2016 49th Hawaii International Conference on System Sciences

How Not to Select Ideas for Innovations: A Critique of the Scoring Method

Graham Horton, Jana Goers, Stefan Werner Knoll
Computer Science Department
University of Magdeburg, Germany
graham.horton@ovgu.de

Abstract

We are interested in the problem of evaluating and selecting one or more ideas to be pursued in the idea stage of the innovation process.

The so-called scoring method, which is based on Multi-Attribute Utility Theory is very commonly used for this task. We know of many corporations that select their innovation projects using this technique.

We present original arguments from three different sources that illustrate some of the severe difficulties associated with this method: theoretical considerations, simulation experiments and laboratory studies.

We conclude that the scoring method should not be used for this task and consider instead the little-known lexicographic approach, which does not suffer from the disadvantages of the scoring method. We suggest that the lexicographic method may be the method of choice for the given application.

1. Introduction

1.1. Idea Evaluation in the Innovation Process

Corporations use a structured process to discover and develop innovations for products, services and business models. This innovation process is typically divided into an idea stage, in which ideas are generated and evaluated, and an implementation stage, in which the future products, services etc. are developed and launched [5].

At the beginning of the idea stage, ideas are only superficially described and no information about them has yet been collected. These raw ideas have to be filtered quickly, since there may be a very large number of them, and typically a simple GO/NO-GO decision is made. On the other hand, at the end of the idea stage, the number of ideas will have been considerably reduced, and they will have been thoroughly researched and documented. Here, evaluation and selection can be very involved, since the decision to proceed to the implementation phase can be very expensive.

The goal of each step of the idea phase is to reduce the number of ideas under consideration by adding appropriate evaluation criteria, gathering data and eliminating ideas that fail to meet the criteria. We concur with Girotra et al [13] that *For an organization interested in the quality of the best identified ideas, the fidelity of the evaluation process it employs is thus crucial.*

One very commonly used approach is the so-called scoring method (also known as utility analysis or decision matrix). By (implicitly) assuming the existence of a cardinal utility function, a simple procedure is obtained, which is transparent and easy to execute using spreadsheet software. The authors know several international corporations that use scoring models for evaluating innovation ideas. However, the method has several disadvantages, some of which are based on theoretical considerations, others of which are of a more psychological or practical nature. In the authors' experience, practitioners are not aware of these limitations, which can lead to substantial hidden risks.

Lexicographic methods were developed many decades ago and have been applied to various decision-making problems, most often supplier selection or purchasing decisions. These methods do not suffer from many of the disadvantages of the scoring method. Nevertheless, they are less widely used, and to our knowledge they have never been applied to selecting innovation projects.

1.2. Goals and Overview

In the next section, we provide the background for the paper, including descriptions of the scoring and lexicographic methods and a discussion of the idea selection task in the innovation context.

In the subsequent sections, we describe our laboratory and simulation experiments. In the former case, subjects solved an idea evaluation task using both

1530-1605/16 $31.00 © 2016 IEEE
DOI 10.1109/HICSS.2016.37

methods and gave subjective feedback on each. In the latter case, we compute quantitative data about each method.

We cite theoretical objections to the scoring method and present two new arguments concerning its use of the arithmetic mean. The median is shown to be preferable to the mean in both cases.

We conclude that for the task of selecting innovation projects – especially in groups – the scoring method has serious drawbacks, and we propose that a lexicographic approach might be more appropriate.

In addition to their roles as academic scientists, the first two authors are also innovation consultants and are able to draw on their experience with many large corporations. The goal of the studies described in this paper is therefore not only to make a scientific contribution but also to make well-founded recommendations to practitioners and thus enhance the quality of idea evaluation in the real world.

2. Background

2.1. General

The task to be solved is to evaluate a set of proposals for innovation projects according to a given set of criteria. The goal of this evaluation is to obtain a complete or partial ordering of the alternatives in order to select one or more of them for further development. The task thus belongs to the realm of Multi-Criteria Decision-Making (MCDM). MCDM is rich scientific area which has spawned a large amount of theory and many algorithms. A survey of the field can be found in [9].

Many methods proposed in the literature require a substantial amount of mathematical manipulation, and are therefore intransparent and reliant on computer support. The best-known example of such a method is the Analytical Hierarchy Process by Saaty [21], which requires the computation of the eigenvectors of a preference matrix.

When the evaluation task is carried out by a group, hidden profiles [25] become an issue. A hidden profile is present when group members have different mental representations of the ideas or the evaluation criteria. These can lead to differing individual evaluations, which then have to be resolved (or else tolerated). Horton and Görs [14][15] have shown that hidden profiles can be successfully resolved, which in most cases leads to a unanimous evaluation vote.

Briggs et al have introduced a useful six-layer model for collaboration projects [1] which is used by the authors of this paper in a modified form for their own consulting projects. Stated in the Briggs terminology, one goal of the work described in this paper is to develop a new and better Technique for an innovation project toolset.

Kolfschoten et al [18] have introduced a classification of so-called Patterns of Collaboration for categorizing collaborative activities. In their terminology, the Technique we propose falls into both the Evaluate and Reduce categories. Of particular interest is that the lexicographic method achieves a reduction in the number of alternatives on the basis of (what appears to be) only a partial evaluation. It is thus very efficient.

MCDM methods are designed for situations which differ from innovation project selection in three important respects. Firstly, the properties of the alternatives are outside the influence of the decision-maker. Secondly, the decision, once made, ceases to occupy the decision-makers attention. Finally, the properties of the alternatives may compensate each other. A good (and typical) example is a purchasing decision for a computer, where memory capacity, processor speed, screen resolution and price are all known and fixed. A decision-maker may consider a trade-off between memory and speed, and will purchase the alternative that represents the best compromise between all criteria. In any case, once the decision has been made, it can be forgotten and the outcome can no longer be influenced.

2.2. Selecting Innovation Projects

Innovation project selection is typically a group decision made in a meeting that includes executive-level management or business unit leaders, as well as product, innovation and business development managers. Thus the issues of transparency, hidden profiles and aggregation of judgments are concerns. This meeting may last two or three hours and may be facilitated by an internal or external innovation expert. The result of the meeting is a prioritization of the projects into the categories GO and NO-GO (and perhaps also REWORK). GO decisions may lead to the commitment of significant resources.

The selection task for innovation projects differs from the typical situation for which MCDM methods are designed: the decision-makers are themselves responsible for the implementation of whichever projects they select, and the properties of the alternatives (such as profitability) are subject to a significant degree of uncertainty.

We propose that attributes of ideas for innovation projects should not be compensatory, i.e. a weakness (or strength) with respect to one criterion should not be allowed to "balance" a strength (weakness) with respect to a different criterion. A useful analogy is

the decathlon event in athletics. In order to win the event, a decathlete's overall score is a combination of partial scores in ten different sub-events, and a strength in one discipline "compensates" a weakness in another. By contrast, the winners of the individual, specialized events such as discus or javelin will usually be better in that discipline than the decathlete.

Innovation projects are often initiated with specific goals in mind. These may derive from the current market situation or from corporate strategy. Some examples from the authors' own consulting experience are…

- *What should our next-generation product be that maintains our technological leadership?*
- *With what new service offer can we earn 50 million Euros per year within the next five years?*
- *How can we improve our chances of winning the forthcoming Request For Tenders?*
- *We want to develop patentable inventions in order to protect our market position.*

In each case, the leading evaluation criteria arise naturally from the project goal. We propose that the ideas that best fulfil these criteria (and do not violate any must-have conditions such as budget or strategic fit) are the ones that should be selected. In our experience, groups of experts (when not using a formal method) will do just this. In other words, we propose that winning innovation ideas should be high-jump or javelin specialists, and not decathletes.

2.3. Utility and Scale

Utility is an artificial construct that is used primarily in mathematical economics to model choice. It is intended to represent the benefit that is achieved by making a particular selection: given a choice, a purchaser will select the alternative that maximizes his or her utility.

The theory of measurement describes different types of scale. In order of increasing modeling power, these are *nominal*, *ordinal*, *interval* and *ratio*, whereby the latter two are known as cardinal scales. A nominal scale is essentially equivalent to categories (*French*, *German*, *Italian*), whereas an ordinal scale allows comparisons (*better than*, *equivalent to*, *worse than*). An interval scale such as the Likert scale often used in surveys or the Celsius scale of temperature uses numerical values, but there is no zero, and multiplication and division are not defined. A ratio scale is continuous, has the notion of zero and allows all of the well-known arithmetic operations. Most physical measures (*length*, *mass*, *velocity*) are ratio scales. The arithmetic mean is defined only on the ratio scale; the median is defined on both cardinal and ordinal scales.

Both cardinal and ordinal utilities have been proposed, and there is a large body of literature comparing and contrasting the two. The question of the appropriateness of each in economic decision-making is at least 60 years old [24].

For both theoretical and practical reasons, cardinal utilities are now hardly used in economics. One textbook [22] states unequivocally: *Economists today generally reject the notion of a cardinal, measurable utility*. The main objections against cardinal utility are that the utility values are without meaning and that complicated preference elicitation techniques are required to determine them.

Despite this, the widely-used scoring method is built on cardinal, measurable utilities, whereas the lexicographic method only requires the ordinal scale.

2.4. Performance and Sufficiency Criteria

In practice, evaluation criteria fall into two categories, which we will denote as performance criteria and sufficiency criteria.

Sufficiency criteria are those for which a minimum level must be achieved. One common example in innovation is strategic fit: *Does the proposed innovation fit into our business strategy?* Sufficiency criteria have a pass/fail nature. Performance criteria are those for which hold: *The more, the better*. Examples are *gain in market share*, *increase in sales* or *amount of media coverage*.

In our experience, criteria that should be treated as sufficiency criteria are often treated as performance criteria. One of our clients, a multi-billion international corporation, evaluates its innovation projects with a 41-criterion scoring model using a five-point Likert scale. This catalogue includes both criteria that are clearly of performance type such as *size of market* and *profit potential* and criteria that we contend should be of type sufficiency such as *Do we have the right organizational structure?* and *Are there any hurdles to implementation?* We claim that measuring the appropriateness of organizational structure on a five-point scale and then combining the result with profit potential is inappropriate and could easily result in less-than-optimal choices.

We suggest that it would be both conceptually preferable and more efficient to first establish minimum levels for as many criteria as possible and use these to filter out unsatisfactory alternatives. The remaining alternatives could then be more easily compared on the basis of performance criteria alone. This assumes, of course, that a clear goal has been established for the innovation project.

2.5. Scoring Method

The scoring method (also known as utility analysis and decision matrix) is a method for evaluating a set of alternatives according to a set of criteria. It is the simplest such method based on Multiple Attribute Utility Theory (MAUT) [7]. MAUT assumes the existence of a cardinal utility function that measures the value of each alternative with respect each criterion. Since multiplication is used by the method, the utility must be defined on a ratio scale. This assumption is the source of much of the controversy surrounding the method.

The scoring method is widely used in practice, not only for decisions such as *Where should we build the new hospital?* or *Which supplier should we select?* but also for innovation project selection. It is also commonly seen in scientific publications. For example Kolfschoten et al [19] recommend in their Guideline 14: *For complex quality assessments use a multi criteria decision matrix to capture scores and enable rapid calculation of group assessments.*

We denote the number of alternatives by m, the number of criteria by n and the number of decision-makers (DM) by d. The scoring algorithm assigns weights w_i to the criteria C_i, $1 \le i \le n$. These weights are used as scaling factors in order to assign different degrees of importance to each criterion. Then, numerical evaluations e_{ij} are awarded to each alternative A_j, $1 \le j \le m$ with respect to each criterion C_i. The total scores are then given by $\Sigma w_i\, e_{ij}$ and the alternative with the highest total is deemed to be the best. Both weights and judgments are made on a cardinal scale, typically integers between 1 and 5 or 1 and 10.

Table 1 illustrates a scoring table with n=3 criteria and m=4 alternatives. Criterion C_1 is the most important with a weight of 4, followed by C_3 and then C_2. The best alternative in this example is A_4 with 4x4+2x2+3x3=29 points.

Table 1. Example Scoring Table

			Alternatives			
		Weights	A_1	A_2	A_3	A_4
Criteria	C_1	4	3	1	2	4
	C_2	2	4	2	4	2
	C_3	3	2	5	3	3
Totals			26	23	25	29

One requirement of the method, which follows from Multi-Attribute Utility Theory, is that the criteria be independent of each other. This is seldom true in practice, in particular for innovation projects. For example, *ease of implementation* almost by definition correlates negatively *with degree of innovativeness*.

The scoring method requires the mapping of subjective evaluations to numerical values. Estimates of uncertain attributes such as future profitability or ease of implementation must be mapped by the decision-maker to a number. This has been shown to be both complex and inaccurate [9][2].

A second requirement of the method is commensurability of the criteria. This is a direct consequence of cardinal utility: The judgments awarded are all measured in the same artificial "currency", sometimes referred to as *utils*. In addition, the numerical values must have the same meaning across all criteria: 4 points for the criterion *size of market* have the same meaning as 4 points for *strategic fit*. Furthermore, when including the criteria weights, 2x2 points for *size of market* carry the same impact on the overall result as 1x4 points for *strategic fit*. Both consequences are highly questionable and furthermore almost impossible to achieve in practice.

Scoring belongs to the class of *compensatory methods*, i.e. those in which scores in different criteria can cancel each out. This is the "decathlon" assumption that has already been mentioned.

In the group context, three approaches can be used to merge individual opinions: *Enforced unanimity*, *means of judgments* and *means of totals*. In the first case, a discussion is carried out for each weight and judgment in order to arrive at a single value for each that represents the group consensus. In the second and third cases, decision-makers carry out the method individually and their results are aggregated computationally. Either the means of the judgments and weights are used to compute the totals or the means of the individual totals can be used. We show in Section 3 that this can be inappropriate.

2.6. Lexicographic Method

The lexicographic method corresponds to the alphabetic sorting of words or phrases. It can be carried out using only pairwise comparisons, which is known to be easier than awarding points and closer to humans' intuitive method of comparison [17]. (By contrast, awarding points is perceived to be unnatural: *There are reasons to doubt that the linear value-maximization model accurately reflects the behavior of decision makers in complex multi-attribute choices* [4]).

The method proceeds by first ordering the criteria in order of priority. Then, the alternatives are ranked with respect to the highest-priority criterion. (Only) those alternatives which have not yet have received a

final ranking are then ranked according to the next-highest criterion. This is repeated until all alternatives have received a final rank.

The method thus has a "dictator" criterion: ranks that are awarded on the basis of the top-priority criterion cannot be subsequently modified on the basis of other criteria. There is thus no possibility of compromise, and the lexicographic scheme belongs to the class of so-called *non-compensatory* methods.

The method only needs the much simpler ordinal scale. It only requires ordinal (as opposed to cardinal) utilities, which were developed extensively by Fishburn [11].

Table 2 shows an example with n=3 criteria and m=4 alternatives. Criterion C_3 has the highest priority (value=1), followed by C_1 and lastly C_2.

The algorithm begins with the most important criterion C_3 and all as yet unranked alternatives are ranked with respect to it. In this example, A_1 and A_4 are given the highest ranking and alternatives A_2 and A_3 are tied for the second rank.

Since the ranking is not yet complete, the next most important criterion C_1 is considered. The ranks $A_4 > A_1 > A_3 > A_2$ are awarded and the procedure is complete. The final rankings have been established without referring to Criterion C_2 at all.

The method can clearly be executed using only pencil and paper – not even basic arithmetic operations are required.

Table 2. Example Lex Table

			Alternatives			
		Prio.	A_1	A_2	A_3	A_4
Criteria	C_1	2	2	4	3	1
	C_2	3	-	-	-	-
	C_3	1	1	2	2	1
Ranking			2	4	3	1

3. Medians or Means?

In the case of group decisions (which is almost always the case for innovation projects), the scoring model uses the arithmetic mean to aggregate individual evaluations. By contrast, the lexicographic method would naturally use the median, since the mean is not defined in an ordinal scale.

In the following two subsections, we prove that the median provides a better representation of the group opinion and is more robust with respect to inaccuracies in individual evaluations and conclude that it should replace the mean in the scoring method.

3.1. Aggregation of Votes

We first prove that the median is the optimal choice for the aggregation of individual evaluations in the sense that it would receive more votes in an election than any other evaluation.

Let e_k, $1 \le k \le d$ be the individual decision-makers' evaluations for any alternative and criterion and s_k be the corresponding values of those judgments. Assume for simplicity that d is odd. Let s_{med} denote the evaluation that corresponds to the median. Define three sets of evaluations as follows:

$$E_l = \{e_k : s_k < s_{med}\}$$
$$E_{med} = \{e_k : s_k = s_{med}\}$$
$$E_u = \{e_k : s_k > s_{med}\}$$

(l stands for lower, u stands for upper.) The following inequalities follow from the definition of the median:

$$|E_l| + |E_{med}| > |E_u|$$
$$|E_u| + |E_{med}| > |E_l|$$

Clearly, the decision-makers who voted in E_l and E_{med} will prefer s_{med} to any evaluation $\in E_u$ and conversely, those who voted in E_u and E_{med} will prefer s_{med} to any evaluation $\in E_l$. Thus, the most preferable evaluation is the median.

Table 3 shows an example. In the upper section, the evaluations awarded by five decision-makers (DM) are shown. Permissible values are integers between 1 and 5. The mean is 2 and the median is 1. The lower section of the table shows the votes of the five decision-makers on all possible combinations of preferences. The median wins with 9 votes.

Table 3. Votes for median and mean

k	1	2	3	4	5
e_k	1	1	1	3	4

DM prefers		to this value				
		1	2	3	4	Σ
this value	1		3	3	3	9
	2	2		3	3	8
	3	2	2		4	8
	4	2	1	1		4

3.2. Robustness

For innovation projects it is reasonable to assume that decision-makers can make different evaluations owing to the presence of hidden profiles and to the uncertainties inherent to the task.

We therefore consider the following experiment: We first obtain evaluations from decision-makers, allow for each to vary by ±1 points or ranks and determine the effect of this variability on the overall result.

If we are only interested in the winning alternative, it is fairly easy to see that if the median is used to aggregate the ranks, then only the runner-up can advance to the winning position.

As an example we again use the five evaluations from Table 3 (upper). Now, the first three DMs can give values of 1 or 2, the third can give 2, 3 or 4 and the last can give 3, 4 or 5. This yields a total of 72 combinations. Table 4 (lower) shows how these are distributed among the ranks: only ranks 1 and 2 are possible. By contrast, using means, every change will affect the overall result, and the same set of variations can yield means from 1.8 up to 3.2.

Table 4. Robustness

means	1.8	2.0	2.2	2.4	2.6	2.8	3.0	3.2
count	1	5	12	18	18	12	5	1

medians	1	2	3	4	5
count	9	63	0	0	0

We therefore conclude that the median is more robust with respect to evaluation uncertainty than the mean.

4. Simulation Experiments

We created Monte Carlo simulations [1] of the lexicographic and scoring methods. In the case of scoring, we were interested in the effect of inaccuracies in the scores on the winning alternative. For the lexicographic method, we were interested in the amount of data needed to complete the analysis.

For our simulation of the scoring method, we assume that individual scores e_{ij} are given on a five-point integer scale $1 \le e_{ij} \le 5$. We then populated the scoring table with random values and determined the winning alternative. Then random manipulations of the scores were made, allowing each to change by ±1 (changes in end-of-range scores of 1 and 5 were limited to permissible values as appropriate). Finally, the number of alternatives whose total weighted scores that now surpass the original winner was counted. This procedure was repeated 10,000 times using independent random numbers in order to obtain a high level of statistical significance.

Table 5 shows the results of this experiment for different problem sizes ($6 \le n=m \le 22$). Shown is the probability p_i that the winner of the scoring method is invincible (more precisely, it is the proportion of replications in which the originally winning alternative was not overtaken by another). In addition, the α=0.05 confidence intervals are given.

For the smallest problem, the probability is only 0.64, and this value decreases even further as the number of alternatives and criteria increases.

Table 5. P(Winner is unique)

m, n	6	10	14	18	22
p_i	0.64	0.57	0.51	0.50	0.47
C.I.	0.026	0.023	0.021	0.020	0.019

Figure 1 shows a breakdown of the data for a problem of size m=10, n=6. Shown are the probabilities for the size of the winning set, i.e. the number of alternatives that can become winners as a result of variations of ±1 in the scores. The probability that the winner is unique is only 0.58, and the probability decreases as the size of the winner set increases. Even for the extreme case, where all 10 alternatives can become the winner, the probability is non-zero.

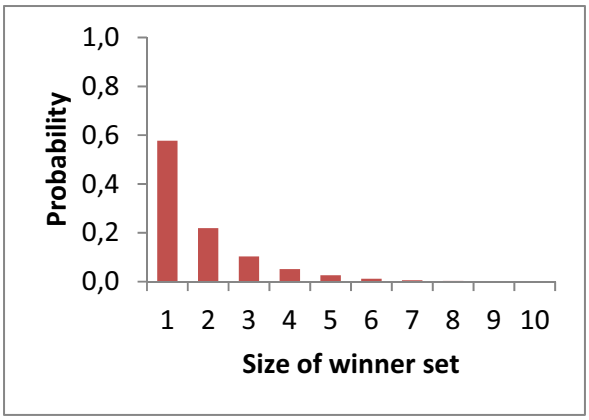

Figure 1. Distribution of winner set size

This has significant consequences for management practice. If the assumption is valid that an inaccuracy of ±1 in the scores and weights is to be expected, then the winning alternative produced by the scoring method is highly uncertain. There are many indications – from the scientific literature, from the experiment described in the next section and from

our own consulting experience – that this assumption is indeed justified, and perhaps even excessively conservative.

A Monte Carlo simulation of the lexicographic method was built to determine the number of evaluations needed to complete the algorithm. The independent variable in this case is p_u, the probability that any given rank is unique in its row of the table. The simulation performed the lexicographic method using random values for the judgments and 10,000 independent replications were made in order to obtain a high statistical significance.

Table 6 presents results obtained from the simulation of a task of size $n = m = 6$. Shown are the mean values for the number of data points needed together with the α=0.05 confidence intervals.

Table 6. # of Lex data points (p_u)

p_u	1.0	0.8	0.6	0.4	0.2	0.0
Data	6.00	8.26	11.8	17.6	26.4	36.0
C.I.	0.00	0.09	0.15	0.21	0.22	0.00

For p_u =1, the first row of the evaluation table will already contain a complete order and no further effort is required. For p_u =0, no distinction between ranks is made, and the table is filled completely (# data points = n x m=36). Intuitively, the amount of data required will decrease as p_u increases. Our experience is that p_u is approximately equal to 0.5 in practice; the empirical experiment for a 5x5 evaluation task (see Section 5.3) yielded p_u =0.4.

Table 7 shows analogous data for varying problem sizes where n=m and p_u was set at 0.5. As before, 10,000 independent replications were used. The number of data points D is shown divided by n (i.e. it counts table rows). The α=0.05 confidence intervals are also shown. It shows that the amount of data needed (measured in table rows) approaches an asymptotic value of approximately 2.6 rows of data.

Table 7. # of Lex data points (n,m)

n, m	4	8	12	16	20
D	2.15	2.48	2.55	2.59	2.60
C.I.	0.029	0.029	0.025	0.022	0.020

The important implication of these results is that for any reasonable value of p_u, the effort required for the evaluation is linear in the number of alternatives m only. By comparison, the scoring method always requires the entire n x m table to be filled.

5. Empirical Experiment

We carried out lab experiments in order to obtain subjective impressions about each method.

5.1. Experimental Setup

Five proposals for innovations in the Computer Science program at a research university in Germany were generated (m=5), together with five evaluation criteria (n=5). The goal provided was that the proposal should "strengthen the department" (which is deliberately open to interpretation). Worksheets were prepared for each method.

Experimental subjects were 30 students (16 female, 14 male) from Bachelor and Master programs. Each method was explained to the participants, who were then given as much time as needed to carry out each method. Subsequently, participants were given a questionnaire containing six statements to be evaluated using a five-point Likert scale. The experiment was within-subjects, and participants completed all tasks individually. No control for ordering effects was performed.

The proposals, criteria and feedback statements are given in the Appendix in Section 8.

5.2. Scoring Method

The variance in the scores was very large: Of the mxnx5=125 possible individual evaluations, only 8 values were unused and of the nx5=25 possible criteria weights, only 4 were unused. The 30 subjects generated 27 distinct rankings, none of which corresponded to the one generated using the mean values.

The mean criteria weights range only from 3.3 to 4.2 (in a possible range of 1 to 5). This makes them appear very similar subjectively. Similarly, in the overall result shown in Figure 2, the overall scores are very close; the winner and runner-up are separated by only 1 point in 125, which can lead to an (unjustified) impression of precision.

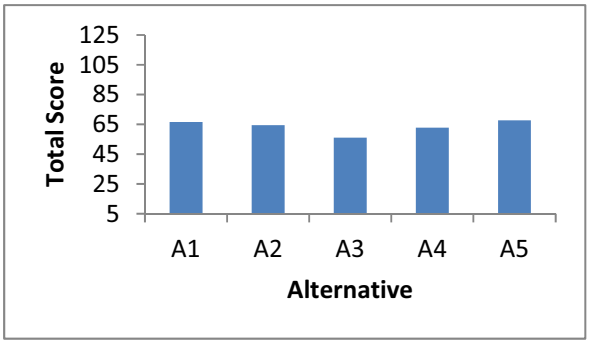

Figure 2. Overall Scores

5.3. Lexicographic Method

Using the lexicographic method, the 30 subjects generated 22 different rankings.

We were interested in the number of entries needed to complete the lexicographic procedure. Table 8 shows possible numbers of data points D needed and the frequency of occurrences O among the 30 subjects. The mean number of data points was 9.7, which corresponds to two rows of the matrix. The median value was 7, i.e. half of the participants completed the exercise with at most 7 evaluations. The frequency of rank duplications was $p_u = 0.4$.

Table 8. # of data points (lab)

D	5-7	8-10	11-13	14-16	17-19
O	17	8	6	1	1

By comparison, the number of data points for the scoring method is 25; the lexicographic method thus has only about 40% of the cost of the scoring method. Assuming that p_u is independent of the number criteria (which we feel to be reasonable), then this advantage grows with the number of criteria.

5.4. Subjective Impressions

Table 9 shows the correlation coefficients for selected pairs of statements S_i from the feedback questionnaire. The statements are provided in the Appendix. All three results are significant at levels of $\alpha = 0.05$ and power $1-\beta = 0.9$.

Table 9. Correlations in feedback

Statements	S_2/S_5	S_2/S_4	S_4/S_5
$r_{1,2}$	0.42	0.76	0.70

Statements S_4 and S_5 refer to the rankings obtained from the scoring and lexicographic methods. S_4 asks which produces the better result and S_5 elicits whether for this innovation task an "all-round" idea or a "focus" idea is more appropriate. An all-round idea is strong with respect to several criteria (the Olympic decathlete), whereas a focus idea is excellent with respect to one criterion (the 100 meter sprint winner). The strong positive correlation ($r_{1,2} = 0.70$) shows that those subjects who consider the focus alternative to be more appropriate also prefer the lexicographic method over scoring.

Statements S_2 and S_4 elicit the appropriateness of points or ranks for individual judgments and of the scoring and lexicographic methods respectively. The correlation coefficient of 0.76 indicates that those subjects who considered ranks to be a better representation of their judgments also considered the lexicographic method to be superior to the scoring method.

Similarly, a high positive correlation was observed between appropriateness of ranks or scores (S_2) and all-round or focus task (S_5).

A further statement S_6 was used to test the effect of framing on the subjects' perceptions of the compensatory effect on the appropriateness of the scoring method. In version S_{6a}, the question was framed as strengths compensating weaknesses, whereas in version S_{6b}, the question was presented as weaknesses compensating strengths. A single-sided Mann Whitney U test was performed on the resulting data, which yielded a significant ($U=59.5$, $U_{crit} = 64$) difference between the two responses in favour of S_{6a}. When presented as a weakening, rather than a strengthening effect, the subjects were less inclined to accept the compensation effect of the scoring method.

Subjects found it easier to award ranks to the alternatives than points (S_1: mean=3.7, CI=0.07, α=0.05), supporting a result already established in the literature. Ranks were considered to be more appropriate then scores (S_2: mean=3.57, CI=0.06, α=0.05). Also, the lexicographic method was found to be easier to carry out overall than the scoring method (S_3: mean=3.73, CI=0.08, α=0.05).

The response counts for all questions are given in Table 10 in the Appendix.

6. Conclusion

6.1. Recommendations for Managers

Based on the results described in this paper as well as others already established in the literature we have six recommendations for managers who carry responsibility for the innovation process:

(1) The scoring model is highly problematic for selecting innovation projects. The problems of non-commensurability, false accuracy, non-independence of criteria and non-robustness with respect to uncertainty in individual scores are difficult, expensive to avoid and can easily lead to incorrect selections.

(2) If the scoring method is used, be aware of the dangers of its compensatory nature: *An excellent design for the walls and roof [combined with a] terrible floor plan do not average out to a satisfactory house.* (Müller-Herbers [16])

(3) Innovations are often pursued in order to achieve a specific goal. Making compromises in selecting innovation ideas therefore appears to be an

error. Managers should commit to a set of performance criteria for innovation projects that is as small as possible and relegate all other criteria to sufficiency type. The latter should then be dealt with first, in order to reduce the size of the idea pool.

(4) Do not award cardinal scores to sufficiency criteria. Instead, treat them as pass/fail decisions. *Have we validated the customer need? Is the product idea well-specified?* and other such early-stage innovation criteria [6] should have Yes/No answers.

(5) Be aware of the possibility of hidden profiles, since these can severely contaminate the group decision. These can be identified and eliminated [14][15], but this requires additional discussion time.

(6) In a group situation, use medians to aggregate individual evaluations rather than means.

(7) Determine ranks by using pairwise comparisons rather than by assigning points. These are both easier to perform and are constant across individuals. The statement *A1 is preferable to A2* has the useful property of having the same meaning when made by different people. On the other hand, every student knows that a top grade has different value depending on which professor awarded it.

6.2. Outlook

It is unknown whether a group can more easily and quickly establish a consensus on a score or a rank. We hypothesize that the rank is preferable. This would add a further argument in support of a lexicographic approach. This question will be the subject of a future study.

Based on the results obtained so far, we will start to experiment with lexicographical evaluation methods in commercial innovation projects in order to gather more empirical evidence about their performance and acceptance.

One current limitation of the lexicographic method is that it forbids equal priorities to be assigned to the criteria. Since this may well be necessary in practice, the corresponding algorithm needs to be developed. This would alleviate concerns about a single "dictator" criterion, since "dictatorship" could be extended to a set of criteria.

Another interesting avenue of investigation would be to develop a lexicographic method that allows satisficing [23] in the manner of the L* method by Encarnación [8], but oriented towards performance criteria. This would permit the use of prioritized satisficing criteria.

Finally, to facilitate practical application, we plan to develop a checklist to aid relegating an apparently performance criterion to a sufficiency criterion, since this appears to be a non-trivial (but critical) question.

7. References

[1] J. Banks, J. Carson, B. Nelson and D. Nicol, "Discrete Event System Simulation", Prentice Hall, 2009.

[2] D. Braziunas, C. Boutilier, "Preference Elicitation and Generalized Additive Utility", AAAI Conference on Artificial Intelligence, Boston, USA, 2006.

[3] R. O. Briggs, G. Kolfschoten, G.-J. de Vreede, C. Albrecht, D. R. Dean and S. Lukosch, "A Six-Layer Model of Collaboration, in [20].

[4] M. Colman, J. A. Stirk, "Singleton Bias and Lexicographic Preferences Among Equally Valued Alternatives", Journal of Economic Behaviour & Organization, Vol. 40, 337-351, 1999.

[5] R. Cooper, "The Stage-Gate Idea-to-Launch Process– Update, What's New and NexGen Systems", J. Product Innovation Management, Vol. 25, No. 3, 213-232, 2008.

[6] A. Day, "Is It Real? Can We Win? Is It Worth Doing? Managing Risk and Reward in an Innovation Portfolio", Harvard Business Review, 85/12, 110-120, 2007.

[7] J. Dyer, "Multiattribute Utility Theory", in [10].

[8] J. Encarnación, "Group Choice With Lexicographic Preferences", Philippine Review of Economics and Business, Vol. XVIII, Nos. 3 and 4, 1981.

[9] H. Fargier, P. Perny, "Qualitative Models for Decision Under Uncertainty without the Commensurability Assumption", Uncertainty in Artificial Intelligence, Stockholm, Sweden, 1999.

[10] J Figueira, S. Greco and M. Ehrgott, "Multiple Criteria Decision Analysis", Springer, 2005.

[11] P. C. Fishburn, "Lexicographic Orders, Utilities and Decision Rules: A Survey". Management Science 20(11), 1442-1471, 1974.

[12] N. Georgescu-Roegen, "Choice, Expectations and Measurability", Quarterly Journal of Economics, Vol. 68, 503-534, 1954.

[13] K. Girotra, C. Terwiesch and K. Ulrich, "Idea Generation and the Quality of the Best Idea", Management Science, 56:591-605, 2010.

[14] G. Horton, J. Görs, "Mining Hidden Profiles in the Collaborative Evaluation of Raw Ideas", Hawaii International Conference on System Sciences (HICSS 47), Hawaii, January 2014.

[15] G. Horton, J. Görs, "A Criterion-Mining Method for Group Idea Selection – Increasing Consensus with Minimal Loss of Efficiency", Hawaii International Conference on System Sciences (HICSS 48), Hawaii, January 2015.

[16] S. Müller-Herbers, "Methoden zur Beurteilung von Varianten" (Methods for evaluating alternatives), School of Architecture and City Planning, Stuttgart University, Germany, 4th edition, 2007 (in German).

[17] R. Kohli and K. Jedidi, "Representation and Inference of Lexicographic Preference Models and Their Variants", Marketing Science 26(3), 380–399, 2007.

[18] G. Kolfschoten, P. Lowry, D. Dean, G.-J. De Vreede and R. Briggs, "Patterns in Collaboration", in [20].

[19] G. Kolfschoten, S. Lukosch, and A. Mathijssen, "Supporting Collaborative Design: Lessons from a case study at the ESA concurrent design facility", Group Decision and Negotiation, Recife, Brazil, 2012.

[20] J. Nunamaker, R. Briggs and N. Romano (Eds.), Collaboration Systems: Concept, Value, and Use, Routledge, 2014.

[21] T. Saaty, "The Analytic Hierarchy and Analytic Network Processes for the Measurement of Intangible Criteria and for Decision-Making", in [10].

[22] P. Samuelson, W. Nordhaus, "Economics", McGraw-Hill, 16th edition, 1998.

[23] H. A. Simon, "Models of Man: Social and Rational", Wiley, 1957.

[24] G. L. S. Shackle, H. Wold and L. J. Savage, "Ordinal Preferences or Cardinal Utility?", Econometrica, Volume 20, Issue 4, 661-664, 1952.

[25] G. Stasser, W. Titus, "Hidden profiles: A brief history.", In Psychological Inquiry, Volume 14, Issue 3-4, pp. 304-313, 2003.

8. Appendix

8.1. Evaluation Task

The five innovation proposals to be evaluated in the lab experiment were:

- A_1: We introduce a double degree with a university in UK, AU, CA or US.
- A_2: We introduce courses offered by visiting faculty from business.
- A_3: We introduce modules on cloud computing and mobile app development.
- A_4: Credits from other universities can be counted towards our degree.
- A_5: Students can freely select courses up 20 credit points from other departments.

The five criteria used to evaluate the alternatives in the lab experiment were:

- C_1: Attractiveness of the department for high-school students
- C_2: Professional relevance of the program
- C_3: Variety of courses offered
- C_4: Broaden the horizon of the students
- C_5: Interdisciplinary cooperation

8.2. Survey Statements

The statements S_1 through S_5 used in the feedback response were as follows (Each response was given on a five-point Likert scale):

- S_1: Which type of evaluation was easier to award? (Points=1, Ranks=5)
- S_2: Which type of evaluation allowed you to express your opinion better? (Points=1, Ranks=5)
- S_3: Which method is easier to carry out? (Scoring=1, Lex=5)
- S_4: Which method yielded a better ranking result? (Scoring=1, Lex=5)
- S_5: What is more appropriate: an "all-round"-idea (good in several criteria) or a "focus" idea (excellent in one criterion)? (All-Round=1, Focus=5)
- S_{6a}: Is it appropriate that a weakness with regard to one criterion can be compensated by a strength with regard to a different criterion? (No=1, Yes=5)
- S_{6b}: Is it appropriate that a strength with regard to one criterion can be compensated by a weakness with regard to a different criterion? (No=1, Yes=5)

8.3. Results from Lab Experiment

Table 10 shows a summary of the subjects' responses referred to in Section 5.4.

Table 10. Subjects' Responses

	Question						
Response	S1	S2	S3	S4	S5	S6a	S6b
1	1	0	2	2	3	1	2
2	4	5	3	3	5	2	8
3	5	8	7	10	8	3	1
4	13	12	7	12	10	7	4
5	7	5	11	3	4	2	0