Expert Systems with Applications 39 (2012) 11048–11054

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

An empirical test of the evidential reasoning approach's synthesis axioms

Ian N. Durbach [*]

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

ARTICLE INFO

Keywords:
Multiple criteria decision analysis
Uncertainty
Evidential reasoning
Belief structure

ABSTRACT

This paper reports results from two empirical tests of the descriptive validity of synthesis axioms used by the evidential reasoning (ER) approach to aggregate performance over multiple criteria. These show that evaluations which invoke the axioms frequently violate them. The two most systematic aspects of the violations are that aggregate evaluations tend to be more favourable than basic evaluations, and that small amounts of ignorance on one attribute may be compensated for by complete assessments on other attributes. The implications for prescriptive use of the ER approach are discussed and some practical assessment procedures suggested.

1. Introduction

The evidential reasoning (ER) approach (e.g. Xu, 2011; Yang, 2001) is a general model for analysing multi-criteria decision problems under conditions of uncertainty. It is based on Dempster–Shafer theory of evidence (Shafer, 1976), which provides an explicit mechanism for dealing with ignorance in subjective probabilities by replacing the notion of probability with a 'degree of belief' that can be used to represent the extent to which a decision maker believes a specific proposition to be true (for example, that a new drug carries "negligible" side effects). The degrees of belief assigned to a set of collectively exhaustive and mutually exclusive hypotheses (by a so-called 'belief function') are allowed to sum to less than one, with the difference revealing the degree of ignorance. Such ignorance may be due to a lack of data or familiarity with the problem at hand, imprecision in assessment, or the absence of certain stakeholders in a group decision.

The ER approach has received substantial attention from researchers (see the review in Xu (2011)), possibly because its general nature means that a wide range of "uncertainties" can be treated within a unified mathematical framework, including probabilistic uncertainty, incomplete information and even complete ignorance, for both qualitative and quantitative evaluations. It is computationally simple to implement (special-purpose software is readily available (Xu & Yang, 2003)) and many applications have been reported in the literature (e.g. Wang, Yang, Xu, & Chin, 2009; Xu et al., 2007). At the heart of the approach is an algorithm for aggregating information across different criteria (hereafter referred to as the ER algorithm). To draw an analogy with decision methods based on utility functions: the belief functions act like marginal utility functions and contain information about intra-criterion preferences, while the ER algorithm acts like the additive, multiplicative, or multilinear aggregation of marginal performance over criteria (Keeney & Raiffa, 1976). The use of the ER algorithm is often motivated on the basis that it satisfies four "synthesis axioms" and therefore "provides a rational means for aggregating multiple attributes" (Yang & Xu, 2002). Indeed, in 2002 the original ER algorithm was modified into its current form at least in part because these synthesis axioms were not satisfied (Xu, 2011).

The objective of this paper is to empirically test whether people follow these four synthesis axioms in their own unfacilitated decision making – that is, whether the axioms are descriptively valid. On the basis of two behavioural experiments conducted as part of the paper, it appears that descriptive violations of all four axioms are fairly common. Of course, the ER approach is *prescriptive* in nature, and any descriptive violations must be interpreted with this in mind. Nevertheless I believe it is of some value to document these descriptive violations, discuss possible reasons behind them and explore possible responses. To again draw an analogy with expected utility theory, it is a well-established fact that the theory does not adequately describe peoples' choices, in the sense that a substantial proportion of people violate one or more of the axioms at least some of the time (e.g. the famous paradoxes of Allais (1953) and Ellsberg (1961)). Despite these descriptive failings, multi-attribute utility theory remains a popular prescriptive decision aid, as any contemporary textbook on the subject will attest to (e.g. Belton & Stewart, 2002). Moreover, the discovery of descriptive violations has led to the development of extensions to the original theory and a host of related non-expected utility theories (e.g. Starmer, 2000). That is, far from weakening expected utility theory (except as an all-encompassing theory of decision-making), discoveries of descriptive failings have with time substantially enriched our understanding of the area.

[*] Tel.: +27 21650508; fax: +27 216504773.
E-mail address: ian.durbach@uct.ac.za

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.03.018

The remainder of the paper is structured as follows. Section 2 describes the ER approach and the associated synthesis axioms. Section 3 describes the two empirical experiments conducted to test the descriptive validity of the axioms, and Section 4 gives the results. Section 5 discusses the implications of the reported violations for prescriptive use of the ER approach, and briefly identifies some possible responses within the ER framework. A final section concludes the paper.

2. The evidential reasoning (ER) approach

We consider a decision problem consisting of I alternatives denoted by a_i, $i \in \{1, \dots, I\}$, each evaluated on J criteria denoted by c_j, $j \in \{1, \dots, J\}$. Let Z_{ij} be the evaluation of a_i in terms of criterion c_j, according to some suitable performance measure. Our concern is with decision making situations in which the values of Z_{ij} for each i are not known with certainty for all j.

In the basic ER approach, a discrete set of evaluation 'grades' $H = \{H_1, H_2, \dots, H_N\}$ are defined for each attribute. The decision maker expresses their degree of belief $\beta_{nij} \in [0, 1]$ that alternative a_i should be assessed to grade H_n on criterion c_j. The distributed assessments for each attribute are combined using a recursive algorithm into a general assessment $\beta_{ni\cdot}$ representing the belief that alternative a_i achieves grade H_n.

2.1. The ER aggregation algorithm

The algorithm is based on first multiplying marginal beliefs with importance weights on each attribute, i.e. $m_{nij} = w_j\beta_{nij}$, with weights normalised to sum to one. The probability mass that remains unassigned to any individual grades (i.e. the difference between 1 and the sum of weighted beliefs) is then partitioned into what is due to non-zero weights on other attributes, i.e. $\bar{m}_{Hij} = 1 - w_j$, and what is due to incomplete assessment, i.e. $\tilde{m}_{Hij} = 1 - \sum_n w_j\beta_{nij}$. The quantity $\bar{m}_{Hij}$ can be interpreted as unassigned due to other criteria being important, i.e. because criterion c_j plays only one part in the assessment. The quantity $\tilde{m}_{Hij}$ is assigned due to incompleteness in the current assessment.

The ER algorithm then uses the following recursive equations to aggregate the basic probability masses described above. First, let G_{nij}, $\overline{G}_{nij}$, and $\widetilde{G}_{nij}$ denote the combined probability masses generated by aggregating the first j criteria. For $j = \{1, 2, \dots, J - 1\}$, the $(j + 1)$th criterion is then combined with the first j criteria as follows (the order of aggregation has no effect (Yang & Xu, 2002)):

$$G_{ni1} = m_{ni1} \tag{1}$$

$$G_{Hi1} = m_{Hi1} \tag{2}$$

$$\overline{G}_{Hi1} = \bar{m}_{Hi1} \tag{3}$$

$$\widetilde{G}_{Hi1} = \tilde{m}_{Hi1} \tag{4}$$

$$K_{i(j+1)} = \left(\sum_{p=1}^{N} \sum_{n=1, n \neq p}^{N} G_{pij} m_{ni(j+1)} \right)^{-1} \tag{5}$$

$$G_{ni(j+1)} = K_{i(j+1)}(G_{nij}m_{ni(j+1)} + G_{Hij}m_{ni(j+1)} + G_{nij}m_{Hi(j+1)}) \tag{6}$$

$$\widetilde{G}_{Hi(j+1)} = K_{i(j+1)}(\widetilde{G}_{Hij}\tilde{m}_{Hi(j+1)} + \overline{G}_{Hij}\tilde{m}_{Hi(j+1)} + \widetilde{G}_{Hij}\bar{m}_{Hi(j+1)}) \tag{7}$$

$$\overline{G}_{Hi(j+1)} = K_{i(j+1)}(\overline{G}_{Hij}\bar{m}_{Hi(j+1)}) \tag{8}$$

$$G_{Hi(j+1)} = \widetilde{G}_{Hi(j+1)} + \overline{G}_{Hi(j+1)} \tag{9}$$

The final quantity G_{niJ} is the probability mass assigned to grade n, $\widetilde{G}_{HiJ}$ is the unassigned probability mass due to incompleteness, and $\overline{G}_{HiJ}$ is the unassigned probability mass that must still be redistributed among the N grades. This is done proportionately so that

$$\beta_{ni\cdot} = \frac{G_{niJ}}{1 - \overline{G}_{HiJ}} \tag{10}$$

$$\beta_{Hi\cdot} = \frac{\widetilde{G}_{HiJ}}{1 - \overline{G}_{HiJ}} \tag{11}$$

While the ER algorithm is not particularly transparent (in the sense of easily explained to decision makers), central to the aggregation is the following (given for the two-attribute case but readily extended), from which some insight may be gained:

1. The product of m_{nij} with any of m_{nik}, $\bar{m}_{ik}$ or $\tilde{m}_{ik}$ is taken to be evidence in support of the aggregated assessment $m_{ni\cdot}$.
2. The product of $\tilde{m}_{ij}$ with any of $\bar{m}_{ik}$ or $\tilde{m}_{ik}$ is taken to be evidence in support of the aggregated assessment of incompleteness $m_{Hi\cdot}$.
3. The product $\bar{m}_{ij}\bar{m}_{ik}$ is reallocated back to all individual grades (including the incomplete grade) in proportion to their current assessments, i.e. $\beta_{ni\cdot} = m_{ni\cdot}/(\bar{m}_{ij}\bar{m}_{ik})$, $\beta_{Hi\cdot} = m_{Hi\cdot}/(\bar{m}_{ij}\bar{m}_{ik})$.

Extensions to the basic approach outlined above allow for grading systems to differ between attributes as well as for quantitative attributes (e.g. Xu, 2011). Both require an additional step in which equivalences are established between overall grades and grades used to describe marginal performances on attributes. These can be done using either rule-based or utility-based transformations (Yang, 2001). Further extensions allow unallocated beliefs to be redistributed to a restricted grade 'interval', i.e. subset of grades (Xu, Yang, & Wang, 2006), and for beliefs to be expressed as intervals (Wang, Yang, Xu, & Chin, 2006) or fuzzy linguistic terms (Yang, Wang, Xu, & Chin, 2006) rather than crisp values.

2.2. The synthesis axioms

As Xu (2011) states, "the rationality of the ER algorithm is checked by using the following four synthesis axioms". Others have proved that the combined degrees of belief generated by the ER algorithm satisfy the axioms (Huynh, Nakamori, Ho, & Murai, 2006; Yang & Xu, 2002), although it is neither the only algorithm to do so nor would these be the only axioms that the algorithm satisfies. The axioms are:

Independence: If no basic criterion is assessed to an evaluation grade then the general criterion should not be assessed to the same grade either:

$$\sum_{j=1}^{J} \beta_{nij} = 0 \Rightarrow \beta_{ni} = 0$$

Consensus: If all basic criteria are precisely assessed to an individual grade, then the general criterion should also be precisely assessed to the same grade:

$$\sum_{j=1}^{J} \beta_{nij} = J \Rightarrow \beta_{ni} = 1$$

Completeness: If all basic criteria are completely assessed to a subset of grades $\mathcal{N}$, then the general criterion should be completely assessed to the same subset as well:

$$\sum_{j=1}^{J} \sum_{n \in \mathcal{N}} \beta_{nij} = J \Rightarrow \sum_{n \in \mathcal{N}} \beta_{ni} = 1$$

Incompleteness: If basic assessments are incomplete, then the general assessment should also be incomplete:

$$\beta_{Hij} > 0 \Rightarrow \beta_{Hi} > 0$$

The essential logic behind the axioms is well summarised by Xu (2011): "if an alternative has a good (or bad) performance on a sub-criterion, then it must be good (or bad) to a certain extent overall". It is the descriptive validity of this claim which I aim to test using the experiments reported in the following sections.

3. Design of experiments

3.1. Experiment 1: property evaluations

3.1.1. Experimental task

In this experiment, subjects are first shown 16 separate images of rooms in various houses or apartments (specifically, the bathroom, bedroom, kitchen, and lounge area). The images are presented sequentially. For each image, subjects are asked to evaluate the rooms by allocating 100 points across the following four qualitative grades: unappealing/unsatisfactory, adequate/satisfactory, luxurious, or not sure/cannot tell. Subjects are specifically told that they have a total of 100 points to allocate across the four grades, and that, for each grade, they should give a number indicating the degree to which they think the grade matches their opinion of the image. They are also instructed that their responses should be integers from 0 to 100, and must sum to 100. Although brief, with this wording and accompanying description, I hope to capture their 'degree of belief' in the conventional sense of the term, in at least an approximate way.

After seeing all 16 individual images, subjects are shown a further 8 panels, each of which contains a collection of between two and four of the individual images. Images on the same panel are always of different room types, i.e. subjects will never be shown a panel with two kitchen images, for example. Subjects are asked to evaluate their *overall impression* (emphasis given in task) of the house or apartment in question, based on all the images that are shown to them in the panel. Apart from this addition, the task is the same as those involving individual images. Subjects are given the same set of instructions as described above, and the same set of four assessment grades is used.

Our interest is of course in whether the overall assessments of the composite images on the panels synthesise the assessments from the individual images in such a way that is consistent with the ER algorithm's synthesis axioms. In order to increase the likelihood of subjects using the "not sure/cannot tell" option and therefore provide more data with which to test the incompleteness axiom, some of the images were manually altered to appear extremely blurred.

3.1.2. Recruitment

Both experiments reported on were conducted over the World Wide Web using the Amazon Mechanical Turk platform (AMT, see http://www.mturk.com). AMT is a web-based labour market originally created to facilitate crowdsourcing (Howe, 2008) of tasks that are either easier or cheaper for humans to perform than for machines. Typical tasks involve image labelling (as in the current study), translation, or various types of classification. Compensation varies with task time and difficulty, but is typically between $0.01 and $0.10 per task. When choosing a task to work on, workers (that this, the people performing the tasks on AMT for financial compensation) select the tasks they wish to do. They may view descriptions or see previews of the task before choosing to accept a task.

AMT is used here as a convenient pool of subjects willing to participate in laboratory-style behavioural experiments for a small payment. Such web-based experimental platforms are becoming increasingly popular with behavioural researchers, with cited advantages (Goel, Reeves, Watts, & Pennock, 2010) the speed and inexpensiveness of conducting experiments, as well as obtaining a potentially more diverse sample than would be typical with university-based lab experiments. While response quality in relatively anonymous web-based environments may be a concern, a safeguard when using AMT is that the person requesting the task can assess the completed tasks and deny payment if performance is deemed unsatisfactory. This is rarely done in university-based lab experiments. In our experiments, for example, payment was withheld (and the response excluded from the results) if beliefs summed to less than 100. Others have shown that good performance can be obtained using crowdsourcing when it takes the same amount of time to perform the task faithfully as to cheat – as is the case in our experiments, to rough approximation (Kittur, Chi, & Suh, 2008).

For this experiment 100 workers were recruited, each earning $0.02 per successfully completed task, i.e. $0.48 would be earned for completing all 24 tasks. From these, a total of 1688 valid responses from 74 unique workers were obtained (425 aggregate evaluations). The remaining responses were discarded for various reasons, most commonly because the worker did not complete the final few tasks involving composite images, or incorrectly answered one or more of the tasks involving individual images. In both these cases it would have been impossible to evaluate the synthesis axioms. The average time taken to complete one of the 16 single-image tasks was 31 s, while tasks with more than one image were completed in an average of 38 s.

3.2. Experiment 2: evaluations of internet service

3.2.1. Experimental task

In the second experiment, subjects answered a short survey about their internet connection. The subject of the survey was chosen on the basis that it was certain that all AMT workers possessed an internet connection, and that it might be considered fairly important to them since at least some of their income is derived from it. Subjects were asked the following five questions (emphasis provided in task):

1. How would you rate your happiness with the *speed* of your internet connection?
2. How would you rate your happiness with the *reliability* of your internet connection (how often you encounter connection problems)?
3. How would you rate the *pricing* of your internet connection?
4. Are there any other features of your internet connection that are important to you? If so, please list these and indicate how satisfied you are with these features?
5. How satisfied are you *overall* with your internet connection?

Subjects were asked to respond to each question by allocating 100 points across the following six qualitative grades: extremely unsatisfied, unsatisfied, about average, satisfied, extremely satisfied, or don't know. Subjects were given the same instructions as for the previous experiment. That is, that they had a total of 100 points to allocate across the grades, and that their responses should indicate how well each grade described their opinion on that particular attribute. In addition, they were also instructed that they would only be paid if they completed all of the questions listed above.

In this experiment an aggregate question about any 'other' important features was added to avoid the possibility of attributing any discrepancy between the overall and synthesised evaluations to the omission of attributes. Because the ER algorithm can be applied at any level of the objectives hierarchy, it does not affect our ability to evaluate any of the synthesis axioms. The order of the question about overall satisfaction – whether it is asked before or after the four attribute-specific questions – is also varied, to test whether this plays any role in the way in which evidence is synthesised. Violations of the completeness axiom are marginally more likely to occur when overall evaluations are asked for first but these differences are not significant at the 5% level ($\chi^2 = 3.6$, $p = 0.06$). No other associations are significant at even the 20% level. Results are therefore pooled across the two order conditions.

3.2.2. Recruitment

For this experiment 200 subjects were recruited using AMT, each earning $0.20 for successfully completing the survey. Subjects were instructed to not attempt the task if they had participated in the earlier experiment. Since each worker possesses a unique identification number, this requirement was enforced (one worker who had participated in the earlier experiment was removed from the analysis). Half of the 200 workers saw the overall evaluation question before the attribute-specific ones, and half after. From these, a total of 161 valid responses from obtained, 79 giving their overall evaluations first and 82 giving them last. The remaining responses were discarded for similar reasons to Experiment 1. The average time taken to complete the task was just under 4 min.

4. Results

The results of the two experiments are given in Table 1. Because subjects make more than one aggregate evaluation in Experiment 1, both the proportion of judgements violating the synthesis axioms as well as the number of subjects who make one or more violations are shown. For Experiment 2 these numbers are identical, since each subject makes only one aggregate evaluation.

It is clear from Table 1 that many of the aggregate evaluations fail to invoke the conditions of the synthesis axioms, particularly the consensus and completeness axioms, and therefore cannot be used to assess the validity of the axioms. Of the evaluations which do invoke each of the axioms though, a significant proportion are violations. While it seems reasonable to suggest that the exact proportion of violations will depend on the context of the problem, in the two experiments conducted here the proportion of violations ranges between around 15% for the independence axiom to around 50% for the incompleteness axiom in Experiment 2. Furthermore, the analysis of individual subjects, i.e. the bottom half of Table 1, indicates that these violations cannot be attributed to a small proportion of subjects each making many violations. Rather, a substantial proportion of subjects – between 25% and 50% depending on the axiom – exhibit at least one violation. It therefore seems reasonable to suggest that in many instances decision makers will make aggregate evaluations in a way that is different from that stated by the synthesis axioms. Table 2 shows the correlations between binary indicators of violations of each of the axioms. The large positive correlation between violations of consensus and completeness is to be expected since the former is a special case of the latter. There are also some moderate positive correlations involving one of the consensus and completeness axioms, but these are based on only a small number of actual violations. Otherwise, there is little evidence to suggest any particularly strong associations between violations of the different axioms.

Table 2
Correlations between violations of the synthesis axioms. Above-diagonal and below-diagonal entries are for Experiments 1 and 2, respectively.

	Indep.	Cons.	Comp.	Incomp.
Independence	–	0.30	0.39	0.04
Consensus	0.38	–	0.77	−0.05
Completeness	0.28	−0.04	–	−0.06
Incompleteness	−0.02	−0.03	−0.13	–

Having established that each of the synthesis axioms are violated at least some of the time, a natural question arises as to the size and nature of these violations. Unfortunately the consensus and completeness axioms are invoked so infrequently that our sample size is too small to analyse further, but Fig. 1 show box-and-whisker plots of the size of any violations, for the independence and incompleteness axioms. For the independence axiom, the plot shows the overall belief assigned to a grade which was not assigned any belief on any basic criteria (implying, by the independence axiom, that the overall grade should also be assigned a zero belief). For the incompleteness axiom, the plot shows the proportion of total belief assigned to ignorance over all basic criteria, in cases where overall assessments were complete, i.e. contained no ignorance. Note that in both cases the plots are based only on assessments which violate the axioms.

It appears that violations of incompleteness, where they occur, are often fairly small in magnitude. This suggests that when decision makers form aggregate judgments, small amounts of ignorance on one attribute may be compensated for by complete assessments on other attributes. This is supported by Fig. 2, which shows the number of evaluations violating and supporting the incompleteness axiom as a function of how many basic attributes have received incomplete assessments. In both experiments the majority of violations take place when only a single basic attribute (out of four) is incomplete. If more than one basic attribute is incomplete, decision makers are much more likely to leave their overall evaluations incomplete as well.

Violations of independence, on the other hand, can often be large in magnitude. This suggests that, even if a decision maker assigns no weight to a grade on any basic attribute, they may still assign a substantial amount of belief to that grade when making aggregate judgments. An initial suspicion was that decision makers who assess some basic attributes as very poor and some as excellent may give an intermediate overall assessment, but this turned out to be relatively rare (5 out of 25 violations in Experiment 1; 2 out of 13 violations in Experiment 2). Instead there is a tendency for those violating the independence axiom to give more favourable overall assessments than basic assessments, in the sense of allocating relatively more belief to the more favourable grades. This is also true of those who do not violate the axiom, but to much lesser extent. The

Table 1
Results obtained from the two experiments.

	Independence		Consensus		Completeness		Incompleteness	
	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2
Total number of judgements	425	161	425	161	425	161	425	161
Judgements invoking axioms	128	103	8	10	27	35	329	37
Judgements violations axioms	25	13	3	2	5	7	97	19
% violations (all judgements)	6%	8%	1%	1%	1%	4%	23%	12%
% violations (invoking axioms)	20%	13%	38%	20%	19%	20%	29%	51%
Total number of subjects	81		81		81		81	
Subjects invoking axioms	41		6		17		74	
Subjects with ⩾1 violation	17		2		4		37	
% violations (all subjects)	21%		2%		5%		46%	
% violations (invoking axioms)	41%		33%		24%		50%	

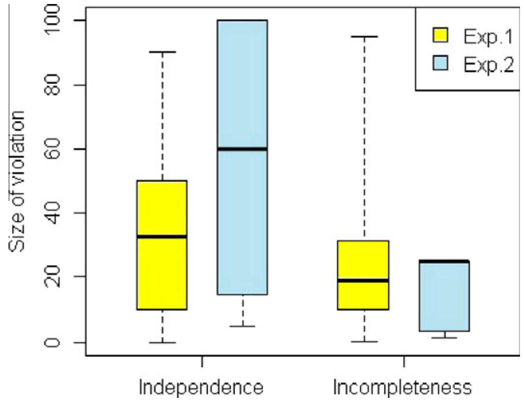

Fig. 1. Distribution of the size of violations for the independence and incompleteness axioms.

effect is shown in Table 3. It is difficult to offer any conclusive explanation for the observed pattern here, but one suggestion is that some decision makers may view consistent performance as a goal in itself – leading them to view consistently mediocre performance, for example, as better than mediocre in the aggregate. It is also conceivable that in other problem contexts decision makers might prefer very good performance on at least one of the basic attributes, i.e. consistency may be viewed as negative.

5. Implications for prescriptive decision aid

It is clear from the results above that the synthesis axioms do not describe the way that all decision makers aggregate evidence. From a descriptive perspective, decision makers sometimes give complete aggregate evaluations even when they have given incomplete basic evaluations (violating incompleteness); they assign belief in the aggregate to a grade that was not allocated any attribute-specific belief (violating independence); and they assign no belief in the aggregate to grades that were allocated some attribute-specific belief (violating consensus or completeness). These violations occur frequently enough that the associated axioms cannot be considered valid in the descriptive sense. That is, they are not axiomatic in the sense of being either self-evidently or universally true. The questions that remains is what effect, if any, this has on the use of the ER approach as a prescriptive model.

In addressing this question it is useful to consider responses to violations of the axioms of expected utility. These have elicited three broad categories of response. The first is that the axiomatic violations are 'not that bad', even in a descriptive sense: that the majority of decision makers are at least in tentative agreement with the most controversial of the axioms, independence (French, 1995) and that at least some apparent violations can be classified as decision making with error rather than axiomatic violation (Carbone & Hey, 1994). The second response is pragmatic: that while the axioms are descriptively invalid and systematically violated, the theory retains a prescriptive value because it allows a simple and coherent framework for constructing preferences (Belton & Stewart, 2002). The final response views the violations as prescriptively as well as descriptively undesirable, and seeks to extend the decision model so that it is able to accommodate the violating behaviour.

Table 3
Average belief allocated to favourable grades in basic and overall evaluations. Overall evaluations in both experiments tend to be more favourable than basic evaluations, but the difference is much more pronounced among evaluations violating the independence axiom.

	Experiment 1		Experiment 2	
	Basic	Overall	Basic	Overall
Violating	8.6 (2.5)	21.3 (3.8)	39.5 (9.3)	56.2 (10.4)
Other	18.1 (0.8)	20.3 (1)	58.7 (2.3)	63.9 (2.8)

Of these, the first response does not seem convincing in the case of the synthesis axioms. Firstly, the proportion of judgements invoking the axioms which then violate them is in several cases quite large – of the order of 30–40%. Then, it is not only one of the axioms which are periodically violated. Substantive proportions of violations for all four synthesis axioms are observed. Even under the assumption of the limited dependence between violations of different axioms shown in Table 2, this means that the probability that a single evaluation will satisfy all four axioms is even smaller than those obtained for the individual axioms. The one behaviour which may have been considered as "decision making with error" – namely that good and bad basic evaluations cancel out and result in a mediocre aggregate evaluation – was not observed with any great frequency, and even here it not clear that an error is involved. While one cannot discount the possibility that subjects may have completed the tasks carelessly, the simplicity of the tasks (relative to a typical real-world decision problem) also means that *more* violations might occur in real-world decision-making.

The second, pragmatic response is perhaps the one that most practitioners of the ER approach would be comfortable with. While not clear, it is certainly arguable that violations of incompleteness are due to ignorance below a certain threshold being considered negligible when making aggregate evaluations, i.e. due to limitations in information processing capability. Thus the axiom may

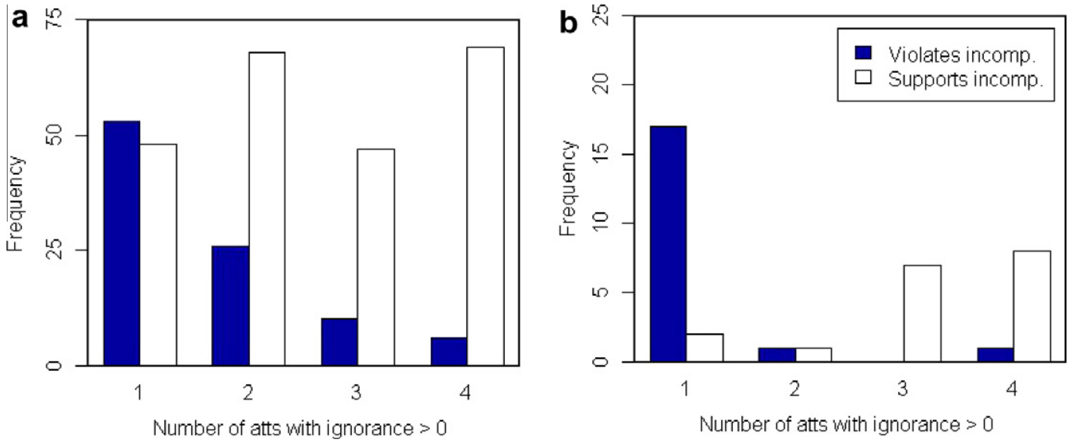

Fig. 2. Illustration of how violations of the incompleteness axiom decline if ignorance is recorded for more than one basic attribute, for (a) Experiment 1 and (b) Experiment 2.

retain prescriptive value. It seems harder to argue in favour of the independence axiom, where there seem no compelling reason why a grade cannot be assessed in the aggregate unless it has been assessed on at least one basic attribute. As mentioned, decision makers may place some kind of a premium on consistent performance, or perform some kind of interpolation between extreme performance on basic attributes. Neither of these violations of independence seem especially objectionable from a prescriptive perspective. From a practical point-of-view, the consistency between the grading of different attributes (e.g. a "good" internet speed; a "good" price) that is demanded by the independence axiom seems unlikely to be matched in practise.

One possibility is to develop practical procedures for assessing the extent to which violations occur, and to employ these as a part of the problem structuring process. Practitioners using the ER approach might ask some simplified questions (along the lines of the tasks presented in the two experiments), and assess the extent of any violations of the synthesis axioms. Severe violations which persist even after discussion between the facilitator and DM might suggest the use of a different MCDA technique, since the ER algorithm is unlikely to faithfully represent the preferences of the DM. Relatively minor violations, on the other hand, may be used as feedback in aligning the grading systems used to evaluate basic attributes, or might even be ignored in some situations.

Addressing the violations by adapting the ER algorithm seems, at the present time, a difficult proposition. Firstly, while some patterns in the violations have been demonstrated, none are what might be considered "systematic". Extending the algorithm to incorporate more detailed behaviour requires a good understanding of why certain violations occur. For example, while the independence axiom may be relaxed in order to allow aggregate evaluations to be made to a grade that has not been used in the basic assessments, it is not clear under what circumstances this allocation should be made or how it should be computed. Also the ER algorithm is already quite complex and non-transparent when compared to other decision aids. Further extensions may have substantial negative impacts on ease-of-use, confidence in the approach, facilitator–DM interaction, and other 'soft' issues.

6. Conclusion

This paper has reported results from two empirical tests of the descriptive validity of the synthesis axioms used by the ER approach to combine evidence across multiple criteria into a single assessment. The four synthesis axioms (independence, consensus, completeness, incompleteness) indicate primarily that no (all) belief should be assigned to an overall grade if no (all) basic attributes have been assessed to that grade(s), and that any incompleteness is preserved in the aggregate assessments.

The results show that while the majority of evaluations do not invoke the conditions of the axioms, those that do frequently violate them. The proportion of violations in the two experiments ranges from 15–20% for the independence axiom to 30–50% for the incompleteness axiom. None of the axioms therefore seem to hold descriptively to the extent required by an axiomatic system. The two most systematic aspects of the violations are:

1. Aggregate evaluations tend to be more favourable than suggested by the basic evaluations, so that aggregate assessments are often made to higher grades that were not employed during the basic assessments, violating the independence axiom.
2. When decision makers form aggregate judgments, small amounts of ignorance on one attribute may be compensated for by complete assessments on other attributes, violating the incompleteness axiom.

These descriptive violations by no means invalidate the use of the ER approach as a prescriptive decision aid – all decision models currently in use are either based on axioms that are periodically violated or do not have any explicit axioms at all. However they do suggest that practitioners exercise some caution when using the approach. Some decision makers will aggregate performance in a way that is quite different from that stated by the synthesis axioms, and it is not clear that they make any kind of definite error in doing so. A pragmatic middle-ground that has been suggested here is to incorporate some assessment of the synthesis axioms into the problem structuring phases. Violations may suggest that grading systems on the basic attributes have not been properly structured, or that the decision maker places some premium on consistently good performance. In more extreme cases the violations might be due to behaviour that is incompatible with the ER algorithm, necessitating another approach.

The experiments reported in this paper are a first attempt at evaluating the synthesis axioms and identifying any systematic violations, with a view to making the ER approach more robust to different types of behaviour. Because some of the axiomatic conditions were invoked relatively seldom in the two experiments – a clear limitation of the design used – it has not been possible to uncover much detail about the specific nature of all violations. There therefore remains a great deal of scope for future research, both to design more detailed experiments and to extend the ER algorithm in ways suggested by these experiments.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Americaine. *Econometrica, 21*, 503–546.

Belton, V., & Stewart, T. J. (2002). *Multiple criteria decision analysis: An integrated approach*. Boston: Kluwer Academic Publishers.

Carbone, E., & Hey, J. (1994). Estimation of expected utility and non-expected utility preference functionals using complete ranking data. In B. Munier & M. Machina (Eds.), *Models and experiments in risk and rationality* (pp. 119–140). Dordrecht: Kluwer Academic Publishers.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics, 75*, 643–669.

French, S. (1995). Uncertainty and imprecision: Modelling and analysis. *Journal of the Operational Research Society, 46*, 70–79.

Goel, S., Reeves, D. M., Watts, D. J., & Pennock, D. M. (2010). Prediction without markets. In *Proceedings of the 11th ACM conference on electronic commerce* (pp. 357–366). ACM.

Howe, J. (2008). Crowdsourcing: Why the power of the crowd is driving the future of business. *New York Crown Business*.

Huynh, V. N., Nakamori, Y., Ho, T. B., & Murai, T. (2006). Multiple-attribute decision making under uncertainty: The evidential reasoning approach revisited. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 36*(4), 804–822.

Keeney, R., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. New York: John Wiley & Sons.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceeding of the 26th annual SIGCHI conference on human factors in computing systems* (pp. 453–456). ACM.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature, 38*, 332–382.

Wang, Y. M., Yang, J. B., Xu, D. L., & Chin, K. S. (2006). The evidential reasoning approach for multiple attribute decision analysis using interval belief degrees. *European Journal of Operational Research, 175*(1), 35–66.

Wang, Y. M., Yang, J. B., Xu, D. L., & Chin, K. S. (2009). Consumer preference prediction by using a hybrid evidential reasoning and belief rule-based methodology. *Expert Systems with Applications, 36*(4), 8421–8430.

Xu, D. L. (2011). An introduction and survey of the evidential reasoning approach for multiple criteria decision analysis. *Annals of Operations Research*. http://dx.doi.org/10.1007/s10479-011-0945-9.

Xu, D. L., Liu, J., Yang, J. B., Liu, G. P., Wang, J., Jenkinson, I., et al. (2007). Inference and learning methodology of belief-rule-based expert system for pipeline leak detection. *Expert Systems with Applications, 32*(1), 103–113.

Xu, D. L., & Yang, J. B. (2003). Intelligent decision system for self-assessment. *Journal of Multi-Criteria Decision Analysis, 12*(1), 43–60.

Xu, D. L., Yang, J. B., & Wang, Y. M. (2006). The evidential reasoning approach for multi-attribute decision analysis under interval uncertainty. *European Journal of Operational Research, 174*(3), 1914–1943.

Yang, J. B. (2001). Rule and utility based evidential reasoning approach for multiattribute decision analysis under uncertainties. *European Journal of Operational Research, 131*(1), 31–61.

Yang, J. B., Wang, Y. M., Xu, D. L., & Chin, K. S. (2006). The evidential reasoning approach for MADA under both probabilistic and fuzzy uncertainties. *European Journal of Operational Research, 171*(1), 309–343.

Yang, J. B., & Xu, D. L. (2002). On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty. *IEEE Transactions on Systems, Man and Cybernetics, Part A, 32*(3), 289–304.