IMP: a decision aid for multiattribute evaluation using imprecise weight estimates

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

e-mail: a.t.jessop@durham.ac.uk
telephone: +44 (0) 191 3345403
fax: +44 (0) 191 3345201

accepted for publication in *Omega*

IMP: a decision aid for multiattribute evaluation using imprecise weight estimates

1. Introduction

Differentiating multiattribute alternatives may be simple, almost intuitive, if the number of attributes is small. As the number of attributes or the number of alternatives or both increases we cannot apprehend all data unaided and so some formalism, a table of attribute values, is helpful. But cognitive limits may mean that we still cannot cope with all attributes for all alternatives [1] and so we use a model which is an acceptable abstraction of our decision process. A popular model finds a score for each alternative as the weighted sum of scaled attribute values.

$$y_j = \sum_i w_i x_{ij} \quad ; \quad i = 1\ldots n \, , j = 1\ldots m \tag{1}$$

with $$\sum_i w_i = 1 \quad ; \; 0 \le w_i \le 1 \tag{2}$$

where y_j is the score for alternative j
x_{ij} is an appropriately scaled measure of the value of attribute i for alternative j
w_i is the weight attached to attribute i

In this modular design elements of the decision process are disaggregated so that judgemental tasks are within the cognitive bounds of the user. This strategy is effective as a way of dealing with decision complexity [2], although some may nevertheless prefer holistic evaluations [3].

The elements introduced in the model – value functions, weights and aggregation – and the interactions between them are not necessarily straightforward. In particular, a number of factors may contribute to imprecise judgements about weights. The effects of cognitive limits on human information processing have been well known for a long time [4] as have the biases common in human reasoning about probabilities [5]. Different judgements are likely to be made depending on mood [6], how questions are framed [7,8] and the method of elicitation and subsequent calculation [9-11].

There is no agreement about an appropriate response to judgemental imprecision in general and about weights in particular. One view is that any attempt at quantification is foolish [12,13], that what makes judgements about values different from judgements about events or facts is that values need to be explored in sensitivity analyses until a resolution is reached. The alternative view is that while imprecision may be reduced it is not necessarily eliminated and that this residual should be described and modelled, most obviously by specifying probability distributions. This approach has frequently been reported, usually using simulation to generate probability distributions for scores or ranks.

The argument against the probabilistic approach is that users may find probability a too abstract concept. Biases in the heuristics used to reason about probabilities were identified by Tversky and Kahneman [5] and stimulated a large body of research [14]. The situation is summarised in Hogarth's well-known conclusion that "man, as a selective, step-wise information processing system with limited capacity, is ill-equipped for assessing subjective probability distributions" [15]. Although users may make the necessary probability specifications it may not be clear either to themselves or to others just what they have done and so commitment to a decision is undermined.

The argument in favour is that the implications of imprecision need to be dealt with in ways that are in some sense comprehensive and consistent: comprehensive in that a probability distribution encompasses all plausible outcomes and not just those used in some sensitivity tests; consistent in that calculations avoid judgemental biases. There are other advantages. Being able to give ranges rather than precise values reduces the stress of elicitation [16] and in group decision avoids a too early commitment and unnecessary conflict [17].

In multiattribute evaluation two questions arise. First, given the judgemental uncertainty of weight estimation is it reasonable to believe that two alternatives may be differentiated by their scores? Second, if they can, is it meaningful to say that, in the context of the problem, they are

materially different? The first question – can these scores be differentiated – may be answered by a significance test. The second question – is this difference meaningful – is better answered by considering scores.

This paper is concerned with the first of these questions. Weights are modelled by a joint probability distribution which permits the probabilistic estimate of differences between scores without the need for simulation. Pairs of scores which cannot be differentiated result in a clustering of alternatives. More precision in weight estimates gives fewer and smaller clusters. These clusters may be sufficient to identify, say, an initial screening [18] of candidates to construct a short list for further examination. The model used – IMP, for IMP*recise multiattribute evaluation* – is the development of an earlier model [19] but now incorporating a number of improvements and extensions.

Judgemental uncertainty needs to be described quantitatively for modelling but it is likely that it is first conceptualised using words or phrases. The relation between the two languages – words and numbers – arises particularly when considering weights and probabilities but the difficulties are quite general. Language theorists have long studied these issues and their approach is used here to provide a useful analytical framework. The next two sections show this framework and its application to multicriteria problems.

In the rest of the paper a model is described for using probabilistic weight assessments derived from judgements. The model is applied to a shortlisting problem both for an individual assessor and for several, illustrating how the approach allows the monitoring of the usefulness of the aggregation of assessors' judgements.

2. Words and numbers

2.1 Imprecise description

Giving numerical judgements is not easy: "The major simplification associated with eliciting parameter imprecision in a prespecified format … is the natural language dialogue needed to establish a model of the situation" [20]. The difficulty is well captured by Williamson [21] who writes that "precise assertions … are more likely to be false, but more useful if true" but that "vagueness is a precondition of the flexibility of ordinary language". It is this tension between the desire for decisiveness and the ambiguity of language which complicates attempts at elicitation and interpretation, a difficulty compounded in group decision making when the words may have "local and cultural meanings which some participants may not understand" [22].

Imprecision in language use arises because of both the *inconsistency* of application by individuals and groups and the *vagueness* or ambiguity of application by individuals (a familiar dichotomy [23]).

2.2 Inconsistency

In the study of language usage the proportion of people applying a particular label to a stimulus defines the consistency profile [24] or cue validity [25]. (This has also been used for decision support [26]). The same idea describes the consistency of application not just of a word but of a number, as in a weight assessment, and not just by different people but by the same person giving several assessments.

In giving weight estimates person p makes $n_p > 1$ single valued numerical responses to the same or equivalent questions. These responses have mean $\mu_{C,p}$ and variance $\sigma^2_{C,p}$.

In a language or user group G the inconsistency is the aggregation of the responses of the members. Group inconsistency is described by a probability ditribution with

$$\left.\begin{array}{ll} & \text{mean} = \mu_{C,G} = \sum_p \pi_p \mu_{C,p} \\ \text{and} & \text{variance} = \sigma^2_{C,G} = \sum_p \pi_p (\sigma^2_{C,p} + \sigma^2_{M,p}) \\ & \\ \text{where} & \pi_p = n_p / \sum_j n_j \\ \text{and} & \sigma^2_{M,p} = (\mu_{C,p} - \mu_{C,G})^2 \end{array}\right\} \quad (3)$$

It has often been found that there is much less consistency between members of a group than for individuals [27], in part because $\mu_{C,p} \neq \mu_{C,G}$. The larger these differences the less justified is it to call the group a coherent language (or user) group.

2.3 Vagueness

An individual may not be sure how appropriate it is to apply a label (in this case, a weight). Degree theorists wish this uncertainty about labelling to be quantified by a membership function of some measurable characteristic: 1 if the label certainly is appropriate and 0 if it certainly is not. This idea is most familiar in multiattribute problems as the motivation for fuzzy analysis, though to acknowledge vagueness is not to accept fuzzy analysis [28].

Deciding the points at which a label becomes inappropriate gives a range [29]. It is common to give an intermediate value and then to assume a triangular function. The intermediate value may be that which is judged more appropriate for the label than any other but this doesn't mean it is judged to be a perfect exemplar (membership=1). Wallsten et al [30] describe an experiment in the analysis of which "these scale values, normalized to be nonnegative with an arbitrary maximum of 1 … can be taken as the membership function". The key word is "arbitrary".

Requiring an assessor to distinguish between "degree of truth = x" and "probability = p" may be optimistic, so that when Jackendoff writes that a fuzzy set is "a set whose membership is defined not categorically, but in terms of the degree or probability of membership" [31] he may well be articulating a common habit of mind.

The vagueness of articulation by person p may be described by a probability distribution [32] with mean and variance $\mu_{V,p}$ and $\sigma^2_{V,p}$. In the case that there are $n_p > 1$ vague estimates simple averaging provides a good combination of probability estimates [33] and so $\mu_{V,p}$ and $\sigma^2_{V,p}$ are just the mean values of the n_p individual means and variances.

The vague articulation of group G has

$$\left.\begin{array}{ll} & \text{mean} = \mu_{V,G} = \sum_p \pi_p \mu_{V,p} \\ \text{and} & \text{variance} = \sigma^2_{V,G} = \sum_p \pi_p \sigma^2_{V,p} \end{array}\right\} \quad (4)$$

This estimate of variance assumes independence between group members. The plausibility of this assumption is discussed below in light of some experimental results.

2.4 Imprecision

Inconsistency describes the variability with which a value (a word or a number) is used in repeated application. Vagueness of articulation describes inherent uncertainty about a single judgement. The characteristic pattern of usage is a combination of the two:

imprecision = inconsistency (of use) + vagueness (of articulation)

While inconsistency and vagueness may be related we assume here that this is a negligible effect and so the means of the distributions describing vagueness and imprecision are the same and

individual imprecision for person p has mean and variance

$$\left.\begin{array}{l} \mu_{I,p} = \mu_{C,p} = \mu_{V,p} \\ \sigma^2_{I,p} = \sigma^2_{C,p} + \sigma^2_{V,p} \end{array}\right\} \quad (5)$$

imprecision for group G has mean and variance

$$\left.\begin{aligned} \mu_{I,G} &= \textstyle\sum_p \pi_p \mu_{I,p} \\ \sigma^2_{I,G} &= \textstyle\sum_p \pi_p (\sigma^2_{I,p} + \sigma^2_{M,p}) \end{aligned}\right\} \quad (6)$$

In multiattribute problems this provides a framework for the probabilistic description of weight estimates.

2.5 Probability is imprecise

Inconsistency is the sum of the variances of a number of point estimates. Vagueness is more difficult because probabilities, although numerically precise, are not always easily understood so that individuals are both vague and inconsistent in their assignment of labels to probability values [34,35]. Over a century ago Karl Pearson, discussing significance testing, described a probability of 0.1 as "not very improbable" and one of 0.01 as "very improbable" [36], an interpretation degree theorists would recognise. In the assessment of probabilities interpersonal variability is usually higher than intrapersonal variability in the assignment of numbers to probability statements [27], especially if users are from different cultures [37].

It follows that how a receiver interprets a probability value will not necessarily be the same as the (vague) intention of the sender, so that using probability to communicate imprecision is not straightforward. This leads to the apparent paradox that people often prefer to give probability estimates as words but to receive them as numbers [38].

The use of a standardised lexicon may help [39] by acknowledging that people feel more at ease giving verbal estimates but that numbers are needed for calculation [40], though this is more likely to be of use in making assessments than in reporting results.

Because of these difficulties vague probability assessments ought to be treated as parameters in a sensitivity analysis.

3. Application to multicriteria problems

There exist a great many recommendations of how best to ask questions the answers to which may be interpreted as judgements about the values of weights. Although weight values are likely to vary depending on the method of elicitation [9-11] it is unusual for different methods to be used in any single application, though inconsistency due to method variation would fit the framework above.

Imprecision is most easily expressed by giving just upper and lower limits. These may be treated as deterministic bounds as in the ARIADNE model [20] which allow ranges to be set as constraints and minimum and maximum values of scores found for each alternative. If these ranges do not overlap then a clear preference has been established and a decision may be made. If not, making the constraints tighter and rerunning the model might lead to a decision, provided that such tightening is possible. Testing the effects of changing limits recognises the vagueness of the initial values.

This use of limits also underpins the PAIRS model [41] and preference programming generally [42], with application to SMART and SWING models [43].

The SMAA model [44] and its derivatives [45,46] explore the weight space by finding weight combinations which result in a given alternative achieving a particular rank . The proportion of the whole weight space volume defined by those combinations is a measure of support for that rank. This *acceptability index* may be interpreted as the probability of the ranking given specified probability distributions for the weights. In particular, if no distributions are given it is assumed that all weights are specified by maximally ignorant uniform distributions over the range [0,1]. In the SMAA-2 model [47] weights may be described in a number of ways, such as point estimates, or as an ordering or, if there is no preference information, the uniform weight distribution is used [48].

The Analytic Hierarchy Process, AHP, has attracted a body of work relevant to this paper. The reciprocal matrix method estimates values for n weights from $n(n\text{-}1)/2$ judgements of weight ratios, a problem with positive degrees of freedom. This allows for an estimate of inconsistency whereas with

direct assessment in models such as SMART, having no degrees of freedom, only vagueness may be assessed.

AHP is also noted for providing a guide for users in giving weight ratios from 1 to 9. The adoption of this lexicon leads to a method which may be seen as "essentially qualitative and not realistically quantitative" [49] and that such judgements "should be treated as qualitative information without associating any quantitative meaning" [50]. These verdicts may be a little harsh but do draw attention to the difficulty involved in language use in pursuit of constructing a "model that computes with words directly" [51], which seems an impossible goal. Shirland et al [52] report the use of a number of suggested ratio values from 1-3 to 1-9 with and without suggested meanings and that "results are reasonably consistent across rating systems" but that "respondents find the 1-9 scale mentally taxing". Inconsistencies in using lexicons [53] indicate that their use may impair communication between assessors. (The reciprocal matrix method has been used to construct a lexicon [54]).

A number of studies of AHP have used simulation to find the effect of vague weight estimates, frequently using uniform or triangular distributions [54-57]. The focus is usually on the effect on ranks [58-60] so that the probability of an alternative having a particular rank is given, as with the SMAA models. These and other papers [61,62] have suggested alternative probability distributions for describing joint imprecision about weights.

Simulation is unnecessary with linear models such as (1). Rosenbloom [58] uses three point estimates and from them obtains means and variances analytically using methods familiar in decision and risk analysis. This approach is adopted here.

4. Method

4.1 Overview

If the uncertainty about weights can be described by a probability distribution with variance/covariance matrix σ_{ij}, the variance of the estimate of the score for alternative k is

$$\mathrm{var}(y_k) = \sum_i \sum_j \sigma_{ij} x_{ki} x_{kj} \qquad (7)$$

The difference of scores for alternatives a and b has mean $(y_a - y_b)$ and variance

$$\mathrm{var}(a,b) = \sum_i \sum_j \sigma_{ij} (x_{ia} - x_{ib}) (x_{ja} - x_{jb}) \qquad (8)$$

The Dirichlet distribution has been used to model imprecise probabilities [63,64] and in combining expert opinion [65] as well as modelling weights [66]. It is a multinomial form of the Beta distribution and so has Beta marginal distributions which, having limits [0,1], are appropriate models for single weight estimates as they have been for imprecise probabilities. Model parameters are found so that, first, the marginal means are the same as the assessed mean weight values and, second, so that the marginal variances are as close as possible to the variances of the imprecise weight estimates. The steps in the model are

1. Obtain a 3-point estimate for each weight
2. Infer mean and variance for each
3. Fit a Dirichlet distribution
4. Calculate means and variances of the differences between pairs of alternatives
5. Display differences, subject to imprecision, to aid discrimination

4.2 The Dirichlet distribution

The Dirichlet density

$$f(w_i) \propto w_i^{u_i - 1}(1-w_i)^{\sum_j u_j - u_i - 1}$$

has marginal mean and variance

$$\mu_i = u_i / \sum_j u_j \quad (9)$$
$$\sigma_i^2 = \alpha\mu_i(1-\mu_i) \quad (10)$$

where $\alpha = 1 / (1 + \sum_i u_i)$ (11)

The mean values (9) are used in the evaluation (1). Because of (2) weights are negatively correlated:

$$\sigma_{ij} = -\alpha\mu_i\mu_j \quad ; \; i \neq j$$
and $$r_{ij} = -[\, \mu_i\mu_j / (1-\mu_i)(1-\mu_j) \,]^{0.5}$$

In this way the Dirichlet model accounts for correlations between weight estimates even though marginal judgements will almost certainly have been made disregarding them.

While any non-negative parameter values are permissible, values $u_i<1$ give U-shaped Beta distributions and for $u_i=1$ a highly skewed distribution with a mode of 0 (or 1) and resembling a right triangle. It is not plausible that assessors had either in mind when making judgements and so the restriction $u_i>1$ applies in what follows.

Mean weight values (9) are determined by the *relative* values of the parameters. The precision of the estimate (10) is mainly determined by the *sum* of the parameter values, so that α summarises the overall precision of the estimates.

4.3 Three-point estimates

In describing their experiences in eliciting subjective probabilities Kadane and Wolfson [67] offer a number of recommendations among which are that experts should be used, that quantiles of observable quantities should be elicited, and that estimates of variance and higher moments of the distribution should not be sought.

In risk analysis and project planning judgements are made of the values of quantities which are observable, if only in retrospect, and about which experts have knowledge and this permits assessment of judgemental accuracy [68]. If these estimates are based on recollection the availability heuristic [5] would support an interpretation of best estimates as modes. There is no reason to think this argument also holds for weight estimates which encode vague judgements of values rather than estimates of verifiable quantities. It is the mean value used in (1) which the assessor will have in mind and so the best weight estimate, M, will be taken as a mean.

In providing three-point estimates for activity times in PERT analyses [69-72], and in decision analysis [73], low and high estimates, L and H, have been interpreted as percentiles of a probability distribution determining a c% confidence interval and the formula

$$\text{estimated standard deviation, } s = b(\text{H-L}) \quad (12)$$

used where b depends on the confidence level c. In the literature values for b have been given for particular values of c. The results may be generalised by generating values of L and H for Beta parameters = 2,3…9 and for c = 80, 85, 90, 95, 99. Values of b were found by regression of s against (H-L) as in (12). The results were good: $r^2>0.99$ except for c=99 when r^2=0.94. Figure 1 shows the relation which may be used to estimate b from c:

$$b = 1.066 - 0.00853c \quad (13)$$

4.4 Fitting the Dirichlet distribution

In the Dirichlet equations (9) and (10) put $\mu_i = w_i = M$. To most closely respect the judgemental variances minimise $\sum_i(\sigma_i^2 - s_i^2)^2$ to give

$$\alpha = \sum_i w_i (1 - w_i) s_i^2 / \sum_i [w_i (1 - w_i)]^2 \qquad (14)$$

The summations are over only those weights for which three-point estimates have been made. For example, in direct assessments where an anchor value is used uncertainty will be expressed only about other values but not about the anchor itself, which is fixed. This does not prevent an estimate for α being made so that variances and covariances are found even for those (few) weights for which only a point estimate is available because it is an anchor or when the number of judgements is high and fatigue is a factor, as with large reciprocal matrices, perhaps. The more general case of missing information is well recognised [74].

If, in (14), the variance s_i^2 is the aggregation of a number, k, of variances (such as for a number of assessors and/or from a number of sources), and assuming independence,

$$s_i^2 = s(1)_i^2 + s(2)_i^2 + \ldots s(k)_i^2$$

then $$\alpha = \sum_i w_i (1 - w_i)[s(1)_i^2 + s(2)_i^2 + \ldots s(k)_i^2] / \sum_i [w_i (1 - w_i)]^2$$

$$= \sum_i w_i (1 - w_i) s(1)_i^2 / \sum_i [w_i (1 - w_i)]^2 + \sum_i w_i (1 - w_i) s(2)_i^2 / \sum_i [w_i (1 - w_i)]^2 + \ldots$$

$$\alpha = \alpha(1) + \alpha(2) + \ldots + \alpha(k)$$

It may sometimes be more convenient to aggregate α values rather than variances.

5. Non-judgemental analysis

An assessor may feel able to give point estimates for weights but does not wish to give lower and upper bounds. We assume that there is no objection in principle to the use of a probabilistic description of vagueness, just a requirement that because of the wish to be minimally judgemental it should contain as little information as possible. The variance of Dirichlet weight estimates (10) is maximised when α is maximised by minimising the parameters u_i (11). The smallest parameter value, $u_{min} = 1+\varepsilon$, is that associated with the smallest weight, w_{min}. Other parameters are scaled to preserve the weights

$$u_i = (w_i / w_{min})(1+\varepsilon)$$

which gives $\sum_i u_i = (1+\varepsilon) / w_{min}$ and, as ε is trivially small,

$$\alpha = 1 / (1 + 1/w_{min}) \qquad (15)$$

There are two special cases. First, there may sometimes be a reluctance to provide any judgements about weights. This may be because the assessor is reluctant to say anything about relative preferences, or it may be that some minimally informed base case is to be established. It is usual to interpret this using a uniform weight distribution justified by either Laplace's indifference principle or by maximum entropy arguments that weights should be minimally different. Since all weights are $1/n$, $\alpha = 1/(1+n)$. Second, the assessor may rank weights. Rank order centroid (ROC) weights [75] have been found to give good estimates [76] in which $w_{min} = 1/n^2$ and so $\alpha = 1/(1+n^2)$.

The larger the problem the greater the precision of these weight estimates, and just by having unequal weights precision is increased.

6. Illustration

6.1 Data and judgement

Figure 2 shows an application of the method.

The task was the initial evaluation of Northern European full-time MBA programmes using data from the 2011 *Financial Times* listing. There were twenty one programmes. For this exercise each was described by seven of the twenty attributes used by the *FT*:

Salary increase
Aims achieved
Employed at 3 months
Women faculty
Women students
International faculty
International students

The value function was chosen to scale attributes to the range [0,1].

Weight values given by an MBA student are shown at the bottom left of the screen. Direct elicitation was used: an anchor value of 100 given to the most important attribute and three-point estimates made for all other weights. The central values are scaled to sum to 1 and are shown next to the input. Low and high values are scaled by the same factor. These are the values L,M,H.

The student agreed to give 5th and 95th percentiles to describe a 90% confidence interval. The resulting Dirichlet model had a parameter $\alpha = 0.0086$.

The results of this sort of evaluation are usually shown as a ranked list based on scores, as in the *FT* and elsewhere, but this is insufficient for it takes no account of imprecision. The two-dimensional chart shows the different programmes separated by distances closely approximating the values

$$z_{ab} = |y_a - y_b| / [\sum_i \sum_j \sigma_{ij} (x_{ia} - x_{ib}) (x_{ja} - x_{jb})]^{0.5} \qquad (16)$$

Giving a value for p, as in statistical significance, identifies pairs of programmes between which discrimination is less easy to justify given the judgmental imprecision about weights: smaller values of p provide greater support for discrimination. In the illustration p=10%.

6.2 Display and interaction

How results of imprecise analysis are shown is important if the analysis is to be useful [77]. The object of the screen design for this simple spreadsheet was that the user could readily see both the judgements made and their results so that exploratory interaction is made easy.

The main output is the chart in the centre of the screen. This shows the alternatives identified either by a sequential identifier or, as here, by their score rank. To help in using this information lines are drawn between pairs which are, in the statistical sense, not significantly different: given the imprecision of the weight estimates there is little support for differentiation based on scores. The display encourages the consideration of clusters rather than ranks, though ranks are shown too. This grouping shows "alternatives whose relative rankings cannot be visibly differentitaed" [78] rather than the more familir clusters based on the similarity of attributes [79].

Changing the p value will alter the membership of justifiable clusters and show the point at which a particular pair may be differentiated. In this way users may explore the effect of setting different p values, the degree of support for differentiation, which is hard to set in advance.

This p value is shown to the right with some basic inputs and the confidence level for the three-point weight estimates. Neither p nor the confidence level are easy to fix despite the ubiquity of commonly accepted values (95% confidence intervals, for example). Such default values may be a useful starting point but exploration of the effect of different values is to be encouraged. It is undesirable that a decision should be based on the unthinking acceptance of default values for what are, after all, vague parameters. At bottom right is the value for the Dirichlet parameter α.

Of the $m(m\text{-}1)/2 = 210$ pairwise comparisons 82% can be differentiated. This value is shown at the bottom right of the screen and indicates overall discrimination.

The result shown in the chart is fairly clear. There are three clusters. The programmes ranked 1 – 4 are undifferentiated and should be considered as a group, as should the other two clusters. It may be that identifying the leading four is enough to make a short list for further investigation.

6.3 Non-judgemental results

Figure 3 shows results of the two non-informative analyses. When point only weight estimates are given the result is, in this case, not much different from that with the three-point estimates. Although the assessor provides only point values the Dirichlet distribution provides maximum variance estimates, as discussed in Section 5.

It is not surprising that giving no weight information at all greatly reduces discrimination.

7. Aggregation

7.1 General structure

The aggregation of a number of assessments is made using the framework described in (5) and (6) and treating Dirichlet marginal distributions as imprecise assessments. Each assessor p makes n_p imprecise assessments for each weight. The means are weight values $w_{i,p}$ with imprecision $\sigma_{\mathrm{I},i,p}^2$ and so, as in (6), the aggregated estimate is

$$\left.\begin{aligned} w_i &= \textstyle\sum_p \pi_p w_{i,p} \\ \sigma_i^2 &= \textstyle\sum_p \pi_p(\sigma^2_{\mathrm{I},i,p} + \sigma^2_{\mathrm{M},i,p}) \end{aligned}\right\} \qquad (17)$$

For a single assessor $\sigma^2_{\mathrm{M},i,p} = 0$.

The parameter α scales variances and so either the variance of each weight estimate from all sources can be found using (17) and then the Dirichlet fitted, or Dirichlet estimates of the components can be found and aggregated (Section 4.4):

$$\alpha = \textstyle\sum_p \pi_p[\alpha_{\mathrm{I},p} + \alpha_{\mathrm{M},p}] = \sum_p \pi_p[\alpha_{\mathrm{C},p} + \alpha_{\mathrm{V},p} + \alpha_{\mathrm{M},p}] \qquad (18)$$

The α values are a convenient way of comparing the relative contributions of the different sources of imprecision.

7.2 Application to reciprocal matrix estimation

A particular application is when weights are given via a reciprocal matrix. Because of the degrees of freedom in this model both sources of imprecision are present. The Simple Normalised Column Sum analysis is used in which each column is scaled to sum to 1 and treated as a single evaluation.

For a single assessor p (18) becomes

$$\alpha = \alpha_{\mathrm{I},p} = \alpha_{C,p} + \alpha_{V,p} \qquad (19)$$

Each column element is a three-point estimate from which a mean and variance is found. The variance is a measure of vagueness and so the mean of these variances taken across all columns gives the values s_i^2 in (14) from which $\alpha_{V,p}$ is calculated. The variance of the mean estimates measures inconsistency. Using these in (14) these gives $\alpha_{C,p}$.

Nine students made weight estimates using the reciprocal matrix method but giving three-point estimates on the 1-9 scale. As in the previous illustration a 90% confidence interval was used. Table 1 shows the results.

The top section of the table shows the mean weight values for each student. For comparison, the weights used by the *FT* are also shown. The *FT* weights do not closely resemble those of the students.

The middle section of the table shows the Dirichlet parameter values. Of the two sources of imprecision, inconsistency $\alpha_{C,p}$ is, in all cases but one, the greater, usually by a large margin. Students were much more inconsistent than their self-assessed vagueness.

The effect of imprecision is shown in the bottom section of Table 1. Discrimination is quite good if only vagueness due to articulation is used but when inconsistency (model estimation error in this case) is taken into account discrimination is much reduced. Figure 4 shows the effect. The shape of the curve is probably general, the parameters being set by the particular weight estimates and also the similarity of the characteristics of the alternatives. This curve could be used as a descriptor of different decision problems.

7.3 Aggregation of assessors

The students' views as expressed in the weights seem broadly to fall into two groups: first, those for whom internationalisation and employment are most important and, second, those for whom salary and the achievement of their aims predominate. It may therefore make sense to aggregate some assessors into groups, with results shown in Table 2. Discrimination is not high. By far the largest source of imprecision is intrapersonal inconsistency. Interpersonal effects are low for the groups with a salary focus, indicating a coherent group.

The effect for those four students with a salary focus is shown in Figure 5(a). It is hard to see how a cluster of alternatives might be identified from this chart. The level of imprecision is fixed by the students' judgements but the level of support for differentiation, p, is not and the effects of different values should be tested. Figure 5(b) shows that if p=0.5 a preferred group begins to emerge. This would usually be thought a recklessly high value for statistical inference, but the purpose here is different: we wish to see the level of support for discrimination. Kreye et al [77] report that in giving forecasts experts and non-experts gave self-assessed confidence levels of about 40%, so p=0.5 may not be so unreasonable.

The programmes ranked in the top five are the same in all cases.

8. Discussion

The probabilistic approach raises five issues: the specification of probability inputs, the probability model, the interpretation of results, the sources of imprecision, the efficacy of aggregation.

8.1 Probability inputs

In the experiments described here subjects had no uneasiness about giving three-point estimates.

Giving marginal estimates for each weight focuses attention on the individual estimates, just as a strategy of disaggregation is meant to do. The probability that all weights will lie within their specified intervals is less than the confidence level for each. In the illustration (Fig 2) 90% intervals were given so that there was a probability of $0.9^7 = 0.48$ that weights would fall within all limits. This is much less than the confidence level of 0.9 used for each weight specification and may come as a surprise to some users. This would be consistent with research that judgements about conjunctive probabilities are heavily influenced by the assessed probabilities of the constituents [80]. Had users wished this conjunctive probability to be 0.9 then each weight estimate should have been at the confidence level of $0.9^{1/7} = 0.99$. In this particular case, discrimination increases a little from 82% to 84%; the fifth ranked MBA becomes a singleton but otherwise there remain four clusters as shown.

The same issue arises in hypothesis testing [81-84] and also, presumably, in PERT and risk analyses where disaggregated elicitation is the norm. While it is easy to point up the effect there seems no justification for altering the elicitation. Asking users to give confidence levels for each weight interval is likely to be the more comprehensible task.

8.2 The probability model

In specifying Bayesian priors

> "It must be stressed that the assessor has no built-in prior distribution which is there for the taking. That is, there is no "true" prior distribution. Rather, the assessor has certain prior knowledge which is not easy to express quantitatively without careful thought." [85]

When dealing with values (weights) rather than events or facts there is no prior knowledge to be had. This is why the idea of vagueness is so important. It is known from the heuristics and biases literature that people do not behave as probability theory recommends, but neither do they adhere to the predictions of fuzzy theory [86]. Both formalisms have been applied to multiattribute problems.

Probability is preferred because it permits the easy combination of vagueness and imprecision and because of the plausibility of the analogy with sensitivity testing and simulation. If we ask what vague descriptions are for it is reasonable to think they are permissive. Assessors may believe that any values within the range are permissible, sometimes in varying degree, as model input to a sensitivity analysis. It would then be quite natural to see how often different scores were found and to collect these results in a frequency distribution. This immediately leads to a probabilistic interpretation.

Distributions other than the Dirichlet have been used in the analysis of reciprocal matrices and some are given in the references cited in Section 3. The column sum analysis lends itself naturally to the linear model of imprecision.

8.3 Interpretation of results

Statistical significance tests have been used to decide whether two alternatives may be differentiated [59,60] but usually as a precursor to finding the probability of a particular rank. (It should be noted that despite their popularity significance tests are controversial [87].) Here we prefer to show clusters for three reasons. First, the analogy with constructing long lists and short lists may help interpretation and understanding. Second, the idea of clusters is familiar in listings provided by the *Financial Times* and others but the gaps are defined only by the appearance of score differences large in comparison with others in the list; a better support is given in the method here. Third, because the display in Fig 2 encourages comparison between pairs of alternatives.

How best to report the role of probability in discrimination of alternatives will depend on the user. For example, simply saying it is has the same form as a test for statistical significance may be sufficient for those familiar with statistical method, while analogy with repeated sensitivity testing may make more sense for others. Avoiding explicit presentation of standard deviations is advisable [88].

The essential point is that smaller values of p offer more support for discriminating between alternatives. Trying different values and seeing what happens to clusters in Figure 2 will show the robustness of a given clustering.

High values of z in (16) occur if either the difference in scores is high or the imprecision is low or both. Figure 6 shows the comparison between one particular alternative – the target, here the MBA ranked 16 – and the others. The vertical axis shows scores on which to base a judgement about material difference. The horizontal axis shows p, support for any differentiation at all. Using a criterial value, $p = 10\%$, gives four quadrants showing the combinations of statistical support and material difference. Discrimination between 16 and 17, 18, 19 and 20 is unlikely to be supported. For 17 (and 18?) the differences also look unimportant because the scores are close to those of the target.

Considering all 210 pairs, and requiring that all differences be at least 5% as well as that $p < 10\%$ gives only an extra 7 nondifferentiated pairs, reducing discrimination from 82% (Fig 2) to 79%. While discrimination is not much altered, taking explicit account of both material and statistical significance may lead to greater confidence in the final recommendation.

Altering the numerator of (16) to $(|y_a - y_b| - \delta)$, p would be the level of support for the non-zero difference δ. This is standard significance testing and suitable for those familiar with the method. For others, keeping the two elements separate may be a more helpful decision support.

8.4 Vagueness and inconsistency

In the illustration, for eight of the nine assessors inconsistency between weight estimates given in a reciprocal matrix was a greater source of imprecision than vagueness of articulation (Table 1): people were less precise than they thought they were.

In Sec 2.4 correlations between vagueness and imprecision were disregarded. Figure 7 shows that this was justified. We may tentatively suggest that inconsistency (from the estimation process) is a greater source of imprecision than vagueness but that there is no consistent relation between them.

Assessors of probabilities are typically overconfident [89-91] which may be due to anchoring and adjustment [5] as well as a conservatism inherent in some cultures [15]. This means that vague estimates are likely to be optimistic and this may account for the result.

It is usual for inconsistency to be recognised by an index, as in AHP. Measuring inconsistency by Dirichlet parameters provides not only a relative indicator but also, because of the probabilistic approach, shows the effect on discrimination between alternatives. Because the α values are additive the relative impact of different effects are easily shown.

Imprecision can be eliminated entirely, of course, by asking just one assessor to make n-1 point estimates. It is clear from Table 1 that disregarding inconsistency and focusing only on vagueness would give greatly improved discrimination (as Figure 2). The argument for doing so is that a method should provide a heuristic framework for decision support [13], that forcing degrees of freedom is artificial and that the subsequent statistical analysis is "unduly conservative" [92]. But methods with positive degrees of freedom naturally invite a statistical approach; it is hard to think of a reason for ignoring the consequences.

Inconsistency accounts for about 80% of the value of α parameters (Table 1) and so is the main factor determining low levels of discrimination. Inconsistency may come either from making a number of estimates as part of the elicitation method (as here) or from making estimates in different circumstances, as when multiple methods are used [93]. In that sense any particular decision problem is one from a number of such evaluations. As an aid to making a particular decision this does not matter, but there may be an argument for measuring these other effects, if practical considerations permit, further to assess robustness.

8.5 Aggregation and groups

Many decisions are group decisions [94,95], either in some form of decision conference or in distributed teams [96]. This process is likely to comprise a number of meetings between which preferences may alter [97]. It is common that a consensus is formed by feedback and discussion between group members [98] where individual judgements are formed and then shared so that, at the collective level, a joint decision is made [99].

Aggregation of individuals into groups permits the use of just one estimate which may, with care, be treated as a consensus. The aggregation may be by behavioural methods (discussion) or mechanical methods (calculation) [100]. A number of computational methods are available to perform this aggregation and to suggest when plausible groups exist (eg. [101]), but an overall consensus is not always possible, though more discussion may help.

If disparate views are nonetheless aggregated the ability to discriminate between alternatives will fall. This is shown in Table 2. The role of the different sources of imprecision are shown by α values. For the salary focus groups interpersonal effects are the smallest component. This is unlike

findings elsewhere that interpersonal effects are the greater ([92] gives an AHP application) and indicates that members of these groups have made similar assessments; they are coherent. The more frequently reported situation is shown for the group with an internationalisation focus and for all students considered together. Judgements are, at this initial stage, too different to justify a group view. The simple aggregation of parameters (18) permits an easy monitoring.

The subjects in the illustration made individual judgements. The groups suggested in Table 2 are post hoc suggestions.

9. Conclusion

This paper presents a model which makes explicit judgemental inconsistency and vagueness as sources of imprecision and provides for their combination.

Imprecision is described probabilistically. Recognising the difficulties implicit in interpreting probability values this decision aid enables easy interaction with probability inputs and outputs to discourage too great a reliance on commonly accepted values and encourage an assessment of sensitivity.

The Dirichlet parameter α provides a means of comparing whole distributions, not just estimates for one weight, and the various sources of imprecision. For individual assessments this shows that imprecision is the far greater effect than vagueness. This raises the issue of whether methods which permit imprecision to be assessed are to be preferred to those that do not (provided that the imprecision really is assessed and not just described as an index). It is common in statistical modelling to prefer more data from which to infer values and estimation error. Users of a decision aid may prefer only to express vagueness, at most. What to do must be a matter of choice of what best supports a decision for a particular individual or group. It is the firm presumption of the method described here that explicitly acknowledging imprecision is better than ignoring it.

References

[1] D. Timmermans, The impact of task complexity on information use in multi-attribute decision making, J Behav Dec Making 6 (1993) 95-111.

[2] O.F. Morera, D.V. Budescu, A psychometric analysis of the "divide and conquer" principle in multicriteria decision making, Organ Behav Hum Dec 75 (1998) 187-206.

[3] H.R. Arkes, C. González-Vallejo, A.J. Bonham, Y-H. Kung, N. Bailey, Assessing the merits and faults of holistic and disaggregated judgements, J Behav Dec Making 23 (2010) 250-270.

[4] G.A. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information, Psychol Rev 63 (1956) 81-97.

[5] A. Tversky, D. Kahneman, Judgement under uncertainty: Heuristics and biases, Science 185 (1974) 1124-1131.

[6] I. Blanchette, A. Richards, The influence of affect on higher level cognition: A review of research on interpretation, judgement, decision making and reasoning, Cognition Emotion 24 (2010) 561-595.

[7] A. Tversky, D. Kahneman, The framing of decisions and the psychology of choice, Science 211 (1981) 453-458.

[8] P. Curşeu, S. Schruijer, The effects of framing on inter-group negotiation, Group Decis Negot 17 (2008) 347-362.

[9] M. Weber, K. Borcherding, Behavioral influences on weight judgements in multiattribute decision making, Eur J Oper Res 67 (1993) 1–12.

[10] P. Bottomley, J. Doyle, Comparing the validity of numerical judgements elicited by direct rating and point allocation: Insights from objectively verifiable perceptual tasks, Eur J Oper Res 228 (2013) 148-157.

[11] P. Bottomley, J. Doyle, A comparison of three weight elicitation methods: good, better, and best, Omega 29 (2001) 553-560.

[12] S. French, Interactive multi-objective programming: its aims, applications and demands, J Opl Res Soc, 35 (1984) 827-834.

[13] M. Morgan, M. Henrion, Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis, Cambridge, Cambridge University Press, 1990, 53-54.

[14] T. Gilovich, D. Griffin, D. Kahneman, Heuristics and Biases: The Psychology of Intuitive Judgement, New York, Cambridge University Press, 2002.

[15] R. Hogarth, Cognitive processes and the assessment of subjective probability distributions, J Am Stat Assoc 70 (1975) 271-289.

[16] A. Jiménez, A. Mateos, S. Ríos-Insua, Monte Carlo simulation techniques in a decision support system for group decision making, Group Decis Negot 14 (2005) 109-130.

[17] L.C. Dias, J.N. Clímaco, Dealing with imprecise information in group multicriteria decisions: a methodology and a GDSS architecture, Eur J Oper Res 160 (2005) 291-307.

[18] Y. Chen, D.M. Kilgour, K.W. Hipel, A case-based distance method for screening in multiple-criteria decision aid, Omega 26 (2008) 373–383.

[19] A. Jessop, Using imprecise estimates for weights, J Oper Res Soc 62 (2011) 1048–1055.

[20] A.P. Sage, C.C. White, ARIADNE: A knowledge-based interactive system for planning and decision support. IEEE Trans Syst Man Cyber 14 (1984) 35–47.

[21] T. Williamson, Vagueness, London, Routledge, 1994, p67.

[22] S. French, D.R. Insua, F. Ruggeri, *e*-participation and decision analysis, Decision Analysis 4 (2007) 211-226.

[23] R.L. Winkler, Uncertainty in probabilistic risk assessment, Reliability Engineering and System Safety 54 (1996) 127–132.

[24] W. Labov, The boundaries of words and their meanings, in: C-J.N. Bailey, R.W. Shuy (Eds.) New Ways of Analysing Variations in English, Washington D.C., Georgetown University Press, 1973.

[25] E. Rosch, Principles of Categorization, in: E. Rosch, E.E. Lloyd (Eds.) Cognition and Categorization, Hillsdale NJ, Lawrence Erlbaum Associates, 1978.

[26] J. Wang, W. Xu, J. Ma, S. Wang, A vague set based decision support approach for evaluating research funding programs, Eur J Oper Res 230 (2013) 656–665.

[27] C.R. Fox, J.R. Irwin, The role of context in the communication of uncertain beliefs, Basic Appl Soc Psych 20 (1998) 57–70.
[28] S. Haack, Deviant Logic Fuzzy Logic, Chicago, The University of Chicago Press, 1996, 229-255.
[29] J. Barnes, Medicine, experience and logic, in: J. Barnes, J. Brunschwig, , M. Burnyeat, M. Schofield (Eds.) Science and Speculation: Studies in Hellenistic Theory and Practice, Cambridge, Cambridge University Press, 1982.
[30] T.S. Wallsten, D.V. Budescu, A. Rapaport, R. Zwick, B. Forsyth, Measuring the vague meanings of probability terms, J Exp Psychol Gen 115 (1986) 348–365.
[31] R. Jackendoff, Semantics and Cognition, Cambridge Mass., MIT Press, 1983, Ch 7.
[32] T. Chávez, Modelling and measuring the effects of vagueness in decision models, IEEE T Syst Man Cy A 26 (1996) 311-323.
[33] R.T. Clemen, R.L. Winkler, Aggregating probability distributions, in: W. Edwards, R.F. Miles, D. von Winterfeldt (Eds.) Advances in Decision Analysis: From Foundations to Applications, New York, Cambridge University Press, 2007.
[34] T.S. Wallsten, The costs and benefits of vague information, in: R.M. Hogarth (Ed.) Insights in Decision Making: A Tribute to Hillel J Einhorn, Chicago, University of Chicago Press, 1990.
[35] S. Fillenbaum, T.S. Wallsten, B.L. Cohen, J.A. Cox, Some effects of vocabulary and communication task understanding and the use of vague probability expressions, Am J Psychol 104 (1991) 35–60.
[36] M. Cowles, C. Davies, On the origins of the .05 level of statistical significance, Am Psychol 37 (1982) 553-558.
[37] L-Y. Lau, R. Ranyard, Chinese and English speakers' linguistic expression of probability and probabilistic thinking, J Cross Cult Psychol 30 (1999) 411-421.
[38] I. Erev, L. Cohen, Verbal versus numerical probabilities: efficiency, biases, and the Preference Paradox, Organ Behav Hum Dec. 45 (1990) 1-18.
[39] E.M. Johnson, Numerical Encoding of Qualitative Expressions of Uncertainty. Arlington Va., U.S. Army Research Institute for the Behavioral and Social Sciences, 1973.
[40] S. Renooij, C. Witteman, Talking probabilities: communicating probabilistic information with words and numbers, Int J Approx Reason 22 (1999) 169-194.
[41] A.A. Salo, R.P. Hämäläinen, Preference assessments by imprecise ratio statements, Oper Res 40 (1992) 1053–1061.
[42] A.A. Salo, R.P. Hämäläinen, Preference programming through approximate ratio comparisons, Eur J Oper Res 82 (1995) 458–475.
[43] J. Mustajoki, R.P. Hämäläinen, A. Salo, Decision support by interval SMART/SWING – incorporating imprecision in the SMART and SWING methods, Decis Sci 36 (2005) 317–339.
[44] R. Lahdelma, J. Hokkanen, P. Salminen, SMAA – Stochastic multiobjective acceptability analysis, Eur J Oper Res 106 (1998) 117–127.
[45] T. Tervonen, J.R. Figueira, A survey on stochastic multicriteria acceptability analysis methods, J Multi-Crit Decis Anal 15 (2008) 1–14.
[46] R. Lahdelma, K. Meittinen, P. Salminen, Ordinal criteria in stochastic multiobjective acceptability analysis (SMAA), Eur J Oper Res 147 (2003) 137–143.
[47] R. Lahdelma, P. Salminen, SMAA-2: stochastic multicriteria acceptability analysis for group decision making, Oper Res 49 (2001) 444–454.
[48] R. Lahdelma, S. Makkonen, P. Salminen, Two ways to handle dependent uncertainties in multi-criteria decision problems, Omega 37 (2009) 79–92.
[49] H.A. Donegan, F.J. Dodd, T.B.M. McMaster, A new approach to AHP decision-making, J R Stat Soc Ser D Statistician 41 (1992) 295–302.
[50] R. Ramanathan, U. Ramanathan, A qualitative perspective to deriving weights from pairwise comparison matrices, Omega 38 (2010) 228-232.
[51] J. Pang, J. Liang, Evaluation of the results of multi-attribute group decision-making with linguistic information, Omega 40 (2012) 294-301.
[52] L.E. Shirland, R.R. Jesse, R.L. Thompson, C.L. Iacovou, Determining attribute weights using mathematical programming, Omega 31 (2003) 423–437.

[53] R.D. Holder, Some comments on the Analytic Hierarchy Process, J Oper Res Soc 41 (1990) 1073–1076.
[54] M. Tavana, D.T. Kennedy, B Mohebbi, An applied study using the Analytic Hierarchy Process to translate common verbal phrases to numerical probabilities, J Behav Dec Making 10 (1997) 133-150.
[55] R.R. Levary, K. Wan, A simulation approach for handling uncertainty in the analytic hierarchy process, Eur J Oper Res 106 (1998) 116–122.
[56] R. Bañuelas, J. Antony, Application of stochastic analytic hierarchy process within a domestic appliance manufacturer, J Oper Res Soc 58 (2007) 29–38.
[57] D. Hauser, P. Tadikamalla, The Analytic Hierarchy Process on an uncertain environment: a simulation approach, Eur J Oper Res 91 (1996) 27–37.
[58] E.S. Rosenbloom, A probabilistic interpretation of the final rankings in AHP, Eur J Oper Res 96 (1996) 371–378.
[59] E.D. Hahn, Decision making with uncertain judgements: A stochastic formulation of the Analytic Hierarchy Process, Decision Sci 34 (2003) 443-466.
[60] I. Basak, Probabilistic judgements specified partially in the Analytic Hierarchy Process, Eur J Oper Res 108 (1998) 153-164.
[61] G. Manassero, Q. Semeraro, T. Toloio, A new method to cope with decision makers' uncertainty in the equipment selection process, CIRP Annals-Manufacturing Technology 53 (2004) 389-392.
[62] M. Escobar, J. Moreno-Jiménez, Reciprocal distributions in the analytic hierarchy process, Eur J Oper Res 123 (2000) 154-174.
[63] P. Walley, Inferences from multinomial data: learning about a bag of marbles, J R Stat Soc Ser B Stat Methodol 58 (1996) 3–57.
[64] J-M. Bernard, An introduction to the imprecise Dirichlet model for multinomial data, Int J Approx Reason 39 (2005) 123–150.
[65] J.R.W. Merrick, Getting the right mix of experts, Decis Anal 5 (2008) 43–52.
[66] J. Butler, J. Jia, J. Dyer, Simulation techniques for the sensitivity analysis of multi-criteria decision models, Eur J Oper Res 103 (1997) 531–546.
[67] J.B. Kadane, L.J. Wolfson, Experiences in elicitation, Statistician 47 (1998) 3–19.
[68] T.S. Wallsten, D.V. Budescu, I. Erev, A. Diederich, Evaluating and combining subjective probability estimates, J Behav Decis Mak 10 (1997) 243–268.
[69] D.L. Keefer, W.A. Verdini, Better estimates of PERT activity time parameters, Manag Sci 39 (1993) 1086–1091.
[70] J.J Moder, E.G. Rodgers, Judgement estimates of the moments of PERT type distributions, Manag Sci 15 (1968) B76–B83.
[71] L.B. Davidson, D.O. Cooper, A simple way of developing a probability distribution of present value, J Pet Technol (1976, September) 1069–1078.
[72] E.S. Pearson, J.W. Tukey, Approximate means and standard deviations based on distances between percentage points of frequency curves, Biometrika 52 (1965) 533–546.
[73] I.N. Durnbach, T.J. Stewart, A comparison of simplified value function approaches for treating uncertainty in multi-criteria decision analysis, Omega 40 (2012) 456-464.
[74] A. Jiménez, A. Mateos, S. Ríos-Insua, Missing consequences in multiattribute utility theory, Omega 37 (2009) 395–410.
[75] F. Barron, Selecting a best multiattribute alternative with partial information about attribute weights, Acta Pschol 80 (1992) 91-103.
[76] F. Barron, B. Barrett, Decision quality using ranked attribute weights, Manage Sci 42 (1996) 1515-1523.
[77] M.E. Kreye, Y.M. Goh, L.B. Newnes, P. Goodwin, Approaches to displaying information to assist decisions under uncertainty, Omega 40 (2012) 682–692.
[78] D-H. Kim, K-J. Kim, K.S. Park, Compromising prioritization from pairwise comparisons considering type I and II errors, Eur J Oper Res 204 (2010) 285-293.
[79] E. Fernandez, J. Navarro, S. Bernal, Handling multicriteria preference in cluster analysis, Eur J Oper Res 202 (2010) 819-827.

[80] M. Bar-Hillel, On the subjective probability of compound events, Organ Behav Hum Perf 9 (1973) 396-406.
[81] L.M. Bland, D.G. Altman, Multiple significance tests: the Bonferroni method, BMJ 310 (1995) 170.
[82] D. Moran, Arguments for rejecting the sequential Bonferroni in ecological studies, Oikos 100 (2003) 403-405.
[83] L.V. Garcia. Escaping the Bonferroni iron claw in ecological studies, Oikos 105 (2004) 657-663.
[84] T.V. Perneger, What's wrong with Bonferroni adjustments, BMJ 316 (1998) 1236.
[85] R. Winkler, The assessment of prior distributions in Bayesian analysis, J Am Stat Assoc 62 (1967) 776-800.
[86] G. Oden, Integration of fuzzy logical information, J Exp Psychol Human 3 (1977) 565-575.
[87] S.T. Ziliak, D.N. McCloskey, The Cult of Statistical Significance: How the standard error costs us jobs, justice and lives, Ann Arbor, The University of Michigan Press, 2008.
[88] I.N. Durbach, T.J. Stewart, An experimental study of the effect of uncertainty rpresentation on decision making, Eur J Oper Res 214 (2011) 380-392.
[89] G. Keren, On the calibration of probability judgements: Some critical comments and alternative perspectives, J Behav Dec Making 10 (1997) 269-278.
[90] B. Fischoff, P. Slovic, S. Lichtenstein, Knowing with certainty: The appropriateness of extreme confidence, J Exp Psychol Human 3 (1977) 552-564.
[91] D. Griffin, A. Tversky, The weighing of evidence and the determinants of coincidence, Cognitive Pschol 24 (1992) 411-435.
[92] J.M. Alho, J. Kangas, O. Kolehmainen, Uncertainty in expert predictions of the ecological consequences of forest plans, Appl Stat 45 (1996) 1–14.
[93] R.L. Winkler, R.T. Clemen, Multiple experts vs. multiple methods: combining correlation assessments, Decis Anal 1 (2004) 167–176.
[94] E. Rokou, K. Kirytopoulos, A calibrated group decision process, Group Decis Negot, published online 2013.
[95] U. Bose, D, Paradice, The effects of integrating cognitive feedback and multi-attribute utility-based multicriteria decision-making methods in GDSS, Group Decis Negot 8 (1999) 157-182.
[96] A. Barcus, G. Montibellar, Supporting the allocation of software development work in distributed teams with multi-criteria decision analysis, Omega 36 (2008) 464–475.
[97] G. Lockett, P. Naudé, The stability of judgemental modelling: an application in the social services, Group Decis Negot 7 (1998) 41-53.
[98] F. Bezerra, P. Melo, J.P. Costa, Visual and interactive comparative analysis of individual opinions: a group decision support tool, Group Decis Negot, published online (2013).
[99] J.P. Costa, P. Melo, P. Godinho, L.C. Dias, The AGAP system: a GDSS for project analysis and evaluation, Eur J Oper Res 145 (2003) 287-303.
[100] A. Wilson, Cognitive factors affecting subjective probability assessment, ISDS Discussion Paper #94-02, Durham NC, Duke University, 1994.
[101] P. Gargallo, J. Moreno-Jiménez, M. Salvador, AHP-group decision making: A Bayesian approach based on mixtures for group pattern identification, Group Decis Negot 16 (2007) 485-506.

	student assessor *p*									
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*FT*
Weights (%)										
Salary increase	8	13	17	18	30	31	34	36	41	54
Aims achieved	9	11	19	36	10	14	25	26	21	8
Employed at 3 months	29	25	17	14	31	32	17	17	16	5
Women faculty	4	4	4	2	2	3	3	2	3	5
Women students	3	3	12	4	6	6	6	3	4	5
International faculty	22	21	3	6	4	5	4	7	7	11
International students	24	24	28	20	17	9	11	10	8	11
Dirichlet parameters										
inconsistency, $\alpha_{C,p}$	0.277	0.028	0.101	0.080	0.054	0.036	0.057	0.168	0.097	
vagueness, $\alpha_{V,p}$	0.014	0.009	0.011	0.028	0.063	0.011	0.015	0.015	0.016	
$\alpha_{I,p} = \alpha_{V,p} + \alpha_{C,p}$	0.291	0.037	0.113	0.109	0.117	0.047	0.072	0.183	0.112	
$\alpha_{cC,p} / \alpha_{I,p}$ (%)	95	76	89	73	46	77	79	92	87	
Discrimination (%)										
using $\alpha_{C,p}$	6	59	25	37	48	55	38	11	23	
using $\alpha_{V,p}$	74	78	72	59	46	71	59	57	56	
using $\alpha_{I,p}$	6	54	23	30	27	48	33	10	20	

Table 1. Evaluation by nine students using reciprocal matrix.

group	salary focus			internationalisation focus	all
	e,f	*g,h*	*e,f,g,h*	*b,c,d*	
α values					
inconsistency, α_C	0.045	0.113	0.079	0.070	0.100
vagueness, α_V	0.037	0.015	0.026	0.016	0.020
interpersonal, α_M	0.002	0.000	0.013	0.025	0.038
total, α_I	0.084	0.128	0.118	0.112	0.158
relative contribution					
inconsistency, α_C	54	88	67	63	63
vagueness, α_V	44	12	22	14	13
interpersonal, α_M	2	0	11	23	24
total, α_I	100%	100%	100%	100%	100%
Weights					
Salary increase	30	35	33	16	25
Aims achieved	12	25	19	22	19
Employed at 3 months	31	17	24	19	22
Women faculty	3	2	3	3	3
Women students	6	4	5	6	5
International faculty	4	6	5	10	9
International students	13	10	11	24	17
discrimination %	36	18	24	20	10

Table 2. The effects of aggregation.

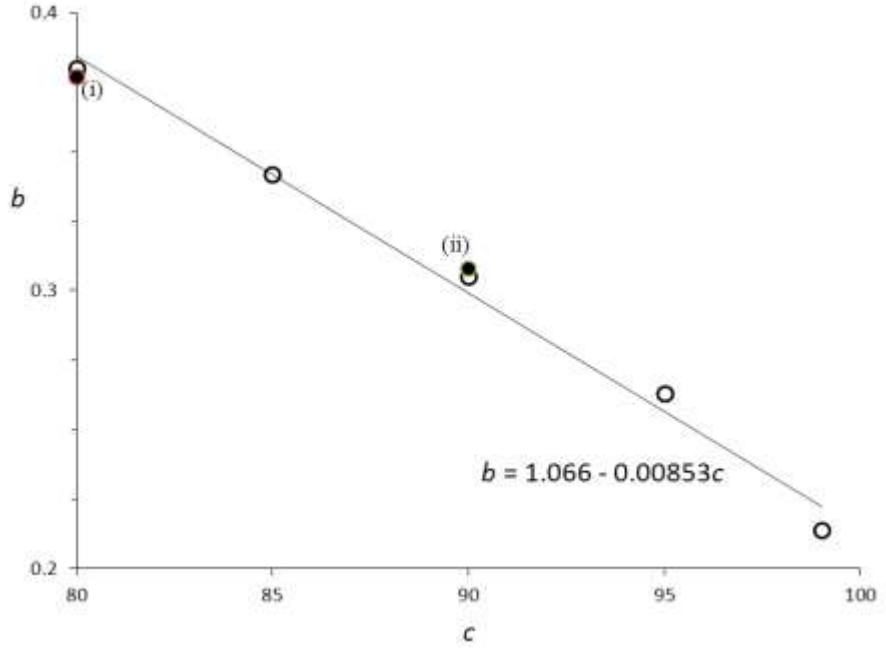


Figure 1. Relation between b and c for inferring standard deviation from L and H.
Note some previous estimates of b: (i) 0.377 [67,68] ; (ii) 0.308 [69]

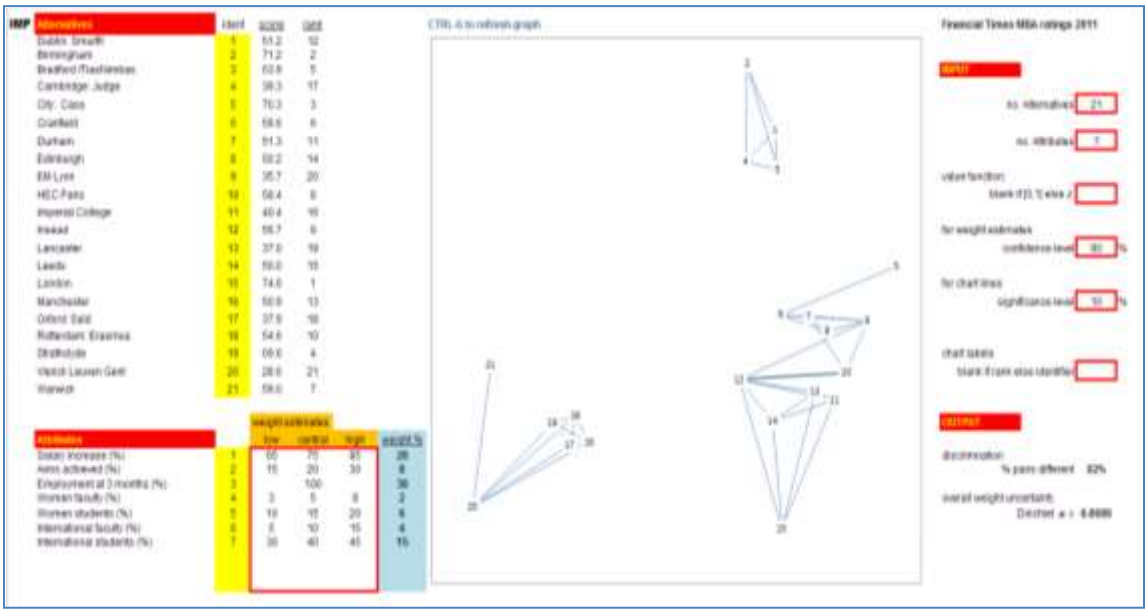
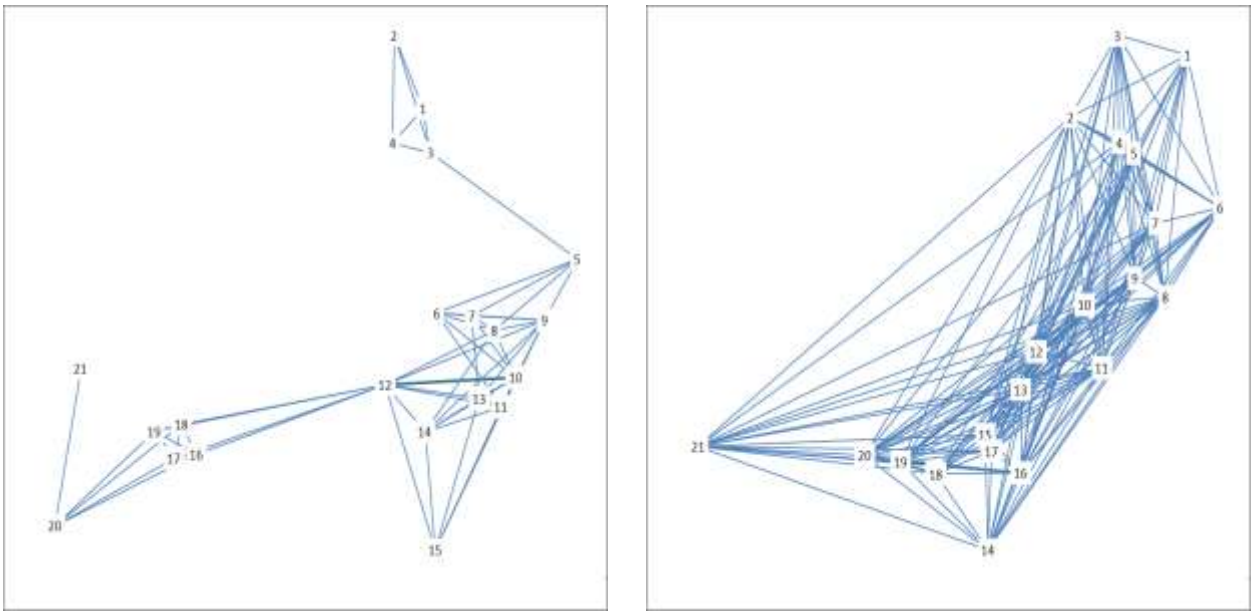


Figure 2. Screen design for interactive model.



(a) single point weight estimates
α = 0.0185; discrimination = 72%

(b) no weight estimates
α = 0.1247; discrimination = 16%

Figure 3. Results for non-informative analyses for a single user (p = 10%).

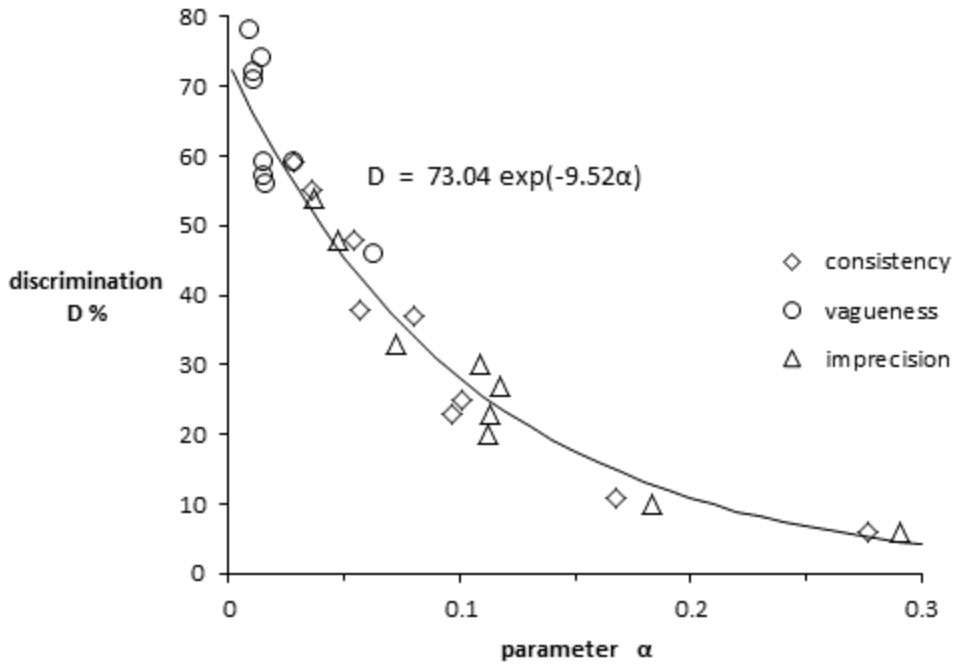
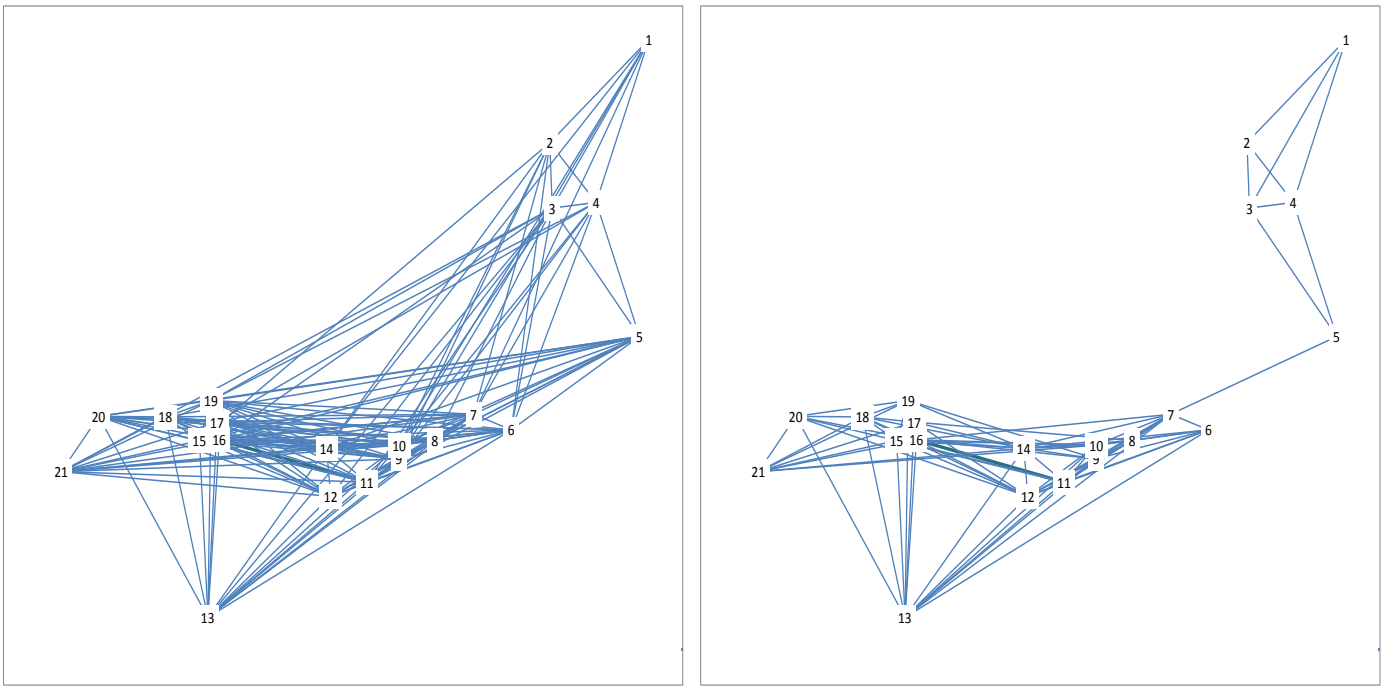


Figure 4. Relation between discrimination and Dirichlet parameter α (Table 1).



(a) $p = 0.1$; discrimnation = 24%

(b) $p = 0.5$; discrimnation = 57%

Figure 5. Results for aggregated results: salary focus (Table 2, students *e*–*h*).

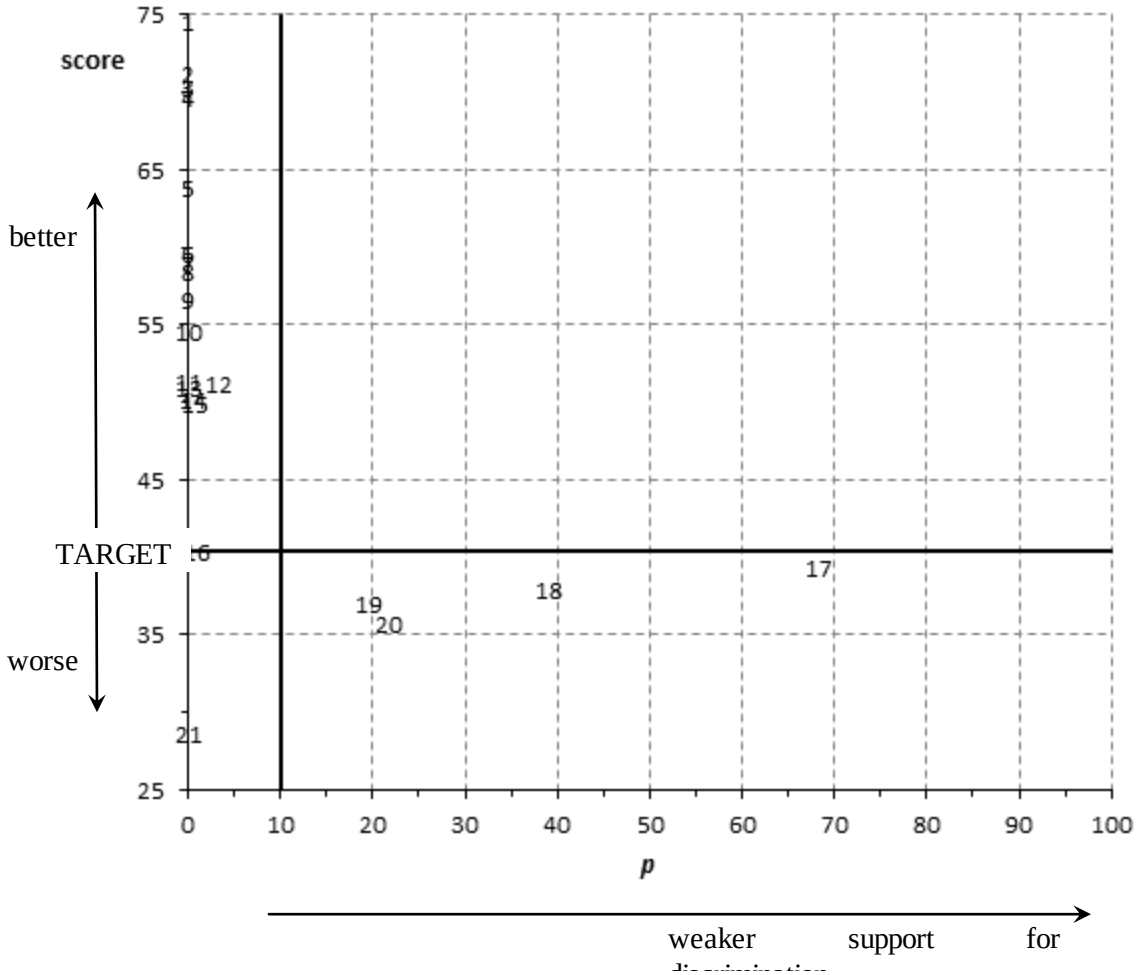


Figure 6. Identifiers show rank. Target = 16, from data in Figure 2. (based on Figure 1.1 of [87])

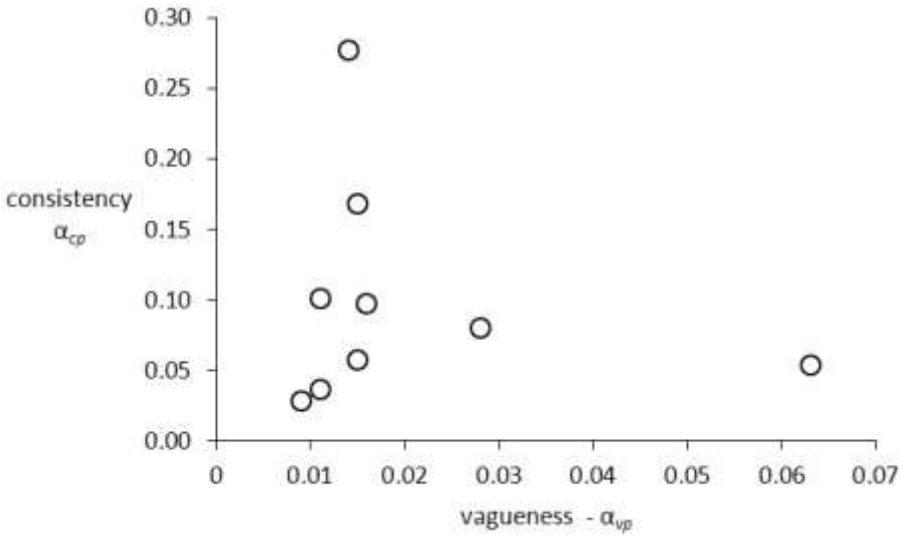


Figure 7. Dirichlet parameters for consistency and vagueness assessments: $r^2 = 0.03$ (Source: Table 1)