

An Evaluation of Multi-Criteria Methods in Integrated Assessment of Climate Policy

MICHELLE L. BELL*, BENJAMIN F. HOBBS, EMILY M. ELLIOTT, HUGH ELLIS, and ZACHARY ROBINSON

Department of Geography and Environmental Engineering, The Johns Hopkins University, USA

ABSTRACT

Those who conduct integrated assessments (IAs) are aware of the need to explicitly consider multiple criteria and uncertainties when evaluating policies for preventing global warming. MCDM methods are potentially useful for understanding tradeoffs and evaluating risks associated with climate policy alternatives. A difficulty facing potential MCDM users is the wide range of different techniques that have been proposed, each with distinct advantages. Methods differ in terms of validity, ease of use, and appropriateness to the problem. Alternative methods also can yield strikingly different rankings of alternatives. A workshop was held in which climate change experts and policy-makers evaluated the usefulness of MCDM for IA. Participants applied several methods in the context of a hypothetical greenhouse gas policy decision. Methods compared include value and utility functions, goal programming, ELECTRE, fuzzy sets, stochastic dominance, min max regret, and several weight selection methods. Ranges, rather than point estimates, were provided for some questions to incorporate imprecision regarding weights. Additionally, several visualization methods for both deterministic and uncertain cases were used and evaluated. Analysis of method results and participant feedback through questionnaires and discussion provide the basis for conclusions regarding the use of MCDM methods for climate change policy and IA analyses. Hypotheses are examined concerning predictive and convergent validity of methods, existence of splitting bias among experts, perceived ability of methods to aid decision-making, and whether expressing imprecision can change ranking results. Because participants gained from viewing a problem from several perspectives and results from different methods often significantly differed, it appears worthwhile to apply several MCDM methods to increase user confidence and insight. The participants themselves recommended such multimethod approaches for policymaking. Yet they preferred the freedom of unaided decision-making most of all, challenging the MCDM community to create transparent methods that permit maximum user control. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: integrated assessment; method evaluation; multi-criteria decision-making; climate change; experiment; weight selection; policy analysis

1. INTRODUCTION

“Greenhouse gases,” such as carbon dioxide, methane, nitrous oxide, and water vapor, raise the earth’s temperature by altering the planet’s radiation balance. Anthropogenic emissions of greenhouse gases have increased significantly since the industrial revolution and may enhance the greenhouse effect. If such global warming occurs, grave ecological, social, and economic conse-

quences could result. Impacts could include sea-level rise, changed crop yields, stressed ecosystems, and water shortages. Human health may be affected due to heat-related mortality and changes in air pollution and infectious disease patterns.

Predicting such impacts involves consideration of interactions between terrestrial, atmospheric, and hydrologic systems, as well as social, political, and economic systems. However, climate change experts are faced with numerous uncertainties and disagree on the magnitude, distribution, and timeframe of global warming impacts (IPCC, 1995; Morgan and Keith, 1995). The possibility of global warming has generated much research and political interest. In 1988, the Intergovernmental Panel on Climate Change (IPCC), comprised of 2500 leading climate scientists, was established to assess scientific, technical, and

*Correspondence to: DOGEE, The Johns Hopkins University, 313 Ames Hall, 3400 North Charles St., Baltimore, MD 21218, USA. E-mail: chelbell@jhu.edu

Contract/grant sponsor: National Science foundation; contract/grant number: NSF SBR9634336.

socioeconomic information relating to climate change. At a December 1997 conference in Kyoto, Japan, parties to the United Nations Framework Convention on Climate Change agreed to limit greenhouse gas emissions to below 1990 levels; but the US subsequently announced its withdrawal from the agreement, citing economic impacts and scientific uncertainty. Given the importance and complexity of the issues, climate change policy evaluation may benefit from the application of multi-criteria decision-making (MCDM) methods.

Integrated assessment (IA) aids the understanding of climate change consequences by using comprehensive models with interrelationships and feedbacks among system components (Dowlatabadi and Morgan, 1993; Parson and Fisher-Vanden, 1995; Shlyakhter *et al.*, 1995). The basic function of most IAs is system modelling, the simulation of the response of physical, biological, and/or social systems to changes in inputs, assumptions, and policies. The ultimate purpose of such efforts is to provide policy-makers with understanding of how assumptions and policies affect system behaviour and associated impacts (Gardiner and Ford, 1980). Some IAs also incorporate decision evaluation, the comparison of options in terms of their risks and performance on important criteria and the application of value judgments to rank or screen alternatives (Rotmans and Dowlatabadi, 1998).

IAs are most useful to policy makers if they are explicitly linked to decision-making (Bernabo and Eglinton, 1992; NAPAP, 1991). This linkage can be accomplished by building formal decision analytic capabilities into models, such as multi-objective tradeoff displays or decision trees. Alternatively, "policy-oriented assessments" can provide information on the performance and risks of policy options for use in decision processes taking place outside the IA system (Meo, 1991). A central aim of the workshop and the focus of this article is to help bridge the gap between IA system modeling and policy evaluation.

Climate change policy-makers face unique challenges, such as the lack of a single decision-maker, uncertainties, long time horizons, and the irreversibility of effects. Structured numerical analysis can aid understanding by managing and analyzing information and alternatives (Arrow *et al.*, 1996a). The use of MCDM methods has the potential to improve the quality of decision by providing information on tradeoffs, increasing

confidence in the decision, and documenting the process. MCDM can thereby function as one of the mediums through which decision-makers use and process IA information. Stewart (2000) identifies three distinct roles for decision analysis in such public sector problems: (1) initial impact assessment and screening; (2) "within interest" structuring and evaluation; and (3) "between interest" negotiation and decision-making. MCDM can play each of these roles in IA. This paper explores the use of MCDM methods in IA to aid policy-oriented impact assessment.

Decision analysis has been used previously to compare climate policies under uncertainty (Peck and Teisberg, 1996), address tradeoffs involved in assigning relative responsibility for greenhouse gas reduction (Ridgley, 1996), and evaluate sequential decision strategies for abating climate change (Hammitt *et al.*, 1992; Valverde *et al.*, 1999). For instance, Manne and Richels (1992) compared the economic impacts under uncertainty of three forms of greenhouse "insurance": (1) intensive research to reduce climate and impact uncertainties; (2) development of new energy supply and conservation technologies to reduce greenhouse gas abatement costs; and (3) immediate reductions in emissions to slow climate warming. In this paper, we systematically compare the results of a range of MCDM methods and their potential usefulness based upon a workshop in which climate experts and IA practitioners applied several MCDM methods in the context of IA.

The MCDM methods are summarized in Appendix A. The experiments conducted in the workshop are unusual in that MCDM methods were applied and assessed by climate change *experts*, rather than less experienced subjects, such as students. The specific purposes of the workshop were to: (1) compare method results and predictive validity; (2) evaluate each method's appropriateness and ease of use for climate change policy-making; (3) evaluate multiple visualization techniques for displaying tradeoffs for both deterministic and uncertain scenarios; and (4) expose workshop participants to MCDM methods and their application. This paper presents results from the workshop and discusses their potential implications for the decision analysis community.

Both user evaluations and analysis of method results are necessary to evaluate decision support systems (Evans and Riha, 1989; Gunderson *et al.*, 1995; Hobbs *et al.*, 1992). Therefore, method performance, appropriateness, and ease of use

Table I. Attribute values for climate policy alternatives

Attribute	x_1 : Global Temperature Increase [°C]	x_2 : Annualized SO ₂ Emissions [10 ⁶ tons/yr]	x_3 : Annualized Nuclear Waste [10 ³ tons/yr]	x_4 : Annualized Cost [10 ⁹ \$/yr]	x_5 : Sea-Level Rise [cm]	x_6 : Ecosystem Stress [10 ⁶ hectares]
Base Case	1.35	159.5	11.7	0.0	26.2	3229
\$75/ton CO ₂ Tax	1.33	136.8	15.4	37.0	25.9	3190
\$150/ton CO ₂ Tax	1.29	118.8	19.3	142.7	24.2	3095
\$300/ton CO ₂ Tax	1.15	93.5	26.0	519.8	22.4	2740
Nuclear Promotion	1.24	149.9	22.2	62.1	24.3	2977
Relaxed SO ₂ Standards	1.25	189.9	10.9	-3.6	24.4	3002
Biomass Promotion	1.30	153.4	11.6	7.1	25.4	3121

were assessed through participant feedback (questionnaires and structured discussions), while method results and validity were compared by statistical analysis of weights and policy rankings. Several limitations of the experiments, such as small sample size, prevent definitive conclusions regarding the relative merits of the methods. Nonetheless, such case studies or quasi-experiments can provide useful information (Adelman, 1991). For instance, such studies often possess an ecological validity (realism of problem setting and sophistication of participants) lacked by better controlled experiments, such as those involving large numbers of undergraduates (e.g., Corner and Buchanan, 1997; León, 1997; Stillwell *et al.*, 1987). Results of quasi-experiments with real practitioners and those from controlled experiments with less experienced subjects can yield more definitive conclusions than each type of study alone (Elmes *et al.*, 1995). Many similar field studies have been conducted with useful results (see reviews in Hobbs, 1986; Huber, 1974; John and Edwards, 1978; Leung, 1978; von Winterfeldt and Edwards, 1986).

The following section describes the experimental design and process, including the climate change policies considered, the MCDM approaches explored, and experimental design limitations. Then we describe the testing of several hypotheses and our conclusions.

2. EXPERIMENTAL DESIGN AND PROCESS

2.1. Climate policies considered

The workshop explored the following policies for limiting greenhouse gas emissions and impacts: base case (no new emissions limits); globally

applied tax of \$75, \$150, or \$300 per ton of carbon dioxide (CO₂) emitted; relaxed sulfur dioxide (SO₂) emission standards (which can have a cooling effect); promotion of nuclear power through subsidies for nuclear fuel; and promotion of biomass energy. The policies were compared relative to six attributes, which were chosen prior to the workshop to represent key features of the problem. Although many more attributes would be used in practice, the number was limited to six so that the problem would be manageable given the workshop's time limitations. The attributes considered are temperature increase (from 1990 to 2050), ecosystem stress (in 2050), sea-level rise (from 1990 to 2050), annualized SO₂ emissions (from 1990 to 2050), annualized nuclear waste generation (from 1990 to 2050), and annualized control cost compared to the base case (in 2050). Some exercises used four attributes for simplicity (temperature increase, SO₂ emissions, nuclear waste, and cost). Attribute values were global aggregate estimates obtained from an IA model (Holmes and Ellis, 1996, 1997).

The attribute scores for each alternative along with ranges (minimum, maximum) were given to participants in each MCDM exercise. Uncertainty scenarios were generated with Monte Carlo simulation using probabilistic inputs for climate sensitivity, SO₂ cooling effect, energy efficiency, labor productivity, natural gas reserves, and population growth. The hypothetical climate change scenarios were constructed to provide plausible attribute values for each policy alternative for the purpose of evaluating MCDM methods, not to provide definitive values. Mean attribute values for each of the non-dominated alternatives from results using the uncertain inputs are provided in Table I.

While climate change is a dynamic problem, attribute values were calculated for a specific time (2050) or time interval (change between 1990 and 2050). This approach provided a manageable set of information that we hoped would be familiar and easily understood, as it is comparable to other analyses of climate change impacts (e.g., IPCC estimates of temperature increase). Values for some attributes were discounted so that impacts in the distant future received less emphasis than those in the near future. Issues involved in this crucial weighting judgment are reviewed in Arrow *et al.* (1996b) and Schubert (1994).

We explained the attribute definitions and discounting procedure to the workshop participants. Although no participants questioned the use of a 1990 baseline or the discounting process, our decision to use these metrics affected the attribute values and therefore the value judgments. We do not know how participants' value judgments would differ had we, say, modeled impacts to 2100, used a different discounting procedure, or provided a full time series for each attribute rather than just means (e.g., temperature in each year).

2.2. MCDM methods

The MCDM methods compared in this workshop fall into three groups: weighting methods, deterministic ranking methods, and uncertainty ranking methods. The weighting methods address user preferences among the attributes (i.e., which attributes are more important to the user and by how much?). Deterministic and uncertainty ranking methods combine those preferences in order to rank or screen alternatives. Each method is defined briefly in Appendix A.

The methods selected are useful for evaluating discrete alternatives, can be conveyed to those unfamiliar with decision analysis, and in most cases have been widely applied elsewhere. They represent a range of divergent philosophies regarding decision-making. (For an in-depth discussion of conceptual differences among MCDM methods, see Stewart, 1992). For example, in direct assessments of importance weights, weighting values are transparent and under control of the decision-maker, whereas with the traditional AHP weighting method, weights are inferred from pairwise ratio comparisons of attributes (e.g., attribute i is twice as important as attribute j). However, it is widely argued (e.g., Belton, 1986; Schoemaker, 1981) that the notion of attribute "importance" is vaguely defined in both direct and AHP assess-

ments, and may diverge greatly from the rates at which users are willing to tradeoff one attribute for another. The latter is the precise definition of importance required by value and utility theory methods, and is what swing and tradeoff weighting methods attempt to capture. In this study, a hybrid swing/AHP method is tested that attempts to combine AHP's ease of use with swing weighting's more precise notion of attribute importance.

Several approaches to combining attributes and ranking alternatives also were compared by the workshop participants. Value or utility methods use each alternative's performance on each attribute along with attribute weights to create a performance score for each option. Goal programming methods require the user to set a target value for each attribute and minimize deviations, in one or both directions, from that target. Outranking methods, such as ELECTRE, do not necessarily generate a complete ranking of alternatives as they allow two alternatives to be 'incomparable' (i.e., neither alternative outranks the other). These methods define a kernel or set of non-dominated alternatives.

2.3. The experiment

The workshop was held June 1–2, 1998, at The Johns Hopkins University. The participants included twenty climate change experts, policy-makers, and IA practitioners, from a range of private and public institutions including academic, governmental, national laboratory, and corporate organizations (Appendix B). Of the 16 participants who filled out the questionnaire regarding previous use of MCDM, 44% either had used MCDM before or were familiar with the use of MCDM by their organization, and 19% had experience using MCDM in IA.

Before the formal MCDM methods were introduced, participants completed a holistic assessment of the policy options (ranks of 1–7 and ratings of 0–100). Although time limits presented a challenge, each method was then explained to provide participants with the conceptual understanding necessary to answer the questionnaires that we used to elicit information needed to apply the methods. Additionally, workshop organizers were available to answer questions individually throughout the process to help ameliorate any confusion about the methods or evaluations. After completion of the method questionnaires, MCDM results were calculated and returned to the participants for each group of

methods (weighting, deterministic ranking, uncertainty ranking). Participants then reviewed and revised their weights for the four attribute case (revised weights). They also compared the original holistic assessment to the MCDM method ranks for the deterministic and uncertain cases. The elicitation portion of the workshop then concluded by asking participants to provide revised holistic ranks for the policies (final holistic assessment).

In addition to elicitations, round robin or “nominal group” discussions (Delbecq *et al.*, 1975) were held to record participant views on the methods and their application to IA and to gain insight into the thought processes behind their answers to the questionnaires. In the nominal group discussions, each participant was able to express his/her views for a limited time. After all participants had a turn, each person was given another opportunity to speak. This continued until all views were expressed or time limitations prevented further discussion. At the close of the workshop, participants were asked to complete an evaluation questionnaire in which they rated each method on a scale of 1 (worst) to 5 (best) for a variety of criteria (e.g., ease of use, ability of each method to increase confidence in decision), and to provide suggestions for potential uses for the methods. Similar questionnaires have been used previously to evaluate MCDM methods (e.g., Hobbs and Meier, 2000; León, 1997; Zapatero *et al.*, 1997).

2.4. Limitations of experimental design

This results of this experiment must be viewed in light of its limitations. In particular, the internal validity of the results may suffer from limitations (e.g., small sample size) that preclude the control for alternate hypotheses (Adelman, 1991). We hypothesize that differences in method results can occur because of the fundamental differences in the types of responses each method elicits from the participants. However, it is also possible that the exact wording of questions or even misunderstanding by the participants may have affected our results. As another example, variations in method results also may have been due to an order effect (e.g., simpler methods were presented before more advanced ones). Order effects can arise because the process of completing MCDM exercises, and related discussions can provide insights to the participants. Therefore, the results of methods may diverge due to participants learning and changing their opinions about alternatives or

attributes, rather than because of a true difference in methods.

Although a rigorous experimental design was not possible due to sample size and time limits, the results can nevertheless be helpful to those who would apply MCDM to IA. The experiment's external validity was increased by the use of participants who may actually use MCDM methods to analyze IA results for climate change policy. An identification of differences in method results and appropriateness is useful to potential users by documenting possible method advantages and disadvantages and alerting them to the possibility that different techniques can lead to divergent conclusions. Finally, general insights voiced by the experts regarding how MCDM could be used in IA are an important outcome of this research, and are not subject to these limitations.

3. HYPOTHESES, RESULTS, AND DISCUSSION

This section presents hypotheses tested by the experiment and the associated results. Each of the hypotheses was formulated to illuminate differences among the methods (e.g., perceived ease of use and appropriateness, validity, and results).

Hypothesis 1.

MCDM methods differ in their ease of use and in their ability to aid decision-making in integrated assessment.

This hypothesis addresses the “user-friendliness” of different MCDM methods and whether the methods are appropriate for actual climate change decision-making. After all methods had been applied and results distributed to the participants, evaluation questionnaires asked participants to rate each method from 1 (worst) to 5 (best) for a variety of questions. Figures 1a and b provide the average participant evaluation of each method for these categories.

For the deterministic methods (Figure 1a), holistic assessment was rated higher than all other methods in all categories for which it was evaluated. Holistic assessment was rated significantly higher than other methods for “ease of understanding concepts” (Wilcoxon matched-pairs signed rank test p -values < 0.03), “makes sense” ($p < 0.03$), “skills reasonably acquired” ($p < 0.05$), and “amount of effort required

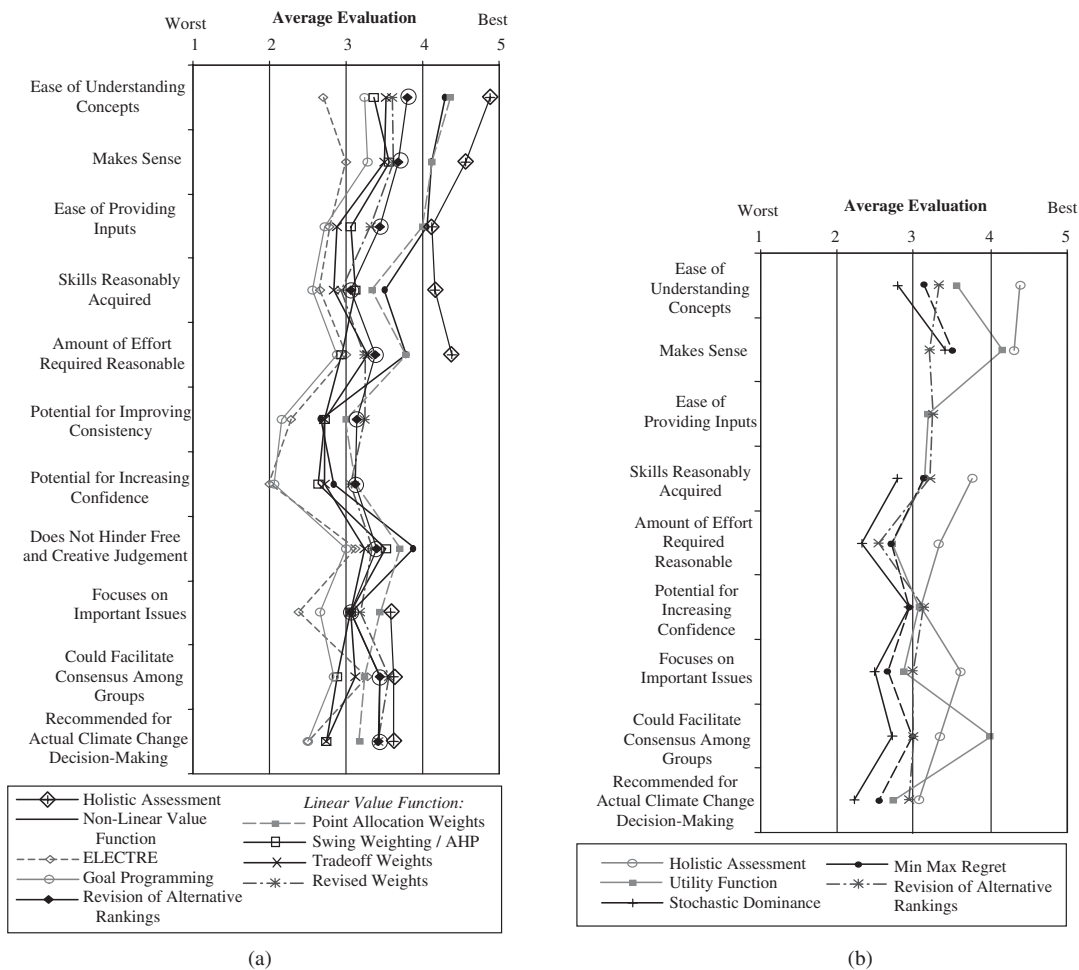


Figure 1. (a) Participant evaluation of certainty ranking methods; (b) participant evaluation of uncertainty ranking methods.

reasonable” ($p < 0.03$). In contrast, goal programming and ELECTRE received poor evaluations across all categories. However the participants suggested that ELECTRE could facilitate consensus among groups better than some other methods. This is likely because the method produces a kernel, rather than definitive ranking, and the kernel can include a diversity of options.

Overall, participants did not believe that the analytical methods had much potential for improving consistency or confidence, as evidenced by the low ratings for these categories. However, this does not indicate that MCDM methods are useless for climate change policy. The primary benefits of MCDM include structuring the problem so that

large amounts of information become manageable, helping users learn how they think about the decision, and exploring value judgements. Participants stated that they gained insights from using the methods, nonetheless they did not believe the methods would help improve consistency or confidence. This raises questions about the benefits of MCDM in IA and suggests that MCDM may be better for helping users think about the decision (e.g., exploring tradeoffs) rather than in forming their actual choice (e.g., improving consistency in decisions).

Holistic assessment also rated highly among the uncertainty methods, receiving the highest rating for most categories for which it was evaluated

(Figure 1b). It rated significantly better than other methods for the “ease of understanding concepts” ($p < 0.04$) and “makes sense” categories ($p < 0.03$, except for the comparison with utility functions). Stochastic dominance had the lowest evaluation for almost all questions. On average, participants felt that all methods had approximately the same potential for increasing confidence.

MCDM methods may be able to facilitate consensus among groups by moving the discussion from alternatives to fundamental objectives and their tradeoffs (Keeney and Raiffa, 1976). By highlighting common interests, the methods discourage focusing on a preferred alternative (Raiffa, 1982). Consistent with this notion, participants believed that the utility function “could facilitate consensus among groups” better than all other uncertainty methods, but not significantly so. However, this method had only the third highest score for “recommended for greenhouse gas evaluation” (out of five methods). This suggests that the ability of an MCDM method to help people better understand their own preferences may be more important than its ability to facilitate negotiation and group consensus.

For the “recommended for actual climate change decision-making” question, holistic assessment received the top average rating for both deterministic and uncertainty methods. The high rating of holistic assessment may indicate the desire of the users to retain control of the decision process. For uncertainty methods, the second highest rated method was reconciliation of alternative rankings, in which users consider and resolve the results of two or more methods. For deterministic methods, the second highest evaluation was a tie between linear value function with revised weights and, again, reconciliation of alternative rankings. This supports the concept that use of multiple methods is beneficial, because participants recommend both use of revised weights (reconciliation of results from several weighting methods) and reconciliation of ranking results from several methods.

Also supporting the use of multiple methods is the fact that except for stochastic dominance, each MCDM method was most highly recommended for decision-making by at least one person. No method dominated the others. For instance, of the 14 people who recommended at least one deterministic approach over the others, 9 rated holistic rating most highly, while 3, 3, and 8 persons gave their highest recommendation to ELECTRE, Goal

Programming, and some version of additive value functions, respectively (some persons used ties, giving more than one method the highest recommendation). Meanwhile, each of these four approaches was also given the lowest recommendation by at least one person. Only one participant strictly preferred holistic assessment to all other approaches.

Participants also ranked three weighting methods (1 for most preferred, 3 for least preferred). All three procedures had roughly the same mean rank: 2.04 for swing weighting/AHP, 2.07 for tradeoff weighting, and 1.89 for direct point allocation. This contrasts with earlier studies where the tradeoff method was relatively disliked (e.g., Hobbs and Horn, 1997).

Hypothesis 2.

MCDM methods have different predictive validities.

MCDM results should be valid (i.e., they should reflect decision-makers’ actual preferences). Because preferences are fundamentally subjective and often imprecise, there exists no universally accepted objective measure of validity (Hobbs, 1986; Larichev, 1992). One type of validity is “predictive validity”—defined here as a method’s ability to predict the final holistic (unaided) ranking of alternatives after an iterative process of applying several MCDM methods and reviewing results (e.g., Corner and Buchanan, 1997; Hadley *et al.*, 1997; Hobbs and Horn, 1997; Lai and Hopkins, 1995). By this definition, a method has a high validity if its results correctly anticipate the user’s final (and presumably most informed) preferences, perhaps because it has helped the user construct them. Note, however, that this does not suggest that methods with poor predictive validity are ineffective at providing insights into the problem.

To evaluate the hypothesis that some methods have higher predictive validity than others, we examined Spearman’s correlations between each method’s policy rankings (1–7) and the final holistic rankings. Such an “intermethod correlation” is defined as the correlation between two methods’ results for a specific user, averaged across all users: $(1/W \sum_{a=1}^W \text{cov}(r_{sa}, r_{ta}) / \sigma_{r_{sa}} \sigma_{r_{ta}})$ where r_{sa} and r_{ta} are the participant a ’s rankings for methods s and t , respectively; $\text{cov}(\)$ is the covariance; and W the number of users (Figure 2).

The deterministic methods differ in their ability to predict the final holistic assessment, sometimes

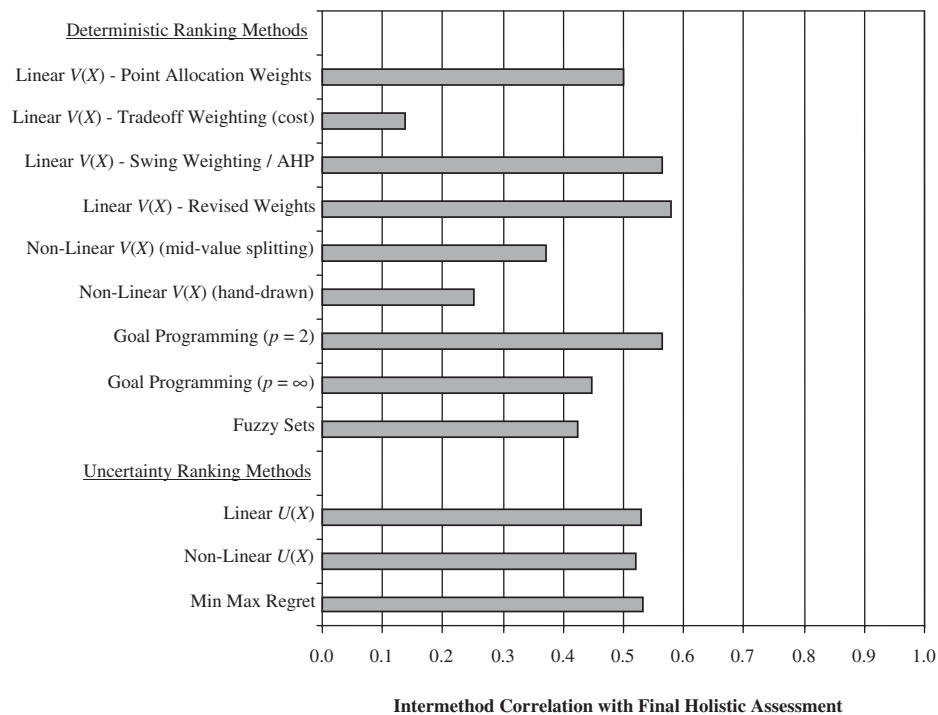


Figure 2. Predictive validity analysis: correlations of method results with final holistic assessment (Note: Deterministic and uncertainty methods were correlated with the deterministic and uncertainty final holistic assessments, respectively. Revised weights used except where specified.).

significantly so. Additive value functions with revised weights and linear single attribute value functions performed best, with swing weighting/AHP and goal programming (exponent = 2) being close behind. However, no method's ranks are highly correlated with holistic assessment's ranks (highest value = 0.58), consistent with previous research (Edwards, 1977; Hobbs, 1986; Hobbs *et al.*, 1992), von Winterfeldt and Edwards, 1986). Possible explanations are that the methods oversimplify sophisticated subjective decision processes or, alternatively, that methods result in more systematic, balanced assessments; of course, Figure 2 cannot distinguish between these hypotheses. Meanwhile, predictive correlations for the linear value function with tradeoff weights are significantly lower than for the linear value functions with point allocation weights, swing weighting/AHP weights, or revised weights ($p < 0.0013$). Surprisingly, the predictive validities for the two non-linear single attribute value function methods (mid-value splitting and hand-drawn value func-

tions) are statistically less than linear single attribute value functions (all with revised weights) ($p < 0.006$).

For uncertainty methods, however, all methods have approximately equal predictive validity (Figure 2). The correlation for all uncertainty methods is 0.52–0.53, and none are statistically different. Thus, our hypothesis, that MCDM methods have different predictive validity, holds for the deterministic methods but not for the uncertainty methods. Assumptions of risk preference differ for each uncertainty method. Min max regret avoids extreme negative outcomes and thereby assumes risk aversion, maximize expected utility assumes risk neutrality when a linear utility function is used, and first-order stochastic dominance makes no risk preference assumptions. Figure 2 indicates that no one representation of risk attitudes did a better job of predicting holistic evaluations than another, and that choice of weighting method affects predictive validity more than choice of risk attitude. However, risk

attitudes should increase in importance if the degree of uncertainty is increased.

The predictive validity of the ELECTRE and stochastic dominance methods cannot be compared in the above manner because they do not yield a full ranking of alternatives. However, we examined their predictive validity by comparing their incomplete rankings to those of holistic assessment. ELECTRE I provides information about whether one alternative “outranks” another. We compared the “kernel” (set of non-outranked alternatives out of the original 7 policies) to the final holistic assessment rankings. The average kernel size for ELECTRE was 3.3. The final holistic assessment’s highest rank (1) was in the ELECTRE kernel for only 50% of the participants, and 50% of the ELECTRE kernels contained the holistic assessment’s lowest ranked alternative (7). This unpromising result could have occurred by chance. One-third of the kernels contained both the highest and lowest ranked alternatives.

Meanwhile, stochastic dominance also does not provide a ranking of alternatives, but shows whether an alternative dominates another. Figure 3 depicts the cumulative distributions for one participant’s utility functions for each policy. In this case, stochastic dominance is evident. For instance, the policy of relaxed SO₂ standards is first-order stochastically dominated by the biomass promotion policy. For this participant, five alternatives are first-order stochastically dominated, therefore the kernel size is two (the \$150 and \$300/ton CO₂ tax policies). The average kernel size for all participants was 3.2 for first-order stochastic dominance and 1.7 for second-order stochastic dominance. Results from this method differed from those of the final holistic assessment. The final holistic assessment’s top ranked alternative was non-dominated for only 28% of the participants, which could have occurred by chance. Meanwhile, the nuclear alternative had an average holistic assessment rank of 5.2 (out of 7), but was non-dominated in 72% of the stochastic dominance results. The results of the last two paragraphs demonstrate that ELECTRE and stochastic dominance produce results that are distinctly different from holistic assessment (i.e., neither has high predictive validity).

In sum, none of the MCDM methods had high predictive validity. It may appear that the methods are not useful if they do not match the user’s final holistic preferences. Nevertheless, insights gained

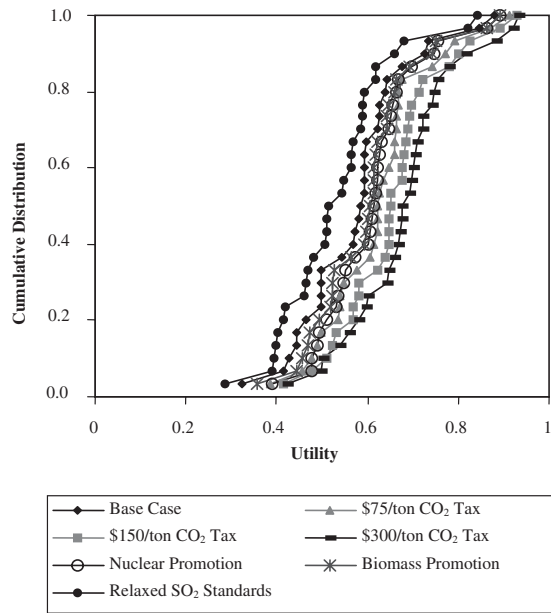


Figure 3. Example cumulative distribution functions for utility function (used for stochastic dominance method).

from using MCDM methods were evidenced by differences between holistic rankings completed before and after applying the MCDM methods and reviewing the results. Initial and final deterministic holistic assessment differed for 94% of participants and were not highly correlated (average correlation of 0.55 between each person’s pair of rank sets), with 53% of participants choosing a different top-ranked alternative. However, these differences between participants’ initial and final holistic assessments also could have occurred by chance (i.e., holistic assessments made one day apart without intervening MCDM assessments might differ just as much); our experiment was not designed to control for this alternative explanation. Meanwhile, holistic rankings of uncertain alternatives taken before and after uncertainty MCDM methods differed for 68% of participants. Yet most differences between participant’s two sets of uncertain holistic evaluations were small (mean correlation of 0.81), so we cannot claim that much learning took part as a result of the uncertainty MCDM exercises.

As we noted previously, the numerical analyses of method results must be viewed in terms of the experiment’s limitations. However, the use of

several methods can help people focus on their objectives and provides an opportunity to think about the problem in different ways (Hobbs and Horn, 1997). This implication for MCDM use is applicable regardless of whether inconsistencies between methods arise from order effects or inherent differences in the methods. In this sense, the use of several MCDM approaches can aid decision-making more than a single approach (Corner and Buchanan, 1997; Simpson, 1996).

Hypothesis 3.

Methods will differ in the convergence of different persons' results.

We hypothesize that some methods will produce more similar results across all persons than other approaches. Such methods might be interpreted as promoting consensus, or alternatively, as obscuring genuine differences of opinions. In order to test this hypothesis and identify such methods, interperson correlations were compared for different methods. We define "interperson correlation" to be the correlation between policy rankings for a given method for a pair of users, averaged across all pairs of users:

$$\frac{1}{\binom{W}{2}} \sum_{a=1}^W \sum_{b>a}^W \frac{\text{cov}(r_{sa}, r_{sb})}{\sigma_{r_{sa}} \sigma_{r_{sb}}}$$

where r_{sa} and r_{sb} are the vectors of policy ranks from method s for participants a and b , respectively. Interperson correlations can be calculated for weight sets as well.

Formal MCDM methods generally have higher convergence of different persons' results than do holistic evaluations (Edwards, 1977; Hobbs and Meier, 2000; von Winterfeldt and Edwards, 1986), perhaps because participants simplify the problem in holistic assessments by focusing on a few criteria. Surprisingly, this was not the case with our experiment, as holistic assessment had a *higher* interperson correlation than most other methods. Interperson correlations ranged from 0.02 (linear utility function) to 0.79 (linear value function with tradeoff weighting) (Figure 4). The correlation of tradeoff weighting results was high because most participants had similar weights for this method, with most placing a high weight on cost (average weight on cost 53%). This correlation is significantly higher than those of other deterministic methods except fuzzy sets ($p < 0.001$). In contrast,

interperson correlations for the goal programming methods (exponent $p=2$ and ∞) are statistically lower than correlations for holistic assessment and the linear value function with point allocation weights, tradeoff weights, or revised weights ($p < 0.03$).

The fuzzy set method chooses the policy that has the highest degree of membership in the set "good decision," as measured by a multivariate fuzzy set membership function. Because of its "min max" operator (see Appendix A), this method tends to choose alternatives that perform moderately well on all alternatives; consequently, the interperson correlation (0.73) was the highest of all methods except the linear value function with tradeoff weights. For all persons, the fuzzy set method showed that the \$75/ton CO₂ tax, \$150/ton CO₂ tax, and nuclear power promotion policies (none of which had the worst attribute value for any attribute) performed better than the base case, \$300/ton CO₂ tax, and biomass promotion options (each of which had the worst possible value for some attribute).

For uncertainty methods, the linear utility function and min max regret methods' interperson correlations are statistically different from the other methods' interperson correlations ($p < 0.001$), with the linear $U(X)$ resulting in the least consensus and min max regret achieving the most. Such differences in interperson correlations have implications for decision-making, as differences of opinion among users present opportunities for discussion and learning.

Hypothesis 4.

Climate change experts are subject to classic weighting biases.

The specific bias addressed in this section ("splitting bias") has been previously identified as a problem for directly assessed weights, such as point allocation (von Winterfeldt and Edwards, 1986). (Another bias associated with tradeoff weighting is mentioned under Hypothesis 5.) The splitting bias occurs when more aggregate weight is given to an attribute when it is "split" into several attributes (e.g., dividing "environment" into SO₂ emissions, nuclear waste, and climate) than when it is categorized as a single attribute (e.g., a single weight is assessed for "environment"). This bias can be tested if weights are chosen by both hierarchical and non-hierarchical methods. In hierarchical point allocation, the

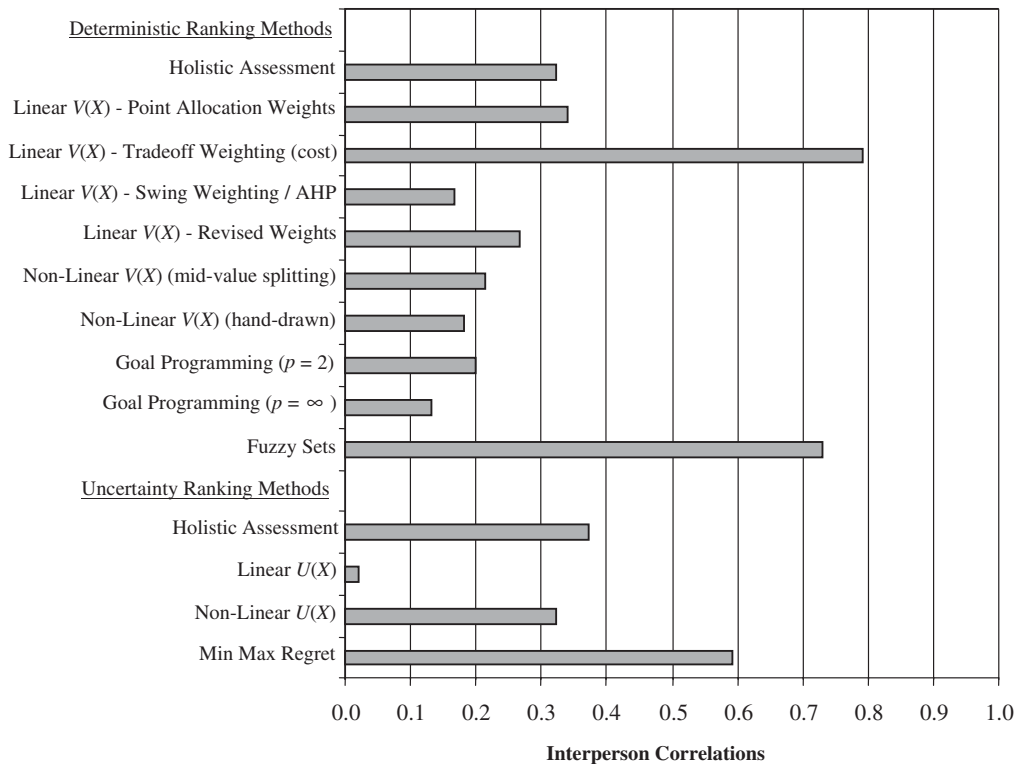


Figure 4. Interperson correlations for MCDM methods.

user first assigns weights to broad categories, then allocates each category's weight among subcategories. In contrast, the non-hierarchical point allocation method presents all attributes simultaneously, without categorization. Criterion information is processed differently when presented in parallel than in sequence (Korhonen *et al.*, 1997). Experiments have found that a hierarchical approach results in less weight for attributes that belong to categories with many attributes than does a non-hierarchical approach (e.g., Eppel, 1992; Hobbs and Meier, 2000; Stillwell *et al.*, 1987).

Most previous experiments that demonstrated splitting bias involved inexperienced subjects (e.g., students). We hypothesize that this bias applies even to the experts in this experiment, and therefore would occur in the actual application of the methods. Although weighting biases are widely recognized from experimental evidence, they are rarely addressed in applications (Pöyhönen and Hämäläinen, 2000).

The participants were divided into two groups, each using a different value tree (Figure 5). Half of the participants applied the non-hierarchical approach by allocating 100 points among the six attributes, with more points indicating a higher significance. The other half applied the hierarchical approach by first allocating 100 points between cost (x_1) and the environmental category; then allocating 100 points among the environmental categories of SO₂ emissions (x_2), nuclear waste (x_3), and climate; and finally allocating 100 points among the climate attributes of temperature increase (x_4), sea-level rise (x_5), and ecosystem stress (x_6). The weight of an attribute on the lower level of the hierarchy was then calculated using the point allocations for each higher level (e.g., weight on sea-level rise = (points for sea-level rise) × (points for climate) / 100 × (points for environment) / 100).

The splitting bias would predict that the sum of weights for x_4 , x_5 , and x_6 would be higher for the non-hierarchical point allocation than for hier-

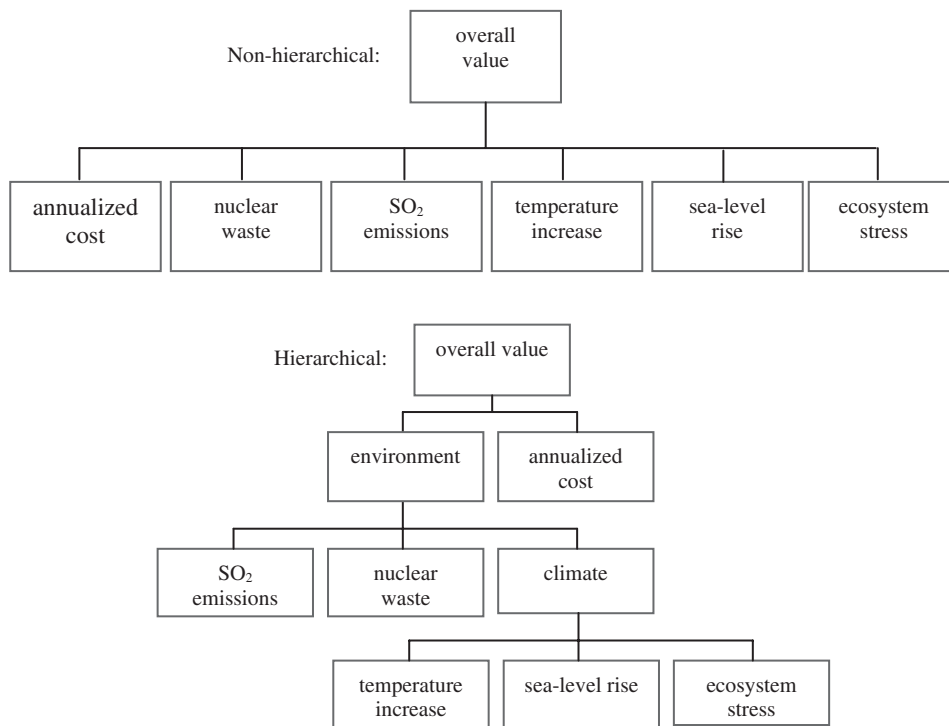


Figure 5. Alternative value trees for splitting bias analysis of point allocation weights.

archical approach. Indeed, the average total weight for attributes x_4 , x_5 , and x_6 for the non-hierarchical method (42%) is significantly higher than that for hierarchical approach (21%) (Mann–Whitney test p -value < 0.008). Thus, we have confirmed that climate change experts are subject to the classic “splitting bias.”

Hypothesis 5.

Weights chosen by different methods differ significantly.

Previous experiments have compared various weighting methods (e.g., Belton, 1986; Hobbs *et al.*, 1992; Pöyhönen and Hämäläinen, 2001). We hypothesize that *how* weights are chosen (the method used) can be as important as *who* chooses the weights, because different methods use different operational meanings of attribute importance. If this is true, the choice of methods is important, as is the issue of validity. If, on the other hand, intermethod correlations are higher than interperson correlations, who performs the assessment matters more than which method is used.

Intermethod correlations between weight sets from various weighting methods show that different approaches did in fact yield different results (Table II). The tradeoff weighting method almost uniformly produced a higher weight for cost (average weight on cost 53%). Earlier experiments have shown that in tradeoff questions, the attribute users are asked to adjust tends to receive higher weight (the scale compatibility bias; e.g., Borcherding *et al.*, 1991), and our result is consistent with that conclusion. The high weight on cost causes tradeoff weights to have low correlations with other methods and also to have the highest interperson correlation (0.46 among weight sets chosen by different people, statistically higher than other interperson correlations, Wilcoxon signed-rank test p -values < 0.001). In contrast, interperson correlations for weights from other weighting methods were essentially zero. Point allocation, swing weighting/AHP, and the final choice of weights (revised weights) yield similar weights: intermethod correlations between each pair of these three methods (0.78, 0.88, 0.89) are statistically higher than intermethod correla-

Table II. Intermethod and interperson correlations of weights from weighting methods

Weighting method	Point allocation	Swing weighting/AHP	Tradeoff weighting	Final choice of weights (revision of weights)
<i>Average intermethod correlations of weights chosen by the same person</i>				
Point allocation	—	0.78	0.58	0.89
Swing weighting/AHP		—	0.39	0.88
Tradeoff weighting			—	0.49
<i>Average interperson correlations of weights for the same method</i>				
	0.05	0.03	0.46	0.06

tions between tradeoff weights and any of these methods (0.39, 0.49, 0.58) ($p < 0.04$). This indicates that tradeoff weighting elicits very different weights.

Comparison of intermethod correlations (range: 0.39–0.89) and interperson correlations (range: 0.03–0.46) indicates that *who* assesses the weights and *which method* is used both have a significant impact on weights. This conclusion is subject to the experiment design limitations discussed earlier. We cannot determine the specific cause of variations in weights. However, we do venture to conclude that using more than one weighting method is preferred to a single approach, because methods frame the problem differently, and thus yield different results and provide opportunities for reflection and learning.

Hypothesis 6.

Ranks resulting from different methods differ significantly.

Similar to the weighting results just described (Hypothesis 5), we hypothesize that which ranking method is used is as important as who performs the method. We tested this hypothesis by comparing the intermethod and interperson correlations for different ranking methods (Table III). If *which method* is used is as important as *who* performs the method, the intermethod and interperson correlations will be of the same magnitude.

The differences among deterministic policy ranks in Table III are the result of the following choices of who or how to rank (in decreasing order of importance):

- choice of person (interperson correlations range: 0.17–0.79, all but two being ≤ 0.34);

- choice between goal programming, value function, or fuzzy sets (intermethod correlations range: –0.11 to 0.56);
- choice between holistic assessment or MCDM method (intermethod correlations range: 0.14–0.58);
- choice between tradeoff weights or more direct weighting methods (intermethod correlations range: 0.51–0.57);
- choice between linear or non-linear value function (intermethod correlation: 0.66);
- choice between AHP or point allocation weights (intermethod correlation: 0.73); and
- choice of exponent (p) in goal programming (intermethod correlation: 0.85).

Factors that affect ranks under uncertainty, in decreasing order of importance, are (Table III):

- choice of person (interperson correlations range: 0.02–0.59);
- choice between holistic or analytical method (intermethod correlations range: 0.52–0.53);
- choice between regret or utility function (intermethod correlations range: 0.74–0.79); and
- choice between linear or non-linear utility function (intermethod correlation: 0.93).

The results show that both choice of method and person are very important. The intermethod correlations for deterministic ranking methods vary from –0.11 to 0.85, indicating that various methods produce different results. Comparison of those correlations to interperson correlations (range: 0.13–0.79) indicate that both *who* applies

Table III.

Intermethod and interperson correlations of policy ranks for deterministic methods ^a										
Deterministic ranking method:	Final holistic assessment	Linear $V(X)$				Non-linear $V(X)$		Goal programming		Fuzzy sets
		Weighting Method				Mid-value splitting	Hand-drawn	$p = 2$	$p = \infty$	
		Point allocation	Tradeoff	Swing/AHP	Revised					
<i>Average intermethod correlations of ranks chosen by the same person</i>										
Holistic assessment	—	0.50	0.14	0.57	0.58	0.37	0.25	0.56	0.45	0.42
Linear $V(X)$: point allocation weights	—	0.57	0.73	0.80	—	—	—	—	—	—
Linear $V(X)$: tradeoff weights	—	—	0.51	0.54	—	—	—	—	—	—
Linear $V(X)$: swing weighting/AHP	—	—	—	0.85	—	—	—	—	—	—
Linear $V(X)$: revised weights	—	—	—	—	0.66	0.63	0.45	0.48	0.21	—
Non-linear $V(X)$: mid-value splitting	—	—	—	—	—	0.72	0.17	0.17	—	-0.11
Non-linear $V(X)$: hand-drawn	—	—	—	—	—	—	0.27	0.24	0.00	—
Goal programming: $p = 2$	—	—	—	—	—	—	—	0.85	0.56	—
Goal programming: $p = \infty$	—	—	—	—	—	—	—	—	0.53	—
<i>Average interperson correlations of ranks for the same method</i>										
	0.32	0.34	0.79	0.17	0.27	0.21	0.18	0.20	0.13	0.73
Intermethod and interperson correlations for uncertainty methods										
Uncertainty ranking method:	Holistic assessment	Linear utility function	Non-linear utility function	Min Max regret						
<i>Average intermethod correlations of ranks chosen by the same person</i>										
Holistic assessment	—	0.53	0.52	0.53						
Linear $U(X)$	—	—	0.93	0.79						
Non-linear $U(X)$	—	—	—	0.74						
<i>Average interperson correlation of ranks for the same method</i>										
	0.37	0.02	0.32	0.59						

^aNote: All methods used revised weights except where noted

the method and *which method* is used both can strongly impact results. Similarly, for uncertainty ranking methods, the intermethod correlations (range: 0.52–0.93) and interperson correlations (range: 0.02–0.59) imply that both the user and the choice of method can significantly influence results, with higher variability among people than among methods.

Fuzzy sets, goal programming, and value functions differ greatly in philosophy and assumptions; as a result, the ranks they select diverge strongly from each other (Table III). Which method is more appropriate depends on which set of assumptions seems most valid for a given situation and person. In general, we believe that multiple methods result in better representations of values because they allow people to compare results and resolve differences.

To further explore some method differences, consider the two methods for eliciting non-linear single attribute value functions. These methods also had different results, although their predictive validities are not statistically different. For the mid-value splitting method, the user specifies an attribute value ($x_{0.5}$), that is halfway in desirability between the best and worst values (x_i^{**} and x_i^*) for each of four criteria. A linear or exponential value function was then fit to the three points. The other method asks the user to directly draw a value function on a graph for each of the criteria. Seven percent of the hand-drawn graphs were shapes that were neither linear nor exponential (e.g., S-shaped) which implies that assumptions of linear or exponential value functions may not be valid. An additional 36% of the graphs were linear or exponential but diverged from the value functions implied by the mid-value splitting method, in that the mid-value points ($x_{0.5}$) of the hand drawn and mid-value splitting graphs differed by more than 10% of the potential range. For example, one value function might be concave and the other convex, or one value function slightly concave and the other very concave. The average intermethod correlation between the ranking results of the mid-value splitting method and direct hand-drawn method is 0.72. Seventy-one percent of these intermethod correlations were above 0.8. Considering just the participants whose correlations fell below 0.8, one-fifth of those users' hand-drawn value functions did not fit the assumption of linear or exponential value functions. Thus, different results arise both from the assumption of linear or exponential value functions used in the mid-value

splitting method and from inconsistent participant responses.

Hypothesis 7.

Visualization methods differ in their ability to aid the decision-making process.

Visualization aids can help users better understand the results of an MCDM analysis, such as tradeoffs among alternatives and how changes in weights affect results. Our experiment explored several standard visualization methods for both deterministic and uncertain results. For deterministic exercises, participants were given a table of attribute values and several visualization aids (bar graphs of attribute values for each policy, Cartesian (XY) plots showing how each policy performs on two attributes at a time, and value path plots). A value path plot is created by normalizing values for each attribute for each alternative, in which the worst value for a given attribute has a normalized value of 0 and the best value has a normalized value of 1. For other values, the normalized value is as follows: $v_i(x_{ij}) = (x_{ij} - x_i^*) / (x_i^{**} - x_i^*)$, where x_i^{**} and x_i^* are the best and worst values, respectively, for attribute i among all alternatives, and x_{ij} is the value for attribute i for alternative j . This allows the viewer to determine the relative performance of each alternative for each attribute. A value path plot is provided in Figure 6 for four of the attributes. It shows, for example, that the \$300/ton CO₂ tax policy performs better than all other alternatives for global temperature change and SO₂ emissions, but performs worse than all others for nuclear waste generation and control cost.

For the uncertainty exercises, participants were given a table of attribute distributions for each option (mean, standard deviation, minimum, maximum), a table of regret values (mean, standard deviation, minimum, maximum), and other visualization aids (Cartesian plots for each pair of attributes that show how each alternative performed under each of the simulations, and box plots). The box plots created for this exercise consist of a central box extending from the 25th percentile ("lower hinge") to the 75th percentile ("upper hinge"); a horizontal line in the box representing the median; "whiskers" which extend from the box to the "lower fence" (equal to the smallest observed value that exceeds $X = \text{lower hinge} - 1.5[\text{upper hinge} - \text{lower hinge}]$) and the "upper fence" (the largest observed value that is no greater than $Y = \text{upper hinge} + 1.5[\text{upper}$

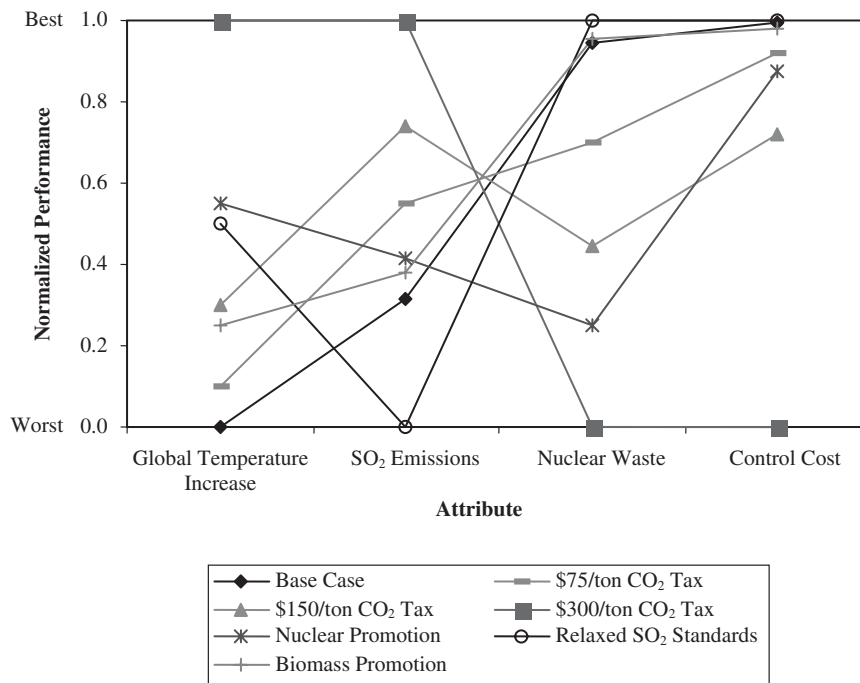


Figure 6. Value path example.

hinge–lower hinge)]; and circles representing outliers (values > upper fence, or < lower fence). Example boxplots are shown in Figure 7.

Participants rated each visualization method (on a scale of 1–5) on several criteria including ease of understanding, whether the participant used the information in completing the exercises, and whether the participant would recommend the visualization method or table for actual decision-making. A score of 5 represented the best possible evaluation, whereas 1 is the worst possible. Means for the visualization methods are shown in Figures 8a and b.

Figure 8a shows that of the deterministic approaches, the table of criteria values performed best on the “ease of understanding” and “reliance of information presented” (meaning more used by participants in the exercises) categories. Interestingly, tables were not the participants’ first or even second highest choice for actual climate change decision-making. The more difficult to understand value path plots were more highly recommended, although their evaluation is significantly better only compared to the *XY* plots ($p < 0.04$).

For the uncertainty visualization methods’ evaluations (Figure 8b), the *XY* plots and table of regret values were significantly easier to understand than the table of criteria distributions or box plots ($p < 0.03$). Nevertheless, the participants relied more heavily on the latter two method ($p < 0.04$). Those two approaches were also more highly recommended for use in actual decision-making than the *XY* plots and tables of regret values ($p < 0.05$).

Hypothesis 8.

Users may be more confident about some value judgments than others, and incorporating precision information can change ranking results.

Although all value judgments are subjective, users may be more sure of some value judgments than others. Yet most MCDM methods require the user to input precise point values (e.g., “How many more times important is criterion *i* than criterion *j*?”). Responses to such questions do not reflect how confident the user is about his/her answers.

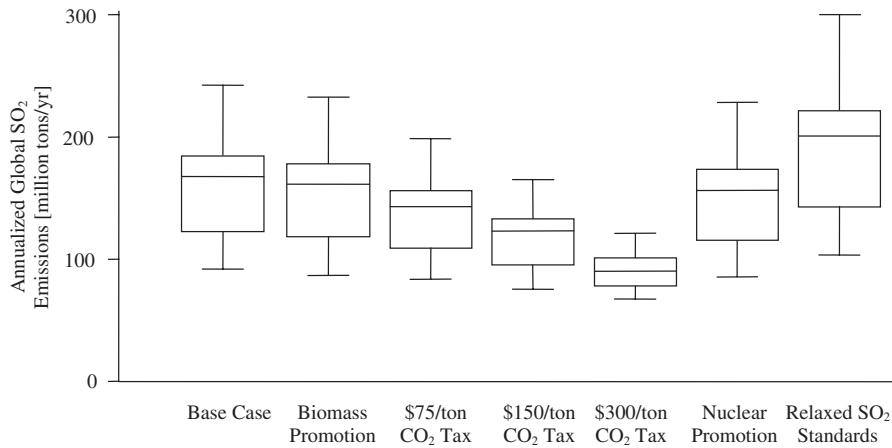


Figure 7. Boxplot examples.

As an alternative to the point estimates traditionally used, we asked the participants to provide a *range* for answers to some exercises in addition to point values. For example, one tradeoff-weighting question asked: “How much would you be willing to spend to lower the year 2050 temperature by 0.1°C? Please provide a range for the answer you provided.” A narrow range indicates that the user is confident about his/her answer.

The point estimates, used alone, provide precise weights for the attributes. The ranges, in contrast, define constraints that the weights must satisfy (e.g., criterion *i* is two to five times as important as criterion *j*). Such ranges could be analyzed in variety of ways. We choose to analyze results as follows: if alternative *A* cannot be better than alternative *B* for any weights that satisfy the ranges provided, then alternative *B* is said to outrank alternative *A* (Sarin, 1977). To determine the outranking relationships, $m(m-1)$ linear programs (LPs) were solved for each person where *m* is the number of alternatives. Each LP determined whether a given alternative *B* outranks another given alternative *A*. In the LP, linear single attribute value functions were assumed, although the method could be applied to non-linear value functions or utility functions. The linear program is as follows:

$$\text{maximize } C_{AB} = V(A) - V(B)$$

- s.t. 1) $\sum_{i=1}^n w_i = 1$, where w_i =weight for attribute *i*, n =number of attributes
 2) $w_i \geq 0 \forall i$

- 3) weighting relationships implied by the ranges provided (e.g., if attribute *i* is 3–5 times as important as attribute *j*, then $3w_j \leq w_i \leq 5w_j$)

If $C_{AB} < 0$, alternative *B* is better than alternative *A* under any *feasible* weights, so *B* outranks *A*. For further discussion of decision-making with incomplete information, see Park *et al.* (1996) and Weber (1987).

This approach can be used to define a kernel of alternatives that are not eliminated by any other alternative. This kernel can be viewed as analogous to the ELECTRE I kernels. By calculating the number of alternatives that outrank a given alternative and the number of alternatives outranked by the given alternative, we can determine the possible ranks (1–7) for each alternative. For example, if policy *A* outranks two other alternatives but is outranked by one other alternative, the feasible ranks for policy *A* range from 2 to 5. Examples of outranking relationships are provided in Figure 9. An arrow indicates that one alternative always outranks another. (Note that we have not shown all arrows; if *A* outranks *B*, and *B* outranks *C*, we omit the arrow from *A* to *C*, since it can be proven that *A* must also outrank *C*.) For instance, in Figure 9, Example 1, the SO₂ emissions alternative always outranks nuclear promotion and the \$75, \$150, and \$300/ton CO₂ tax alternatives. The SO₂ emissions, biomass, and base case alternatives are incomparable (none outrank another). Similarly, the outranking results from ELECTRE I can be used to determine a range of possible ranks for each alternative.

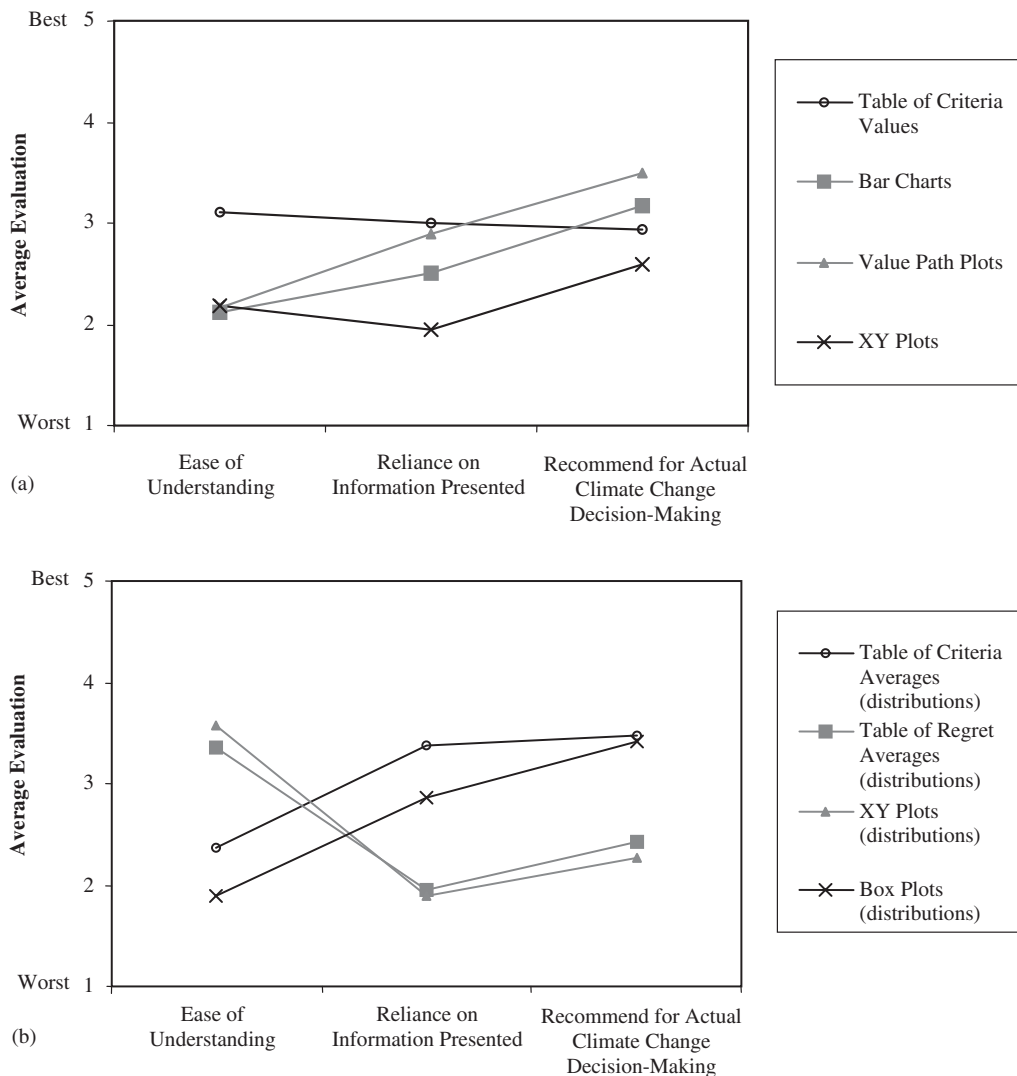


Figure 8. (a) Participant evaluation of deterministic visualization methods; (b) participant evaluation of uncertainty visualization methods.

Kernel sizes and ranges of ranks from ELECTRE I and tradeoff questions with ranges are compared in Table IV.

The average kernel sizes from the tradeoff weighting method with ranges and the ELECTRE method are not statistically different. The tradeoff method with ranges reduced the set of non-dominated alternatives by more than half, while allowing the decision-maker to express their imprecision as well as point estimates. This outcome implies that results from point estimates of weights

(yielding a complete ranking of alternatives) may inspire a false sense of precision in the rankings. In other words, our participants were not certain enough of their answers (did not have sufficiently narrow weighting relationships) to yield complete rankings of alternatives in all cases, although use of a point estimate forces this complete ranking. Advantages of explicit consideration of imprecision include: (1) a more accurate reflection of the state of mind of users, and (2) a screening out of a significant number of alternatives.

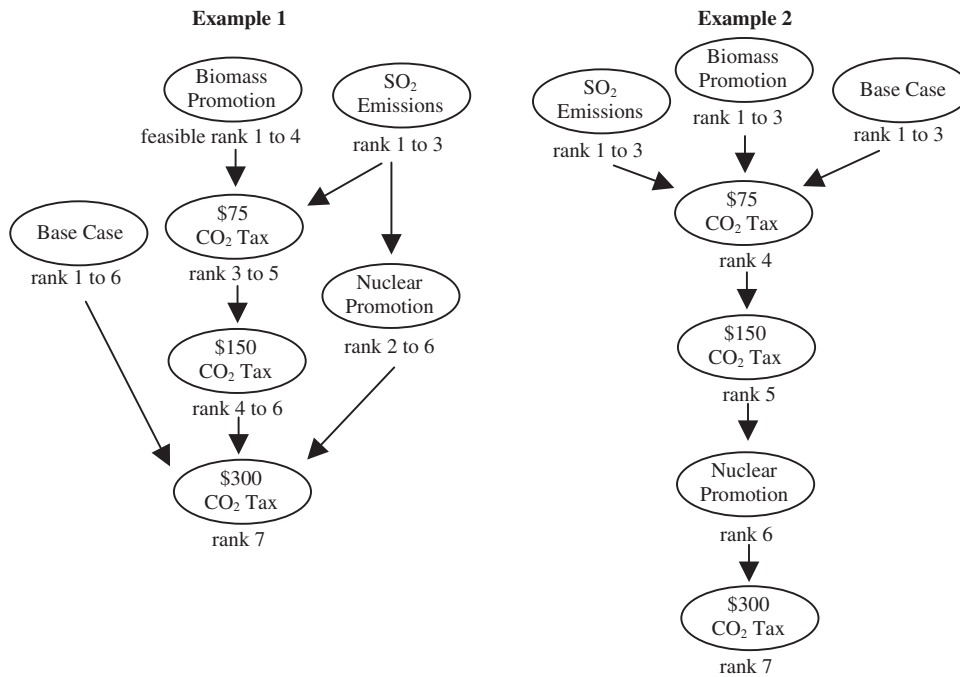


Figure 9. Example outranking relationships resulting from LP analysis of tradeoff questions with ranges.

Table IV. Results from linear value function using tradeoff weighting methods with ranges and ELECTRE I

Analysis of 7 alternatives:	Tradeoff Weighting (linear value function)	ELECTRE I
Kernel size: average (standard deviation)	2.8 (1.7)	3.3 (2.0)
Number of possible ranks for a given alternative: average (standard deviation)	3.2 (2.1)	4.6 (2.0)

4. INSIGHTS INTO THE APPLICATION OF MCDM TO CLIMATE CHANGE AND INTEGRATED ASSESSMENT

The workshop provided an opportunity for climate change experts to state how MCDM methods could be applied to climate change policy and IA, and what the obstacles to such applications might be. Many of their comments are applicable to a wide range of decisions beyond climate change policy. The following conclusions were synthesized from comments written on questionnaires and offered during the formal

discussions. The first conclusion focuses on the formidable task of addressing expert disagreement with respect to climate change impacts, and suggests how MCDM could be used to help identify where such disagreement affects policy choice. The second highlights the challenges facing climate change experts who want to use methods that require decomposition of the problem (i.e., methods that require the user to think of interrelated attributes as independent).

4.1. To apply MCDM methods, decision-makers must have confidence in attribute values, which can be especially challenging in the uncertain realm of climate change

The attribute values we provided were for the purpose of workshop exercises and were not intended to be definitive. However, whether such estimates were to be *believed* was crucial for some participants, even though several workshop exercises explicitly incorporated uncertainty. For example, one participant was skeptical of model results because climate change models oversimplify the complex reality, while another person stressed disbelief about estimates beyond the immediate future, such as predictions of consequences in the year 2100. The consequences of

particular policies are uncertain, as can be seen from present debates on climate change science. Uncertainty results in disagreement among models and experts with respect to the nature, distribution, and timeframe of impacts (Morgan and Keith, 1995). This can be problematic for the use of MCDM with integrated assessment, as the methods require some measure of how each alternative will perform on selected attributes.

Although decision-makers may not agree on the outcomes of alternative scenarios, each person or decision-making entity must have some assumptions concerning what will or might happen if a certain action is taken in order to compare alternatives. Otherwise, the comparison of alternatives becomes difficult or impossible. Incorporating uncertainty can help address this issue, but still experts may disagree on the uncertain impact (e.g., probability distribution for an impact) for a given policy. In other words, including uncertainty in the manner we used (i.e., defining a single probability distribution for each attribute for each policy) is not enough to address expert disagreement in climate change policy.

MCDM could be used to explore how expert disagreement affects the choice of alternative in several ways. Users' value judgments could be applied multiple times, once to each set of attribute values or distributions derived from various IA models or expert opinions, representing a range of possibilities. A comparison of the resulting rankings would identify where model or expert disagreements affect choice of policy. Another approach is to allow users to specify attribute values and distributions according to their own beliefs, and then apply the MCDM methods. Although this may introduce bias, it would do so in a documentable way, and would help reveal how different expert opinions affect the rankings. For example, each user may have different value paths (Figure 6) because they disagree on the impacts, yet their policy ranks may or may not differ.

4.2. MCDM methods that involve decomposition of the problem require the user to separate value judgments from how the system functions, and may be difficult for some users

Climate change decisions are complicated by uncertainties, multiple decision-makers, and social implications, which can be problematic for conventional decision analysis tools that decompose the decision problem into a simpler frame-

work (Jaeger *et al.*, 1998). For instance, MCDM methods often require the user to decompose the problem by valuing a change in one or two attributes while holding other attributes constant (e.g., as in tradeoff weighting). For climate change experts who are highly familiar with feedbacks among attributes, such decomposition proved to be a difficult cognitive task. Some participants felt that the decomposition employed by such analytical MCDM approaches is a serious weakness because it treats problems in isolation, not in the broader context of feedbacks and other decisions. This difficulty may in part explain some of our participants' poor evaluations of the tradeoff weighting method, and their preference for holistic evaluation (Figure 1). One participant said, for example, that he "couldn't buy pair-wise [comparisons] with all else being equal because of [his] preconceptions about how the world works." This suggests that such methods require more time than others to allow the user to feel comfortable with a process that separates value judgments of the attributes from how the system functions.

However, users may gain some understanding because decomposition would encourage them to think about the problem differently, considering the possibility of value independence (e.g., preference independence) where there may still be physical dependence (the performance of attributes are correlated). However, this requires a willingness and ability on the part of the user to think of the system in ways that may not be possible in reality (e.g., increase in temperature with no change in sea-level rise). Ironically, the more knowledgeable a user is about the system, the more challenging this will be. For some decision-makers, methods that do not require decomposition may be more appropriate.

5. RECOMMENDATIONS FOR MCDM EXPERIMENTS

In retrospect, we can identify several ways in which the workshop design can be improved. We hope that future experiments involving MCDM methods can learn from our experience.

5.1. Attributes and general objectives must be clearly defined and encompass the decision-makers' values

To allow greater flexibility, we did not specify the general objectives of climate policy (e.g., to

minimize the net global costs associated with greenhouse gas emissions, to minimize just U.S. costs, or to appease constituents); rather, we let each participant choose weights consistent with their view of the appropriate objectives. However, some participants found the exercises hard to complete without explicit statements about *who* is the decision-maker (e.g., governmental agency) and their objectives (e.g., individual versus societal goals). In addition, the data presented to the participants omitted information some workshop participants considered valuable or necessary (e.g., distribution of costs, nuclear waste disposal methods, fate of tax revenue). Naturally, these participants found the MCDM exercises difficult. This may explain why holistic rankings differed so much from those of other methods (Figure 2); the former may include unstated objectives, while the later were based only on our IA model results.

In an actual policy-making, much effort would go into determining what criteria are important to each decision-maker. This was beyond the scope of our experiment, therefore we were unable to completely address concerns about attribute completeness and specificity. Participant feedback stressed the importance of explicit and comprehensive sets of attributes. While this observation may seem obvious, it demonstrates the need for clearly determining early on what decision-makers deem important. This may be especially challenging in the realm of climate change, as decision-makers' goals are often divergent or unspecific, and who the decision-makers themselves are might not be obvious. In retrospect, participants might have felt more comfortable with the process had we spent time prior to the workshop identifying what objectives were important to participants rather than choosing what we believed to be representative ones. Even though the workshop involved a hypothetical decision and was focused on MCDM methodology rather than the policy decision, some participants found the application of methods to be difficult because they did not participate in the selection of attributes.

5.2. Anchoring, although a potential source of bias, may be necessary or helpful for making decisions

People often provide numerical estimates by modifying numbers that others have provided or suggested, which is also known as the "anchor and adjust" heuristic (Kahneman and Tversky, 1973). For example, a survey which asks, "How much are you willing to pay to eliminate 1 ton of CO₂

emissions: less than \$25, \$25 to \$50, or more than \$50?" will generally elicit lower numbers than a question which uses "less than \$100, \$100 to \$200, or more than \$200?" To prevent such bias, we deliberately did not provide anchors (e.g., U.S. gross national product as an anchor for annualized cost). However, several workshop participants commented that the exercises were difficult without such reference points. For example, one said it was "hard to relate to the numbers." Therefore, even though anchors are a known source of bias, it may be necessary to provide them in order to help participants understand the attributes of the alternatives. If this is true with the knowledgeable experts participating in our workshop, it will be even more applicable if diverse stakeholders or the public at large are involved. Offering a diverse range of possible anchors may lessen the possible bias. If sample sizes permit, offering different groups different anchors could allow for control of and testing for an anchoring effect.

5.3. The experiment's schedule should be designed so that it does not overburden participants

The two-day workshop included presentations of methods, over a dozen MCDM exercises, group discussions, completion of evaluation forms, and breaks and meals. Although this ambitious schedule allowed comparison and evaluation of many methods, it had the drawback of fatiguing the participants. In retrospect, the workshop might have benefited from a less demanding schedule, resulting either from a longer workshop or the application of fewer methods. While this would have resulted in less data or required more time, it would help ensure that participants understood the methods and had adequate time to complete MCDM exercises and review results in a more relaxed atmosphere.

6. SUMMARY AND DISCUSSION

A workshop with 20 climate change experts, IA practitioners, and policy-makers explored the application of MCDM methods to climate change policy and IA. Participants applied various MCDM methods (weighting, deterministic ranking, and uncertainty ranking) to a hypothetical climate policy decision. MCDM methods were compared and evaluated through analysis of method results, opinion questionnaires, and nominal group discussions. A variety of conclusions

can be drawn from the workshop results concerning eight hypotheses about method validity, appropriateness, and differences in choices. In summary, our analysis shows that the methods varied in their predictive validity and convergence of different persons' results. The weighting and ranking methods were found to often yield divergent results for attribute weights and policy rankings, respectively. In particular, goal programming, additive value functions, and fuzzy set analysis often yielded very dissimilar policy ranks, and tradeoff weights gave distinctly different results than other weighting methods. Although tradeoff weighting, in theory, should be most valid (as it directly measures marginal rates of substitution, as required by value and utility theory methods), its weights had the lowest predictive validity while the users preferred weight sets chosen directly or via a hybrid swing weighting/AHP procedure.

Another hypothesis considered whether climate experts are subject to the classic splitting bias when directly selecting weights. This hypothesis was confirmed. Several techniques for visualizing tradeoffs and uncertainties also were compared in the workshop. For use in actual climate change decision-making, participants recommended visualization methods that they also found were harder to understand, suggesting that participants believe the extra effort needed to comprehend more complex visualizations is worthwhile. The final hypothesis concerned the usefulness of information on precision of weights. Participants provided information about how confident they were of their responses to tradeoff questions. Linear programming-based methods used this information to eliminate some alternatives but generally did not provide a complete ranking of policies. We conclude that this approach provides a more realistic representation of the precision of people's preferences, and their implications for policy ranks.

A recurring theme in these results is the benefit of using multiple MCDM approaches. Low predictive validities and intermethod correlations indicated that no single method can be used to identify the best alternative. The outcomes of the various methods often conflicted because each method frames the problem differently. Further, the participants disagreed about which method is preferred for policy making—every method (except for stochastic dominance) had its advocate, and no one method was favored by all.

In their evaluation of methods, many of the participants advocated using several methods in concert. By asking the participants to resolve conflicts among multiple methods, they were forced to reflect on the problem further and to reconsider their judgments and the effects they have on policy choices. Participants recommended using revised weights (reconciliation of weights from different techniques) more than any single weighting method in actual climate change decision-making. Similarly, for deterministic ranking methods, participants recommended reconciliation of multiple methods (where the user reviews results from several methods and selects a final set of policy ranks) over any individual MCDM method. Finally, reconciliation of multiple methods and linear utility functions with revised weights were tied for the most highly recommended uncertainty ranking method.

Yet holistic assessment was, on average, more highly recommended for actual policy making than any MCDM method, including reconciliation, for the both the deterministic and uncertain cases. In other words, participants slightly preferred that decision-makers examine attribute values and decide policy rankings subjectively instead of applying a formal MCDM method. Several participants discussed the danger of using a "black box" approach, and others were uncomfortable with the process (e.g., "This framework doesn't tell me anything, because I don't view the world that way."). Participants stressed the need for adequate time to understand and incorporate all relevant alternatives and decision-maker's values, to complete MCDM exercises, and to explore and discuss results. Approaches that decompose the problem and evaluate attributes or pairs of attributes independently were especially problematic because the climate change experts had difficulty separating value independence from physical independence. In addition, participants had other problems with the MCDM exercises including missing attributes and lack of confidence in the values of the attributes that were included.

Perhaps the biggest obstacle to use of formal MCDM methods in IA is expert disagreement with respect to the performance of attributes for the alternatives. Experts dispute the distribution, nature, and magnitude of impacts and how much uncertainty exists. Some participants disliked our particular IA model and others argued that no model could meaningfully estimate impacts in the distant future, even if uncertainty is considered.

This makes the application of MCDM to climate policy extremely difficult, as all methods require estimates of what will happen under different alternatives. The workshop results suggest that including distributions of impacts is insufficient to address expert disagreement. MCDM often has been used to explore how differences in value judgments affect the choice of policy alternatives; however, for climate policy, MCDM could make a bigger contribution by identifying how policy choices are affected by disagreements over the attribute values for individual policies.

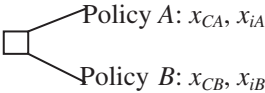
In spite of the shortcomings they perceived in MCDM methods, participants generally felt that the process of working through MCDM procedures and examining results helped them better understand how they think about the problem holistically. For example, one person said he wanted to do the “right” thing when he completed the holistic assessment, so he choose one of the more costly alternatives, whereas with other methods he placed higher weight on the cost attribute. He said that examining results from

these methods caused him to reanalyze how he thinks about the attributes. This further stresses that use of multiple methods can enhance understanding (Brown and Lindley, 1986; Corner and Buchanan, 1997; Simpson, 1996; Hobbs and Horn, 1997). In discussions, participants supported the idea that MCDM provides insights, saying: “structured analysis can help educate [the users about the decision]”, “one useful aspect [of MCDM is the] implications of doing things different ways,” and “[We can] understand the decision process better by looking at decisions with different perspectives. . . That’s good.”

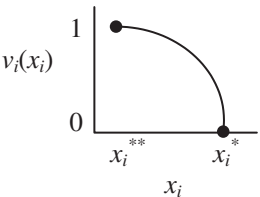
Beyond a method’s numerical output, insight gained from the process of working through methods is a primary benefit of MCDM. This sentiment was echoed in our workshop. Use of multiple methods was supported by analysis of method results, participant evaluations, and participant discussion. However, the workshop also highlighted several challenges that must be overcome for the successful application of MCDM to climate change policy.

APPENDIX A. SUMMARY OF MCDM METHODS COMPARED

All methods used four attributes (temperature increase, SO₂ emissions, nuclear waste generation, and control cost) except where indicated. Methods applied with six attributes also included ecosystem stress and sea-level rise. All methods are defined in Goicoechea *et al.* (1982) unless other citations are given.

Method	Description
Weight Selection Methods	w_i = weight for attribute i x_{ij} = value for attribute i for alternative j
Point allocation	Allocate 100 points among attributes. Performed twice: (a) 6 attributes, (b) 4 attributes
Hierarchical point allocation	6 attributes
Swing weighting/analytical hierarchy process (AHP) hybrid (Hobbs and Horn, 1997)	Used swing method to compare two attributes at a time, stating a ratio of importance. Repeated for all possible pairs of attributes. Inconsistencies resolved via AHP eigenvector method (Saaty, 1980).
Tradeoff weighting (Keeney and Raiffa, 1976)	Users chose one of the following question formats: (1) What value of x_{CA} would make you indifferent between policies A and B ? (x_{CA} and x_{CB} = values for control cost for alternatives A and B , respectively. x_{iA} and x_{iB} = values for attribute i for alternatives A and B , respectively. We supplied the user the values for x_{iA} , x_{CB} , and x_{iB} .)  (2) How much cost are you willing to incur for a $(x_{iA} - x_{iB})$ improvement in attribute i ?

APPENDIX A. (Continued)

Method	Description
	<p>Each question was asked three times to compare the cost attribute to the other three attributes. Weights were calculated as follows:</p> $\frac{w_i}{w_C} = \frac{[x_{cA} - x_{cB}]/[x_C^{**} - x_C^*]}{[x_{iB} - x_{iA}]/[x_i^{**} - x_i^*]}$ <p>x_i^{**} and x_i^* = best and worst values, respectively, for attribute i, among all alternatives. x_C^{**} and x_C^* = best and worst values, respectively, for cost among all alternatives.</p>
Revised weights (final revision of weights)	Participants were given weights resulting from each of the above methods and asked to provide a final set of weights.
Deterministic Ranking Methods	n = number of attributes. Revised weights used except where specified.
Initial holistic assessment	Alternatives ranked from most desirable (1) to least desirable (7), and then rated from most desirable (100) to least desirable (0), using the information provided in Table I. Performed twice: (a) 6 attributes, (b) 4 attributes
Additive linear value function	$\text{MAX}_j V(X_j) = \sum_{i=1}^n w_i v_i(x_{ij})$ <p>$V(X_j)$ = overall value of policy j $v_i(x_{ij})$ = single criterion value function that converts the criterion into a measure of value of worth, $v_i(x_i^{**}) = 1$, $v_i(x_i^*) = 0$, with:</p> $v_i(x_{ij}) = \frac{(x_{ij} - x_i^*)}{(x_i^{**} - x_i^*)}$ <p>Additive value function applied using results of each weighting method.</p>
Additive non-linear value function	<p>Two methods were used to generate $v_i(x_{ij})$, which may be non-linear, for use in additive value function:</p> <p>(1) Mid-value splitting: $x_{i0.5}$ = user-specified value for attribute i that is halfway in desirability between x_i^{**} and x_i^*, $v_i(x_{i0.5}) = 0.5$. Linear value function used if appropriate; otherwise, a_i, b_i, and c_i found such that $v_i(x_i) = a_i + b_i \exp(c_i \cdot x_i)$.</p> <p>(2) Users drew a value function representing $v_i(x_i)$. E.g.,</p> 
Goal programming ^a	<p>(a) $p = 2$, (b) $p = \infty$. g_i = user-specified maximum acceptable value for attribute i.</p> $\text{MIN}_j \sum_{i=1}^n w_i (\text{MAX}(0, v_i(g_i) - v_i(x_{ij})))^p$ <p>Thus, only undesirable deviations from goals are penalized.</p>
ELECTRE I ^a	<p>Alternative A is superior to B (A “outranks” B) if both of the following</p> <p>(1) Concordance: $C(A, B) > P$</p>

APPENDIX A. (Continued)

Method	Description
	<p>P = specified threshold (0.5 used in this experiment)</p> $C(A, B) = \sum_{i \in N} w_i / \sum_{i=1}^n w_i$ <p>N_i = set of attributes for which x_{iA} is better than x_{iB}. If there is a tie, then half of the weight is placed in the denominator. (2) Discordance: $D(A, B) < q_i \forall i$ $D_i(A, B) = v_i(x_{iB}) - v_i(x_{iA})$ q_i = user-specified threshold for tolerable dissent for attribute ELECTRE does not yield a complete ranking of alternatives. The set of alternatives that are not outranked defines a "kernel" of preferred options.</p>
Fuzzy Sets (Bellman and Zadeh, 1970) ^b	$\text{MAX}_j \text{ MIN}_i v_i(x_{ij})^{w_i}$ <p>w_i = weight for attribute i, rescaled so the highest weight for any attribute is 1. $v_i(x_{ij})$ is interpreted as a fuzzy set membership function describing the extent to which j is a "good" solution in terms of attribute i. The above aggregation procedure is one of many possible implementations of fuzzy sets and is often used in electrical engineering applications.</p>
Revision of Ranks and Ratings (final holistic assessment)	Participants were given results for the deterministic ranking methods (except fuzzy sets) and asked to provide a final set of ranks and ratings.
Uncertainty Ranking Methods^c	<p>x_{ijk} = value for attribute i for alternative j for simulation k K = number of simulations x_i^{**} and x_i^* = best and worst values, respectively, for attribute i among all alternatives and simulations</p>
Initial holistic assessment	Alternatives ranked from most desirable (1) to least desirable (7), and rated from most desirable (100) to least desirable (0), using the information provided regarding possible outcomes.
Linear utility function	<p>$U(X_j)$ = expected utility of alternative j $u_i(x_{ijk})$ = single criterion utility function, $u_i(x_i^{**}) = 1, u_i(x_i^*) = 0$</p> $U(X_j) = \sum_{k=1}^K \sum_{i=1}^n w_i u_i(x_{ijk}) / K$ $u_i(x_{ijk}) = \frac{(x_{ijk} - x_i^*)}{(x_i^{**} - x_i^*)}$
Non-linear utility function	<p>Gamble method: User specified a value ($x_{i0.5}$) for attribute i so that he/she is indifferent between a deterministic alternative ($x_{i0.5}$) and a gamble (50/50 chance of x_i^{**} or x_i^*). Repeated for all attributes. Linear value function used if appropriate. Otherwise, $a_i, b_i,$ and c_i found such that $u_i(x_i) = a_i + b_i \cdot \exp(c_i \cdot x_i)$</p>

APPENDIX A. (Continued)

Method	Description
Min Max Regret (Loomes and Sugden, 1982)	<p>R_{jk} = regret; loss in utility under scenario k if policy j is chosen as opposed to the best alternative under that scenario.</p> $R_{jk} = \text{MAX}_h U(x_{hk}) - U(x_{jk}), \text{ where } U(x_{jk}) = \sum_{i=1}^n w_i u_i(x_{ijk})$ <p>This method chooses the alternative j that minimizes the maximum regret among all scenarios k:</p> $\text{MIN}_j \text{MAX}_{k=1}^K R_{jk}$
Stochastic Dominance (Becker and Soloveitchik, 1998; Zeleny, 1982)	<p>Let $F_j(U)$ = cumulative probability distribution for $U(x_{jk})$, where $U(x_{jk})$ is estimated using the K values of $U(x_{jk})$ and $x_{jk} = \{x_{ijk}, \forall i\}$. Alternative A first-order stochastically dominates (fsd) alternative B if: (a) $F_A(U) \leq F_B(U), \forall U \in [0,1]$; and (b) $\exists U \in [0,1]$ such that the inequality is strict.</p> <p>First-order stochastic dominance makes no assumptions about risk preference; if A fsd B, then any $U(\cdot)$ that is a positive monotonic transformation of the original $U(\cdot)$ will result in A having a higher expected utility than B.</p> <p>Alternative A second-order stochastically dominates (ssd) alternative B if: (a) $\int_0^U F_A(v) dv \leq \int_0^U F_B(v) dv, \forall U \in [0,1]$; and (b) $\exists U \in [0,1]$ such that the inequality is strict. Second-order stochastic dominance assumes the decision-maker is risk-averse (i.e., if A ssd B, then any $U(\cdot)$ that is a positive concave transformation of the original $U(\cdot)$ will result in A being preferred).</p>
Revision of Ranks and Ratings (final holistic assessment)	Participants were given results for the above uncertainty ranking methods and asked to provide a final set of ranks and ratings.

^aThese are variations of traditional goal programming and ELECTRE I procedures, which subjects preferred in previous experiments (Hobbs *et al.*, 1992; Hobbs and Meier 2000).

^bParticipants did not review results for this method or evaluate its appropriateness or ease of use.

^cTo simply the assessment, uncertainty methods considered only four alternatives (base case; \$75, \$150, and \$300/ton CO₂ tax).

APPENDIX B. ORGANIZATIONS
REPRESENTED AT THE WORKSHOP

American University
Argonne National Laboratory
Battelle Pacific Northwest National Laboratory
Carnegie Mellon University
Case Western Reserve University
Charles River Associates
The H. John Heinz III Center
Johns Hopkins University
Lumina Decision Systems
Margaree Consultants, Inc.
The RAND Corporation
Resources For the Future
U.S. Environmental Protection Agency
U.S. General Accounting Office
U.S. Global Research Program

ACKNOWLEDGEMENTS

We would like to thank workshop participants for their indispensable contributions, patience, and good humor. We also thank Richard Anderson for valuable insights, and the editor and two reviewers for their helpful comments. This research was supported by the National Science Foundation (NSF SBR9634336).

REFERENCES

Adelman L. 1991. Experiments, quasi-experiments, and case studies: a review of empirical methods for evaluating decision support systems. *IEEE Transactions on Systems, Man, and Cybernetics* **21**: 293–301.

- Arrow KJ, Parikh J, Pillet G, Grubb M, Haites E, Hourcade JC, Parikh K, Yamin F. 1996a. Decision-making framework for addressing climate change. In *Climate Change 1995: Economic and Social Dimensions of Climate Change*, Intergovernmental Panel on Climate Change (IPCC), Bruce J, Lee H, Haites E (eds). Cambridge University Press: New York.
- Arrow KJ, Cline WR, Maler K-G, Munasinghe M, Squitieri R, Stiglitz JE. 1996b. Intertemporal equity, discounting, and economic efficiency. In *Climate Change 1995: Economic and Social Dimensions of Climate Change*, IPCC, Bruce J, Lee H, Haites E (eds). Cambridge University Press: New York.
- Becker N, Soloveitchik D. 1998. Dynamic multiple-objective optimization of environmental regulation policies. *Journal of Multi-Criteria Decision Analysis* 7: 13–19.
- Bellman RE, Zadeh LA. 1970. Decision-making in a fuzzy environment. *Management Science* 17: 141–164.
- Belton V. 1986. A comparison of the analytic hierarchy process and a simple multi-attribute value function. *European Journal of Operational Research* 26: 7–21.
- Bernabo JC, Eglinton PD. 1992. *Joint Climate Project to Address Decision Makers' Uncertainties*. Science and Policy Assoc., Inc.: Washington, DC.
- Borcherding K, Eppel T, von Winterfeldt D. 1991. Comparison of weighting judgments in multiattribute utility measurement. *Management Science* 37: 1603–1619.
- Brown RV, Lindley DV. 1986. Plural analysis: multiple approaches to quantitative research. *Theory and Decision* 20: 133–154.
- Corner JJ, Buchanan JT. 1997. Capturing decision maker preference: experimental comparison of decision analysis and MCDM techniques. *European Journal of Operational Research* 98: 85–97.
- Delbecq AL, Van de Ven AH, Gustafson DH. 1975. *Group Techniques for Program Planning—A Guide to Nominal Group and Delphi Processes*. Scott Foresman and Company: Glenview.
- Dowlatabadi H, Morgan MG. 1993. Integrated assessment of climate change. *Science* 259: 1813–1814.
- Edwards W. 1977. How to use multiattribute utility measurement for social decision-making. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-7: 326–340.
- Elmes DG, Kantowitz BH, Roediger III HC. 1995. *Research Methods in Psychology*, 5th Edition. West Publishing Co.: St. Paul, MN.
- Eppel T. 1992. *Description and Procedure Invariance in Multiattribute Utility Measurement*. Purdue University School of Management: West Lafayette, IN.
- Evans GE, Riha JR. 1989. Assessing DSS effectiveness using evaluation research methods. *Information and Management* 16: 197–206.
- Gardiner PC, Ford A. 1980. Which policy run is best, and who says so? *TIMS Studies in Management Studies* 14: 241–257.
- Goicoechea A, Hansen D, Duckstein L. 1982. *Multi-objective Decision Analysis with Engineering and Business Applications*. Wiley: New York.
- Gunderson DE, Davis DL, David DF. 1995. Can DSS technology improve group decision performance for end users? An experimental study. *Journal of End User Computing* 7: 3–10.
- Hadley CF, Schoner B, Wedley WC. 1997. A field experiment comparing anchored and unanchored criteria weights in the analytic hierarchy process. *Journal of Multi-Criteria Analysis* 6: 140–149.
- Hammit JK, Lempert RJ, Schlesinger ME. 1992. A sequential-decision strategy for abating climate change. *Nature* 357: 315–318.
- Hobbs BF. 1986. What can we learn from experiments in multiobjective decision analysis? *IEEE Transactions on Systems, Man, and Cybernetics* SMC-16: 384–394.
- Hobbs BF, Chankong V, Hamadeh W, Stakhiv EZ. 1992. Does choice of multicriteria method matter? An experiment in water resources planning. *Water Resources Research* 28: 1767–1780.
- Hobbs BF, Horn GTF. 1997. Building public confidence in energy planning: A multimethod MCDM approach to demand-side planning at BC gas. *Energy Policy* 25: 357–375.
- Hobbs BF, Meier PM. 2000. *Energy Decisions and the Environment: A Guide to the Use of Multi-Criteria Methods*. Kluwer Academic Publishers: Boston.
- Holmes KJ, Ellis JH. 1996. Potential environmental impacts of future halocarbon emissions. *Environmental Science and Technology* 30: 348–355.
- Holmes KJ, Ellis JH. 1997. Simulation of halocarbon production and emissions and effects on ozone depletion. *Environmental Management* 21: 669–685.
- Huber GP. 1974. Multiattribute utility models: A review of field and field-like studies. *Management Science* 10: 1393–1402.
- Intergovernmental Panel on Climate Change (IPCC), Working Group I. 1995. *Climate Change 1995: The Science of Climate Change*. Houghton JJ, Meiro Filho LG, Callander BA, Harris N, Kattenberg A, Maskell K (eds.). Cambridge University Press: New York.
- Jaeger CC, Renn O, Rosa EA, Webler T. 1998. Decision analysis and rational action. In *Human Choice and Climate Change*, Rayner S, Malone E (eds.). Battelle Press: Columbus, Ohio.
- John R, Edwards W. 1978. Importance weight assessment for additive riskless preference functions: A review. Research Report 78-5. Social Sciences Research Institute, University of Southern California: Los Angeles, CA.
- Kahneman K, Tversky A. 1973. On the psychology of prediction. *Psychological Review* 80: 237–251.
- Keeney R, Raiffa H. 1976. *Decisions with Multiple Objectives*. Wiley: Cambridge, Massachusetts.
- Korhonen P, Larichev O, Mechitov A, Moshkovich H, Wallenius J. 1997. Choice behavior in a

- computer-aided multiattribute decision task. *Journal of Multi-Criteria Decision Analysis*: 233–246.
- Lai S, Hopkins L. 1995. Can decisionmakers express multi-attribute preferences using AHP and MUT? An experiment. *Environmental Planning B* **22**: 21–34.
- Larichev OI. 1992. Cognitive validity in design of decision-aiding techniques. *Journal of Multi-Criteria Decision Analysis* **1**: 127–138.
- León OG. 1997. On the death of SMART and the birth of GRAPA. *Organizational Behavior and Human Decision Processes* **71**: 249–262.
- Leung P. 1978. Sensitivity analysis of effect of variations in form and parameters of a multiattribute model: A survey. *Behavioral Science* **23**: 478–485.
- Loomes G, Sugden R. 1982. Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal* **92**: 805–824.
- Manne AS, Richels RG. 1992. *Buying Greenhouse Insurance: The Economic Costs of Carbon Dioxide Emission Limits*. MIT Press: Cambridge, Massachusetts.
- Meo M. 1991. Policy-oriented climate impact assessment. *Global Environmental Change* **1**: 124–138.
- Morgan MG, Keith DW. 1995. Subjective judgments by climate experts. *Environmental Science & Technology* **29**: 468A–476A.
- National Acid Precipitation Assessment Program (NAPAP) 1991. *The Experience and Legacy of NAPAP, Report to the Joint Chairs Council of the Interagency Task Force on Acidic Deposition*. NAPAP Oversight Review Board: Washington, DC.
- Parson WA, Fisher-Vanden K. 1995. *Searching for Integrated Assessment: A Preliminary Investigation of Methods, Models, and Projects in the Integrated Assessment of Global Climate Change*. Consortium for International Earth Science Information Network: University Center, MI.
- Park KS, Kim SH, Yoon WC. 1996. An extended model for establishing dominance in multiattribute decision-making. *Journal of Operational Research* **47**: 1415–1420.
- Peck SC, Teisberg TJ. 1996. Uncertainty and the value of information with stochastic losses from global warming. *Risk Analysis* **16**: 227–235.
- Pöyhönen M, Hämäläinen RP. 2000. There is hope in attribute weighting. *INFOR* **38**: 272–282.
- Pöyhönen M, Hämäläinen RP. 2001. On the convergence of multiattribute weighting methods. *European Journal of Operational Research* **129**: 569–585.
- Raiffa H. 1982. *The Art and Science of Negotiation*. Belnap Press: Cambridge.
- Ridgley M. 1996. Fair sharing of greenhouse gas burdens. *Energy Policy* **24**: 517–529.
- Rotmans J, Dowlatabadi H. 1998. Integrated assessment modeling. In *Human Choice and Climate Change*, Rayner S, Malone E (eds.). Battelle Press: Columbus, OH.
- Saaty TL. 1980. *The Analytical Hierarchy Process*. McGraw-Hill: New York.
- Sarin RK. 1977. Screening of multiattribute alternatives. *The International Journal of Management Science* **5**: 481–489.
- Schoemaker PJH. 1981. Behavioral issues in multi-attribute utility modeling and decision analysis. In *Organizations: Multiple Agents with Multiple Criteria*, Morse JN (Ed.). Springer-Verlag: New York.
- Schubert R. 1994. Climate change and discount rates. In *Steps Towards a Decision Making Framework to Address Climate Change: Report from the Montreux IPCC WG III Writing Team II Meeting, March 3-6, 1994*, Pillet G, Gassmann F (eds). Paul Scherrer Institut: Switzerland.
- Shlyakhter A, Valverde Jr. ALJ, Wilson R. 1995. Integrated risk analysis of global climate change. *Chemosphere* **30**: 1585–1618.
- Simpson L. 1996. Do decision makers know what they prefer? MAVT and ELECTRE II. *Journal of the Operational Research Society* **47**: 919–929.
- Stewart TJ. 1992. A critical survey on the status of multiple criteria decision making theory and practice. *OMEGA* **20**: 569–586.
- Stewart TJ. 2000. Policy decisions in the public sector: can MCDA make a difference? In *Research and Practice in Multiple Criteria Decision Making*, Haimes YY, Steuer R (eds.). Springer: Berlin.
- Stillwell WG, von Winterfeldt D, John RS. 1987. Comparing hierarchical and nonhierarchical weighting methods for eliciting multiattribute value models. *Management Science* **33**: 442–450.
- Valverde Jr. ALJ, Jacoby HD, Kaufman GM. 1999. Sequential climate decisions under uncertainty: An integrated framework. *The Journal of Environmental Modeling and Assessment* **4**: 87–101.
- von Winterfeldt D, Edwards W. 1986. *Decision Analysis and Behavioral Research*. Cambridge University Press.
- Weber M. 1987. Decision making with incomplete information. *European Journal of Operational Research* **28**: 44–57.
- Zapatero EG, Smith CH, Weistroffer HR. 1997. Evaluating multiple-attribute decision support systems. *Journal of Multi-Criteria Decision Analysis* **6**: 201–214.
- Zeleny M. 1982. *Multiple Criteria Decision Making*. McGraw-Hill: New York.