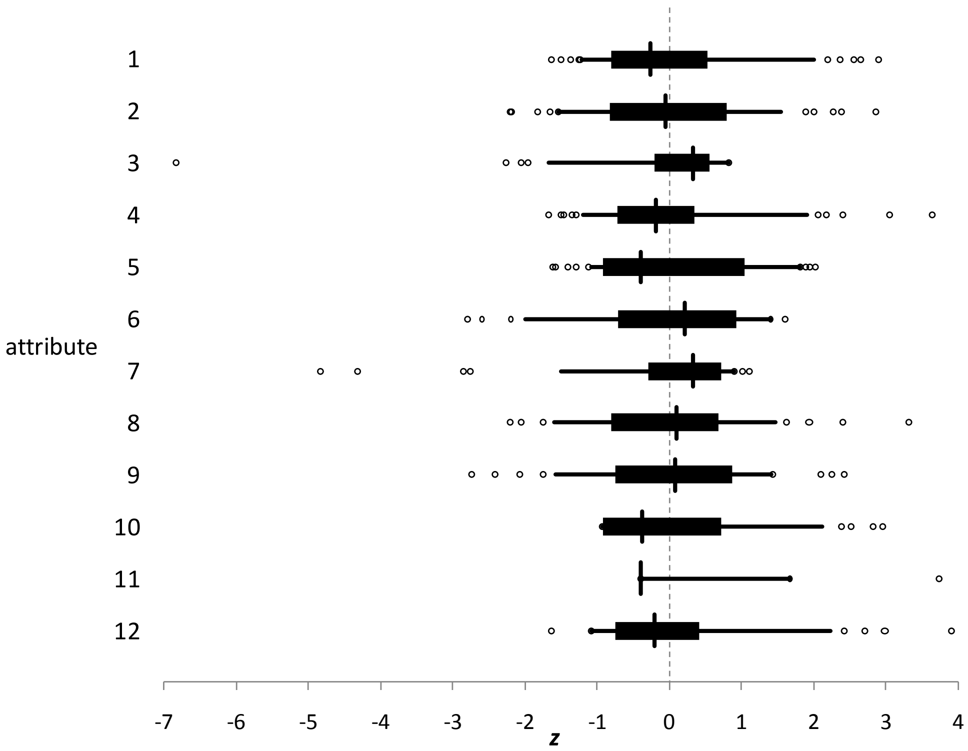

Figure 1. Distributions of attributes. For each attribute the thicker line shows the spread of the middle 50 values, the thinner line the middle 90, and the circles the highest and lowest five values. The vertical line shows the median.

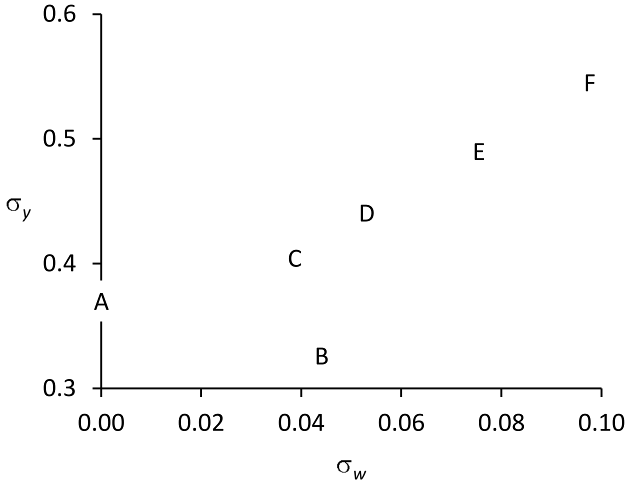

Figure 2. Standard deviations of scores and weights (see Table 3).

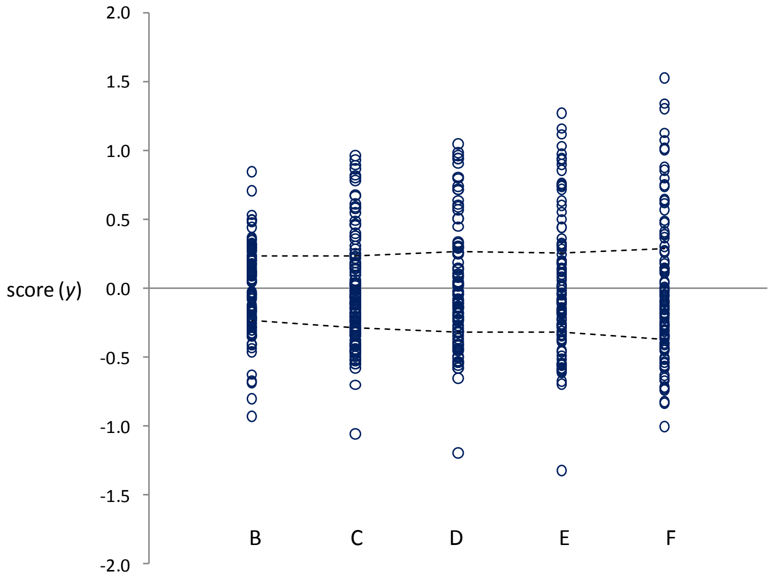

Figure 3. Distributions of scores. Dashed lines join the quartiles.

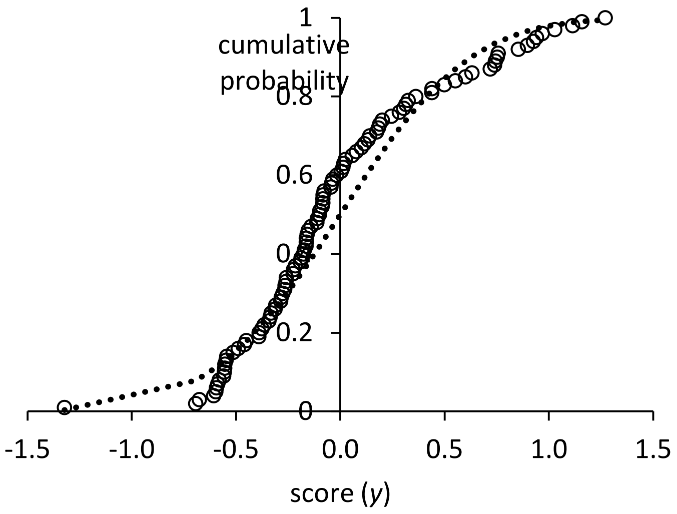

Figure 4. Comparison of E scores with Normal distribution.

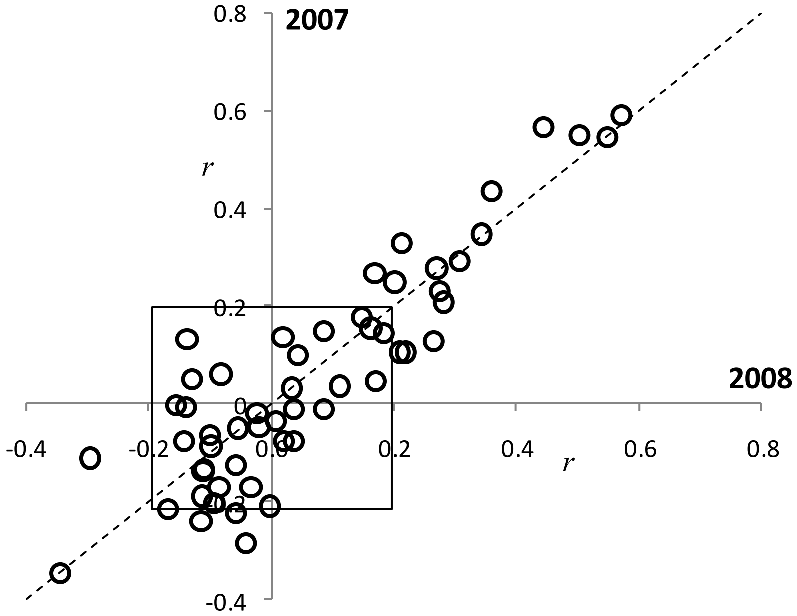

Figure 5. Correlations in 2007 and 2008. The diagonal line shows equal values. The box shows correlations insignificant ($p \geq 0.05$) in both years.

	attribute	*FT* weights		skew
		full set	subset	
		a	*b*	*c*
1	Weighted salary (US$'000)	0.20	0.299	1.02
2	Salary increase(%)	0.20	0.299	0.27
3	Faculty with doctorates (%)	0.05	0.075	-3.67
4	International faculty (%)	0.04	0.060	1.17
5	International students (%)	0.04	0.060	0.65
6	Aims achieved (%)	0.03	0.045	-0.80
7	Employed at three months (%)	0.02	0.030	-2.64
8	Women faculty (%)	0.02	0.030	0.23
9	Women students (%)	0.02	0.030	-0.19
10	International board (%)	0.02	0.030	1.17
11	Languages (number)	0.02	0.030	2.60
12	Women board (%)	0.01	0.015	1.50
13	FT research rank	0.10		
14	International mobility rank	0.06		
15	FT doctoral rank	0.05		
16	Value for money rank	0.03		
17	Career progress rank	0.03		
18	Placement success rank	0.02		
19	Alumni recommend rank	0.02		
20	International experience rank	0.02		
		1.00	1.000	

Table 1. The *FT* data: variables and weights.

	1	2	3	4	5	6	7	8	9	10	11	12
1		**0.29**	**0.23**	**0.28**	-0.05	**0.58**	0.18	**-0.35**	-0.02	**0.25**	0.15	-0.05
2	**0.31**		0.13	**-0.21**	**-0.24**	**0.46**	0.15	-0.11	-0.08	0.04	0.10	-0.13
3	**0.28**	-0.14		0.13	**-0.21**	**0.31**	0.13	**-0.22**	-0.01	-0.01	-0.06	-0.19
4	**0.27**	-0.09	**0.27**		**0.55**	-0.05	0.06	**-0.21**	-0.09	**0.57**	**0.35**	0.11
5	-0.02	-0.12	0.00	**0.55**		**-0.24**	-0.14	0.05	0.00	**0.59**	**0.43**	**0.21**
6							0.16	-0.16	-0.09	-0.05	0.02	-0.16
7	0.15	0.09	0.02	-0.08	-0.11			-0.17	**-0.29**	-0.01	0.05	-0.08
8	**-0.35**	**-0.30**	-0.06	-0.17	0.17		-0.09		**0.33**	-0.17	-0.08	**0.27**
9	-0.02	0.02	-0.14	-0.10	-0.16		-0.04	**0.21**		-0.04	-0.14	0.10
10	**0.20**	0.11	0.04	**0.45**	**0.57**		0.09	-0.03	0.01		**0.55**	0.14
11	0.16	**0.22**	-0.10	**0.34**	**0.36**		-0.13	-0.14	-0.11	**0.50**		0.03
12	-0.06	-0.06	-0.11	**0.21**	**0.28**		0.04	0.17	0.04	0.18	0.03	

Table 2. Correlation matrix for attributes. Significant values ($p \leq 0.05$) shown in bold. Upper half shows correlations in 2007, lower half in 2008.

objective		standard deviations σ_w	 σ_y	skew
A	min. σ_w^2: equal weights	0.000	0.370	0.35
B	min σ_y^2: no weight restrictions	0.044	0.327	-0.23
C	min σ_y^2: a=0.2	0.033	0.405	0.48
D	min σ_y^2: a=0.3	0.053	0.441	0.54
E	min σ_y^2: a=0.4	0.075	0.490	0.61
F	*FT* weights	0.098	0.545	0.77

Table 3. Results of weight & score determinations showing standard deviations of distributions and also the skew of the distributions of scores (see Figure 2).

weights	C	D	E	F
w_1,w_2	0.144	0.185	0.235	0.299
w_3	0.115	0.130	0.141	0.075
w_4 , w_5	0.092	0.091	0.084	0.060
w_6	0.073	0.064	0.051	0.045
w_7, w_8,w_9 w_{10},w_{11}	0.059	0.045	0.030	0.030
w_{12}	0.047	0.031	0.018	0.015

Table 4. Weights.

rank	A	B	C	D	E	F
1	12	12	12	11	11	11
2	5	36	5	5	2	2
3	13	9	13	2	1	1
4	7	5	11	1	5	10
5	15	28	7	12	4	4
6	11	15	1	13	12	12
7	1	42	2	4	10	9
8	36	3	15	7	13	5
9	3	51	4	10	9	6
10	24	10	9	9	7	8

Table 5. MBA programmes ranked 1 to 10 under different weight specifications. The identifying numbers in the table show the position in the *FT* published list.

	A	B	C	D	E	F
A		0.912	0.953	0.894	0.820	0.753
B	0.887		0.810	0.743	0.667	0.600
C	0.955	0.854		0.987	0.953	0.867
D	0.898	0.789	0.985		0.989	0.924
E	0.803	0.690	0.935	0.978		0.970
F	0.723	0.609	0.901	0.951	0.980	

Table 6. Correlation between scores obtained in six evaluations. The upper matrix shows Pearson’s correlation and the lower Spearman’s rank correlation.

Captions

Figure 1. Distributions of attributes. For each attribute the thicker line shows the spread of the middle 50 values, the thinner line the middle 90, and the circles the highest and lowest five values. The vertical line shows the median.

Figure 2. Standard deviations of scores and weights (see Table 3).

Figure 3. Distributions of scores. Dashed lines join the quartiles.

Figure 4. Comparison of E scores with Normal distribution.

Figure 5. Correlations in 2007 and 2008. The diagonal line shows equal values. The box shows correlations insignificant ($p \geq 0.05$) in both years.

Table 1. The *FT* data: variables and weights.

Table 2. Correlation matrix for attributes. Significant values ($p \leq 0.05$) shown in bold. Upper half shows correlations in 2007, lower half in 2008.

Table 3. Results of weight & score determinations showing standard deviations of distributions and also the skew of the distributions of scores (see Figure 2).

Table 4. Weights.

Table 5. MBA programmes ranked 1 to 10 under different weight specifications. The identifying numbers in the table show the position in the *FT* published list.

Table 6. Correlation between scores obtained in six evaluations. The upper matrix shows Pearson's correlation and the lower Spearman's rank correlation.