

Does Decision Analysis Work?
A Research Agenda

by

Robert T. Clemen
Fuqua School of Business
Duke University
Durham, NC 27708-1208

(919) 660-8005
clemen@mail.duke.edu
<http://www.duke.edu/~clemen>

April 15, 1996

Draft: Comments Welcome

Does Decision Analysis Work?

A Research Agenda

ABSTRACT

How effective are decision-analysis techniques? Certainly decision analysis can show an individual how to be coherent in making inferences and choices. On the other hand, behavioral decision theory shows that people are not always coherent or internally consistent. And many other techniques have been developed to help individuals and groups understand their preferences, construct models, and make decisions. But are the techniques really effective in helping people to achieve their objectives? In this paper, we distinguish between two types of effectiveness. First is the notion that individuals must typically make choices before all uncertainties are resolved; such decisions must be made on the basis of preferences over those uncertain prospects, and the question is whether decision analysis leads to options that are broadly viewed as preferable. Second, after all uncertainties are resolved, the decision maker experiences the consequences, and the question here is whether decision analysis really leads to better outcomes: Are users of decision analysis happier, healthier, wealthier, or wiser in the long run? We use the terms weak and strong effectiveness, respectively, to refer to these two ways to think about the effectiveness of decision-making techniques.

Unfortunately, no body of literature exists that demonstrates unequivocally that decision analysis is either weakly or strongly effective. In this essay, we discuss how to determine the answers to these two questions. We review some existing related work, but much of the discussion is prospective, describing a number of possible research paradigms and experimental procedures. The essay concludes with suggestions for specific research to be done and some dilemmas that researchers will face.

Does Decision Analysis Work?

A Research Agenda

1. INTRODUCTION

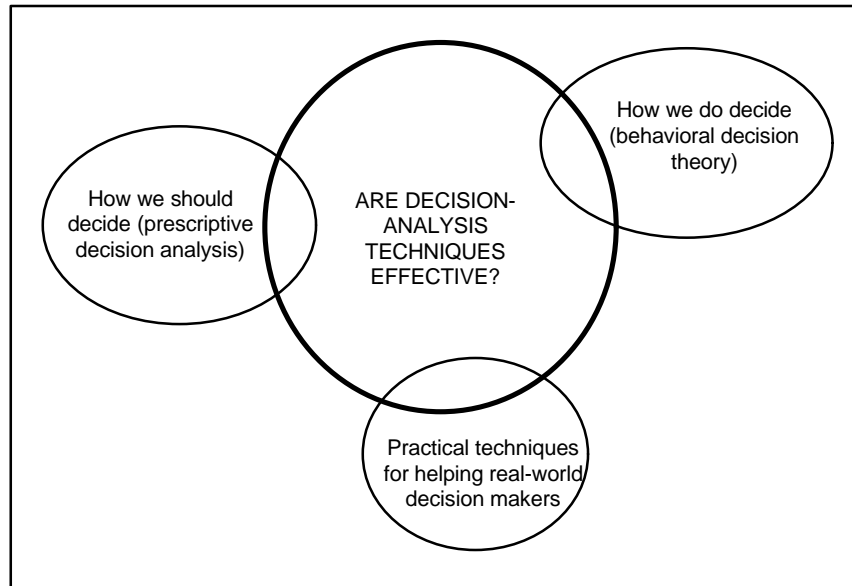
Decision analysts are quick to point out the distinction between decision process and decision outcome, and that even the best decision process can be derailed by an unlucky outcome (Clemen, 1996). So why should we use decision-analysis (DA) techniques? Typical answers include "gaining insight" and "being coherent," but the best reason to use a particular technique would be that doing so would more likely get us what we want. Are there any results indicating that this is the case? Do DA techniques really work? If an individual or a group uses DA, will they be more likely to get what they want? In the long run, are users of DA healthier, wealthier, safer, wiser, happier?

Certainly DA can show an individual how to be coherent in making inferences and choices. That is, adherence to decision-theory principles can promise that your decisions will not be self-contradictory and that the inferences you make will not expose you to the possibility of exploitation by a cunning bookmaker. Thus, DA tells us what one should do on the basis of logical argument. On the other hand, behavioral decision theory shows that people are not always coherent or internally consistent; we do make inconsistent patterns of choices, and our inferences could indeed be exploited. And beyond the axiom-based principles offered by DA, many other techniques have been developed, by consultants and others, to help individuals and groups understand their preferences, construct models, and make decisions.

Exhibit 1 shows the relationships among these three types of contributions to the decision-theory literature. Unfortunately, no body of literature truly unites the three by showing unequivocally whether the use of DA techniques (or other techniques for that matter) actually helps individuals to become happier, healthier, wealthier, or wiser. Despite over four decades of research done on developing DA methods, on understanding behavioral and psychological aspects of decision making, and on the application of DA in practice, no stream of research has been able to show conclusively what works and what does not. Such literature would not only evaluate techniques, but could also provide, on

the basis of results from behavioral decision theory, the reasons behind the failures and successes of the techniques.

EXHIBIT 1. The relationship between prescriptive DA, behavioral decision theory, and other practical techniques. The heavy circle represents research on the effectiveness of techniques; such research could connect the three existing areas into a truly unified body of literature.



Decision theorists often lament the fact that DA has not been widely adopted by corporations and public policy makers, a fact typically explained by how taxing DA procedures can be. But suppose DA could be shown definitively to be of value, and that its value easily overwhelms the typical costs of doing the modeling and analysis; having such “lore” as part of the story to tell about DA would make it a much easier sell!

Are DA techniques effective in getting people what they want? Answering this basic research question is far from straightforward. One cannot, for example, compare two different decisions taken by the same decision maker at the same time; one person can follow only one path in the decision tree, and it is impossible to compare what actually happens with what might have happened for that person. This suggests that any experimental study must use some interpersonal comparisons, but individuals with

different preferences may well prefer different paths for legitimate reasons. Finally, even though one makes a choice now, can future preferences change enough to render the foregoing decision inappropriate? To the extent that preferences change in unanticipated ways, any answers to our fundamental research question will be less meaningful.

Determining whether decision-making techniques are effective is a research problem fraught with challenges. As we work our way through the issues in this line of research, we discuss these challenges. In some cases, satisfactory approaches are easily identified, but in others it is clear that creative new methods must be found in order to address the research question adequately. In spite of this, there are ways to engage in comparative studies using available approaches. We will describe some ways that such studies might be carried out.

The next two sections begin by discussing in more detail just what it means to show that one technique is more effective than another and how this and related questions have been addressed in previous work. Section 4 describes in some detail three different types of studies that might be done to explore the effectiveness of various DA techniques and presents some specific research possibilities. Section 5 describes a general research agenda, and Section 6 concludes with a discussion of some dilemmas that will inevitably haunt researchers as they follow this path.

2. WHAT DOES EFFECTIVENESS MEAN?

2.1. *Strong and Weak Effectiveness.* The simple answer to the question, “What does effectiveness mean?” is that techniques are effective to the extent that they help us achieve what we want to achieve. Thus, we simply measure the quality of the consequences we get -- in terms of what we want to get -- as a function of the technique used. This perspective is clearly consequentialist. That is, it embodies the notion that the ultimate value of expending effort on decision making is because doing so can help one to obtain preferred consequences (Frisch & Clemen, 1994). In Kahneman’s (1994) and Kahneman & Snell’s (1990, 1992) terms, we intend to measure the hedonic experience utility that is obtained as a result of the action chosen and the outcome of any intervening chance events. The implication for studies of effectiveness is that we must focus on what

the decision maker experiences as the consequence of his or her choice and the uncertain outcomes that occur. In short, there is no substitute for a longitudinal study that follows a decision maker and records his or her experiences. The ultimate goal of such studies is to determine whether a decision technique is effective in this sense -- in helping a decision maker to obtain what he or she prefers over the long run -- and we will use the term *strongly effective* to describe such techniques.

Kahneman and Snell distinguish between hedonic experience utility and “decision utility.” The latter is captured by the preferences over the uncertain prospects faced at the time of the decision. Decision analysts call this “expected utility,” and the presumption is that the decision maker’s expected utility at the time of taking action corresponds to the hedonic utility experienced at the time the outcome is reached. A broader view, however, suggests that many phenomena could intervene and lead to poor performance of a decision-making technique, primarily by subverting the decision maker’s attempt to properly represent his or her anticipated experience. Problems could arise in understanding and assessing uncertainty, in identifying dimensions of preference, assessing preference functions (tradeoffs and individual utility or value functions), or in identifying key uncertainties or available alternatives. It is precisely the issue of comparing decision and experience utility that Kahneman and Snell address.

For our purposes, we can use the distinction between decision and experience utility to define a weaker form of effectiveness. If a decision technique leads to choices that are seen as preferable, then we will say that the technique is *weakly effective*. Although this may appear to be trivial -- of course the decision maker prefers what he or she chooses! -- it is not when viewed more broadly. A large part of the DA exercise is the identification of suitable alternatives, but alternatives may be preferred by different people for different reasons. Thus, showing weak effectiveness may be accomplished by showing that DA-generated alternatives are preferred -- or better yet dominant -- to other alternatives as judged by an appropriate population (or sample) of decision makers.

Does strong effectiveness imply weak effectiveness? It is certainly tempting to answer in the affirmative; if a technique produces alternatives that in turn lead to preferred consequences, would it not be the case that the decision makers would have evaluated

those alternatives as having greater expected utility? Unfortunately, no compelling reason exists to believe this would be the case. In fact, one can make the case that a technique might mislead decision makers – that they would perceive the recommended alternatives as dominated although the consequences eventually obtained from these alternatives would clearly be preferred. The issue here is not only the relationship between decision utility and experience utility, but also the extent to which a decision technique can itself lead to a discrepancy between the two.

2.2. *Elements of Value.* Saying that a technique is effective when it helps one to achieve greater utility upon experiencing an outcome begs the question of what contributes to one’s utility. What we want to achieve -- what has value and thus contributes to greater utility -- can typically be expressed as objectives (Keeney, 1992). Although we can legitimately expect different decision makers to have their own objectives in specific contexts, some basic classes of objectives may be common to certain types of decision-making units. Exhibit 2 lists some generic objectives that may be of interest to individuals, small groups, corporations, public-policy organizations, and scientists. In what follows, we will use the term “decision maker” to refer to the decision-making unit, regardless of whether that unit consists of one person or a group of individuals.

The objectives in Exhibit 2 are intended to be representative, not exhaustive. These objectives describe typical reasons why the decision maker cares about any decisions within its purview. For the individual, we might characterize the objectives as “why we live.” In contrast, a small interest group’s objectives can be said to represent “why we join” voluntarily with others in common endeavors. The objectives of policy makers include notions of fairness, efficiency, and the management of externalities, and we might call these objectives “why we govern.” Likewise, the corporation’s objectives could be described as “why we engage in economic activity,” and the scientist’s “why we learn.”

EXHIBIT 2. Typical Objectives of Different Decision Makers

- **Individual:** Happiness, health, wealth, safety, wisdom, love, respect, prestige
 - **Small Interest Group:** Impact on community, influence, social standing, camaraderie, goals specific to group's mission
 - **Public Policy Maker:** Efficient use or allocation of resources, productivity, environmental quality, safety, health, fair decision processes and outcomes.
 - **Corporation:** Profit, market share, stock price (wealth), sales, lower costs, worker satisfaction.
 - **Scientist:** Understanding, insight, predictability, accuracy
-

Measuring a decision technique's effectiveness will require measuring the extent to which objectives are achieved. Thus, Exhibit 2 can provide guidance on the sorts of things that should be measured in order to determine effectiveness. Although specific decision contexts may have specific objectives (e.g., maximize starting salary in the context of an individual's job search), that objective is likely related to one of the objectives in Exhibit 2 (wealth). Knowing what to measure to determine effectiveness is crucial; we want to be sure that we are concerned with the extent to which the decision maker's lot is improved, according to his or her perspective, by using one specific decision technique or another.

Measuring achievement of objectives may be straightforward or may prove to be quite difficult. For example, measuring health may be possible using standard epidemiological survey methods. Likewise, other than prying into a relatively sensitive personal issue, wealth presents no problem. How does one measure happiness, though? Or prestige? For a small group, how does one measure its impact? In policy making, fairness depends on perceptions of the distribution of outcomes as well as the process which led to the allocation. For example, some authors in the area known as "procedural justice" use the concept of "voice" to describe the extent to which stakeholders have control over the decision-making process (e.g., Folger, 1977). Thus, it may be important to measure both outcome and process fairness. Boiney (1995) and Fishburn and Sarin (1994, in press) provide decision-theoretic procedures for evaluating fairness of allocations (including risky allocations) using the concept of envy among stakeholders.

Finding adequate measures of crucial objectives for a specific decision maker thus represents the first challenge of research on effectiveness of decision-making techniques. Meeting this challenge, however, can be done by adopting a technique from DA! When a fundamental objective is not easily measured, Keeney (1992) describes how to specify proxies and how to construct scales. Either approach may prove useful for research purposes, provided the researcher is careful to ensure that the devised scales faithfully represent the decision maker's objectives.

The scientist's objectives focus primarily on the acquisition of knowledge, understanding, and often the ability to predict accurately based on acquired knowledge. Measuring achievement of these objectives is different, and in some ways easier, than doing so for other objectives. For example, one can easily compare forecasts in terms of their accuracy. On the other hand, it is not clear how one measures achievement of understanding or insight. Certain proxy scales can be imagined, such as publications, citations garnered, or even predictability. Although it may be possible to construct a scale to represent level of insight achieved, we leave the specifics of such a scale to the speculation of the reader and the creativity of a future researcher.

3. PREVIOUS STUDIES

It is unfair to claim that no prior work on the effectiveness of various decision-making techniques has been done. Much of behavioral decision theory is broadly relevant to our discussion. In this section we discuss some aspects of the literature that are pertinent to the problem of establishing the effectiveness of DA techniques. This review is not exhaustive but instead selects elements of the existing literature to demonstrate the nature of the connections between the existing literature and the proposed research.

3.1. *Probability Assessment.* In principle, this is one of the easier aspects of DA for which to determine the effectiveness of different techniques. One can imagine studies that use different techniques to elicit judgments and then compare those judgments to actual outcomes. If a specific decision context is used, then the judgments obtained can be used as inputs to the decision, and the average value obtained using different techniques

can be compared. If no decision or value function is specified, it is natural to adopt a strictly proper scoring rule (Savage, 1971).

Forecasting in general is closely related to the problem of probability assessment. Since the mid-1970s, beginning especially with work by Newbold and Granger (1974), considerable work has been done to evaluate different forecasting techniques. Evaluation has typically been in terms of some sort of average error metric, such as mean squared error, calculated for forecasts outside of the fitting data. Performance results have varied somewhat, but a common theme has been that simpler techniques have worked well compared to more complex techniques (Armstrong, 1984). Among the many explanations for this phenomenon, the one I find most compelling is that the underlying processes are not stationary; complex models that require substantial data to estimate a large number of parameters are unable to calculate good estimates of those parameters because of the nonstationarity of the data.

DA relies on the use of probability as a paradigm for modeling the uncertainty surrounding future outcomes. Errors from forecast estimation procedures can be used to construct probabilistic forecasts, although in classical forecasting, this is typically done rather simplistically by considering only the standard error associated with the forecast and applying (implicitly or explicitly) a normal distribution; we might say, for example, that there is only about a 5% chance that the actual value will be more than two standard errors away from the forecast. A more complete approach involves construction of the entire distribution, either using historical data or subjective assessment.

Considerable work has been done on the assessment of subjective probabilities, although much of it has focused on the internal consistency of human assessment. For example, articles in the volume by Kahneman, Slovic, and Tversky (1982) emphasize how heuristic judgment processes lead to cognitive biases. For the most part, this work follows a paradigm of showing ways that human judgment of subjective probability is inconsistent with the probability laws and definitions. One particular arm of the probability-judgment literature has focused on the calibration of judgments (Lichtenstein, Fischhoff, & Phillips, 1982). This work is more relevant to the research agenda we describe here simply because it considers the extent to which a record of probability judgments reflects the pattern of

actual outcomes; those who are more calibrated provide judgments closer to the actual long-run frequency of outcomes.

Wallsten & Budescu (1983) provide an in-depth review of literature on probability assessment. The discussion is framed around the concepts of reliability and validity. Reliability relates to the ability of an individual to replicate his or her assessments over time for the same events. Validity is described in terms of internal consistency of assessments (or consistency with probability theory), calibration, and “external” validity, or the extent to which the assessed distributions agree with distributions constructed on an “objective” basis. In most cases, researchers report conservatism of individuals in modifying judgments in light of new information and, for simpler situations in which the individual has only to assess a single, static distribution, substantial differences between the “subjective” and “objective” distributions.

Is there any evidence that specific procedures can improve probability judgments? Stael Von Holstein (1971a, b) and Schaefer & Borcharding (1973) report that short and simple training procedures can improve the accuracy (calibration) of assessed probability, although the results are far from an overwhelming improvement in performance. Fischhoff (1982) discusses debiasing techniques intended to improve the quality of subjective probability assessments.

Two specific areas have yielded results intriguing for those interested in improving individual probability assessments. The first is from the work by Gigerenzer & Hoffrage (1995), who emphasize that framing judgments in frequency terms (as opposed to the more traditional subjective “degree of belief”) can reduce assessment bias in a variety of situations. The second is the literature on the value of decomposition, the process of breaking down an assessment into smaller and presumably more manageable tasks, making these simpler judgments, and then recombining them using standard probability procedures to obtain the overall probability desired (e.g., Clemen, 1996, Ch. 8). For example, Hora, Dodd, & Hora (1993) indicate that decomposition can improve assessment performance, and Clemen, Jones, & Winkler (1996) show similar results in the context of aggregating expert judgments. Morgan and Henrion (1990) review the empirical support for decomposition in probability judgment.

Many judgments in large-scale analyses are generated by groups of experts, and so research on the accuracy of group judgments is also relevant for DA. Unfortunately, experimental conclusions on group probability judgment have been mixed. For example, a few studies on group-level judgment suggest caution. Myers & Lamm (1975) report evidence that face-to-face interaction in groups working on probability judgments may lead to social pressures that are unrelated to group members' knowledge and abilities. Gustafson, et al. (1973), Fischer (1975), Gough (1975), and Seaver (1978) all found in their experiments that interaction of any kind among experts led to increased overconfidence and hence worse calibration of group probability judgments. More recently, Argote, Seabright, & Dyer (1986) found that the representativeness heuristic is used more by groups than by individuals, presumably leading to more biases (e.g., overconfidence) related to this heuristic. In a related study, Tindale, Sheffey, & Filkins (1990) concluded that groups committed the conjunction fallacy more often than individuals.

3.2. *Preference Assessment.* While probability assessment can be evaluated fairly readily, the study of preferences and preference assessment techniques is more problematic. In this subsection, we sample a few of the studies that have been done that are particularly relevant to the problem at hand. A much more complete description of the literature can be found in von Winterfeldt & Edwards (1986, Chapter 10).

The most popular approach to studying preferences has been to consider the extent to which expressed preferences are internally consistent, as exemplified by the Allais paradox (Allais, 1953; Allais & Hagen, 1979) or by Tversky and Kahneman's (1981) work on framing. An example of such work that is especially appropriate for our discussion here is the work on preference reversals, a phenomenon first described by Lichtenstein and Slovic (1971) in which stated preferences reversed depending on response mode. For example, when presented with a pair of bets, A and B, where A has a higher chance of winning a smaller dollar prize, individuals tend to choose Bet A but at the same time assign a greater selling price to Bet B. The result is robust, having been demonstrated in many different domains and different forms, demonstrating that the elicitation of

preferences using presumably equivalent DA approaches does not guarantee consistent responses.

Ordonez et al. (1995) briefly review the preference-reversal literature and also study whether preference reversals can be reduced by “debiasing” along the lines of Fischhoff (1982). For example, Reilly (1982) showed that teaching subjects how to calculate the expected value of a bet can significantly reduce preference-reversal rates. One approach that Ordonez et al. took was to have subjects perform the judgment tasks simultaneously. Unfortunately, this approach yielded little improvement. Providing financial incentives for consistency, however, did reduce the reversal rate. Moreover, their results are consistent with the notion that the simultaneous judgment tasks, in the presence of adequate financial incentives, leads to a merging of the preference patterns displayed in the different tasks.

DA, on the basis of the axioms of preference, prescribes a number of approaches that are formally equivalent for assessing utility functions. Farquhar (1984) surveys many of the available utility-assessment methods. Hershey, Kunreuther, & Schoemaker (1982) discuss biases induced by different preference-elicitation approaches in spite of their formal equivalence. One such bias appears to be due to the certainty effect (Kahneman & Tversky, 1981), which led McCord & de Neufville (1987) to propose a lottery-equivalent assessment method.

For another example, consider multiattribute preference modeling, in which an individual typically assesses an individual value function for each attribute and weights or scaling constants that are used to combine the individual functions. Fischer (1975) reviews early studies on the validation of multiattribute utility assessment. The typical approach has involved what is called “convergent validity,” which is measured in this case by calculating the correlation between the intuitive rankings of the subjects and the rankings produced by the utility function. Many of these and related studies are described in von Winterfeldt and Edwards (1986).

Although most preference studies have been aimed at understanding and reducing internal inconsistencies, Kimbrough and Weber (1994) describe an experiment with a slightly different orientation. They compared a variety of utility-elicitation approaches,

each one implemented via a computer program. Some approaches confronted subjects with their inconsistencies and forced them to make modifications; these methods produced recommendations -- and by implication utility functions -- that were more acceptable to the users.

3.3. *Decision-Analysis Modeling.* So far we have spoken about uncertainty- and preference-assessment procedures. What about studies that investigate the effectiveness of overall decision-making strategies? For example, how does DA stack up against intuitive decision making or the Analytical Hierarchy Process (AHP) (Saaty, 1980)? Such studies seem to be quite rare, perhaps because behavioral theories are often tested in more microscopic contexts. Nevertheless, the question is a valid one; given what we know about human behavior and prescriptive approaches, it should be possible to design procedures that take into account human abilities and behavioral biases, and it would be appropriate to test whether such procedures do indeed produce better results.

Although such studies are rare, some effort was directed at understanding the effectiveness of MAUD, one of the early multiattribute decision aids. Humphreys and McFadden (1980) found that the program's value lay in the fact that it taught users about decision making rather than in helping to solve a specific problem at hand. John et al. (1983) and Aldag & Power (1986) used panels of experts to judge the quality of decisions made with the decision aid. Although the computer program was viewed favorably by the users, the experts judged that the quality of the decisions reached was the same regardless of the method used.

4. SOME RESEARCH PARADIGMS

Our discussion so far has shown that most of the work on the effectiveness of DA techniques has focused primarily on the psychology of judgment. Because DA is based on a system of axioms, it has been reasonable to study whether normal people behave according to those axioms and, if not, why not? Thus, we have found, for example, that individuals are prone to a variety of logical errors in assessing probabilities, or that the choices they make indicate that their preferences are inconsistent in some way.

Another aspect of much of the work done has been to ask whether DA-based judgments (of uncertainty or preference) reflect intuitive or holistic judgments. Certainly this sort of paradigm makes sense if one wants to know whether people naturally follow the logic on which DA rests. We now know that they do not, and the question now is not whether people naturally follow the axioms, but whether learning to do so can lead them to preferred choices and consequences.

Finally, previous work has largely been motivated by expected utility and hence has focused on aspects of decision making that fall within that model. However, as argued by Frisch and Clemen (1994), many elements of decision making fall outside of the expected utility paradigm *per se*. For example, expected utility sheds little light on how to identify one's objectives or how to find new alternatives. The model's mathematics do shed light on some fundamental structuring issues (value assessment can be separated from uncertainty assessment) and on different assessment procedures that should be equivalent (e.g., pricing out, swing weights, and lottery weights for multiattribute assessment). Not surprisingly, aspects of decision making for which expected utility has a specific implication have been more fully explored.

Unfortunately, virtually no work has been done that demonstrates clearly whether DA is strongly effective -- whether in the long run a user of DA will have better consequences. Relatively little work has been done to show the weaker form of effectiveness, that DA techniques recommend options that are broadly viewed as preferable. In this section, we do two things. First, we describe some research paradigms that must be developed in order to address these questions. Second, we give examples of specific studies that could address the effectiveness of particular techniques.

4.1. *Experimental Paradigms*

4.1.1. Strong Effectiveness: Longitudinal and Simulation Studies. Studies of strong effectiveness must ultimately embrace the challenge of longitudinal studies. If consequences are experienced over time, as they do for many important individual, corporate, or public-policy decisions, then it will be important to follow individuals over that time. As described above, it will be important to identify objectives that are important to the individuals and to track achievement of those objectives.

Aside from the complicated logistics of tracking a group of mobile individuals over long time spans (and of maintaining long-term funding for doing so), an important issue is identifying an appropriate population. Two specific groups come to mind. First are college graduates making decisions about careers. At a large school with a strong alumni program, it would be a relatively straightforward matter to keep track of individuals who had gone through a particular manipulation (training, use of a computer aid, etc.) as they made career choices. A second group would be upcoming retirees for a large corporation; as employees approach retirement, it would be possible to recruit some as participants in a study that manipulates decision techniques for retirement planning.

The logistic difficulties of real-time longitudinal studies reduce their attractiveness. Simulations may provide a suitable alternative. Industrial simulations, for example, are common fare in business curricula; similar games that would be amenable to manipulations in decision-making techniques could provide a test bed for the effectiveness of those techniques. Such games would have at least two advantages; the time dimension is highly compressed, and the environment (including in part the objectives of the participants) can be tightly controlled. Games could be designed around individual decisions, corporate strategy, or public policy; the main necessary ingredients are realistic decision situations and outcomes, along with appropriate incentives to engage the participants in the exercise. An example might be a game that requires participants to make a series of marketing-strategy decisions for their firm in a large industry. An experimental group would use a particular technique (assessment or modeling method or computer-based decision aid), and their results, measured relative to the objectives specified in the game, could be compared with the results of other groups.

4.1.2. Weak effectiveness: Panel Preferences. In contrast to strong effectiveness, studies of weak effectiveness need not be designed to track outcomes and consequences over time. Instead, the question is whether the alternatives generated by various techniques are viewed as preferable. In general terms, we are concerned here with the notion of decision utility, and it would be natural to have a panel of judges -- decision makers sampled from the same population as those who generated the alternatives in the first place -- express their preferences for those alternatives. These preferences could be

based on holistic judgments, full-fledged preference models, or something in between. Holistic judgments would appear to be unsatisfactory; after all, the entire premise of the research challenges the view that holistic judgments adequately represent an individual's preferences. On the other hand, forcing an individual into a specific preference model requires selection of a particular structure and possibly a particular modeling or assessment technique. Thus, it would appear that some in-between approach is needed, one that requires the judge to make relatively easy assessments regarding the alternatives' attributes.

For example, in comparing multiattribute alternatives, we can ask each member of a panel of judges to rate each alternative on a set of relevant attributes. With data of this nature, the researcher can explore all of the dimensions of preference. The strongest result would be to show that a particular decision-making technique tends to generate a high proportion of dominant or efficient alternatives, regardless of the importance (weights) ascribed to the different attributes. In what follows we will use the term *efficiency*, and construct a measure of it, to indicate how close an alternative is to the efficient frontier.

Exhibit 3 demonstrates an analytical approach aimed at measuring the relative performance of the alternatives in terms of efficiency and without imposing constraints on the specific weights on attributes in an additive utility function. A panel member has identified two attributes, X and Y , that are important in evaluating the alternatives and has in fact rated alternatives A, B, and C in terms of X and Y using functions $U(x)$ and $V(y)$. From the graph it is clear that A and B define an efficient set among the three alternatives and that C is dominated by A. Because A and B are both on the efficient frontier with respect to this particular set of alternatives, it makes sense that their efficiency measures E_A and E_B would be 100%. We desire a measure that yields a value of less than 100% for E_C .

Assuming an additive value function, C lies on an indifference curve defined by $aU(x) + bV(y) = t$. Using the same weights a and b , the greatest utility achievable is t^* , represented by the line segment AB. Thus, we can define E_C to be the ratio t/t^* . This is equivalent to calculating the ratio of the distance DC to the distance DE in Exhibit 3.

formula generalizes easily to more than two attributes and to more general forms of multiattribute utility functions.

Why would such an approach be useful? Suppose we have two sets of alternatives, each set generated by decision makers using a particular technique. A panel of judges can score the alternatives on a set of attributes (of their own choosing or attributes predetermined by the experimenter), and then for each judge the data can be used to generate an efficiency measure for each alternative. With efficiency measures calculated for all of the judges, statistical analysis can be used to compare alternatives generated by different decision techniques. The judgmental inputs satisfy our desire for something between holistic judgments of preference and a full-blown preference model; the ratings for each attribute are straightforward judgments that still capture the richness of the judges' preferences. No assessment of utility weights by the judges is needed, and the analysis can indicate whether one technique tends to generate more efficient alternatives than another technique.

The calculation of efficiency measures as described above can also be used to evaluate negotiated agreements, thereby providing a means for comparing agreements created using different negotiation or group-decision techniques. Rather than having a single judge score each alternative on several attributes, we would have representatives of each stakeholder group score each alternative from their own perspective. The score could arise from a holistic judgmental rating or from a carefully constructed preference model, but regardless, it would represent the overall value of the alternative to the stakeholder. To generate efficiency scores, choose one representative from each stakeholder group and use their scores on the alternatives to plot the alternatives. For example, in Exhibit 3, each axis would represent the utility of a particular stakeholder group in the negotiation, and the three alternatives are plotted according to the utilities of the two parties. Alternatives A and B fall on the Pareto frontier, and the efficiency score E_C would be a measure of how close C is to the Pareto frontier. Using an approach like this, with appropriate sampling of judges' preferences for each stakeholder group, one can determine whether one particular negotiation technique tends to generate more Pareto-superior agreements than another technique.

4.2. *Some Research Projects.* The paragraphs above describe some general experimental paradigms that could be used to determine the effectiveness of DA (and other) techniques. In this subsection we speculate on some different specific studies that might be done.

4.2.1. Probabilistic Forecasting Competition. Imagine a complex forecasting situation that involves many kinds and levels of uncertainty, such as forecasting crude oil prices, diagnosing a disease based on patient signs and symptoms, or troubleshooting a computer software installation. A variety of analytical and modeling techniques are available for problems such as these, ranging from the construction of complex belief nets or other artificial-intelligence models to decomposed or even holistic probability judgments made by experts.

Given one or more prespecified domains, a competition could be held, pitting different probabilistic forecasting techniques against each other. It would be most natural to hold such a competition in real time, in which case the result can be used to determine strong effectiveness of the techniques. As indicated above, strictly proper scoring rules provide natural performance measures that can capture both calibration and skill (Murphy and Winkler, 1970). In a probabilistic forecasting environment, enough outcomes would have to be recorded to calculate meaningful average scores for the various techniques. Competitions of this nature have been run in the forecasting field (Makridakis *et al.*, 1982, 1993) and in meteorology (Abramson *et al.*, 1994).

An alternative to a real-time exercise would be to construct a simulation in which participants would have to make probabilistic forecasts and in which their overall performance would be measured by their average scores. For example, a business simulation in which participants must make judgments and take calculated risks could be designed to incorporate participants' skill in assessing probabilities related to aspects of their business such as marketing, R&D, production, or competitive analysis. Aside from creating such a game, the challenge would be to implement different probability-forecasting options within the context of the game in a way that permits experimental manipulation of the those techniques.

Scoring rules need not be the basis of comparing probability-forecasting techniques, especially if the probability forecast can be related to a specific decision context. For example, in the business-simulation game, one might want to measure stock price, profits, market share, or some other objective important to a real or fictitious corporation. Other possibilities are to choose a context like college choice or retirement planning and have proponents of different techniques develop systems that lead users through the necessary uncertainty judgments and modeling before offering alternatives from which to choose. By tracking the experiences of the participants in a longitudinal study, one could measure the effectiveness of the different systems and implicitly of the underlying techniques.

4.2.2. Value Structuring and Creativity. In his book *Value-Focused Thinking* (VFT), Keeney (1992) stresses that his approach is valuable for many things, not least of which is its potential for generating creative alternatives. Creativity in decision making relates to the quality of alternatives generated; a truly creative alternative is one that is both novel and effective in achieving the decision maker's objectives (Clemen, 1996).

How would one determine whether a technique generates creative alternatives? One must start with a relatively unstructured problem that admits the possibility of creative problem solving; highly structured textbook-like problems typically do not provide adequate leeway for the decision maker to find creative answers. But if we want to be able to evaluate creative alternatives, by definition we may be considering alternatives that we have not yet seen, so no scoring system for the alternatives can be established *ex ante*.

Fortunately, a procedure like the one described in the previous subsection for measuring weak effectiveness can be used. Suppose that VFT and other techniques are used to generate alternatives in some decision situation, either real or simulated. A panel of judges can evaluate the alternatives by rating them on each of several dimensions, and the subsequent efficiency analysis can determine whether VFT tends to generate more efficient alternatives than do the other techniques. To the extent that more creative alternatives are also efficient, such a study can indicate the potential of DA techniques to enhance creativity.

4.2.3. Decision Making Tournament. Several different decision paradigms exist, among them DA, possibly represented by experts with different leanings (e.g., Ralph Keeney vs. Ron Howard), the AHP, and goal programming to name a few. A tournament could be held by having proponents of different methods address a prespecified set of decision problems. The quality of the decisions chosen could be evaluated by a panel of judges or, if suitable, by tracking the downstream consequences to the decision makers either in a real-world or simulated environment.

An important issue that must be faced in such a tournament is coming up with decision problems that present a reasonably level playing field for the various techniques to be tested. Such problems would presumably consist of “case studies” that are rich in realistic and detailed information. Care must be taken not to present information in a way that artificially predisposes a decision maker toward a particular technique; for example, expressing uncertainty explicitly in terms of subjectively assessed probability distributions might create a bias toward DA, whereas explicit indication of pairwise comparisons might predispose a decision maker toward the AHP.

4.2.4. Canonical Decision Problems. The tournament just described with technique proponents as participants could be run only once for each set of cases. In any subsequent run with the same set of decision problems, one would have to ensure that the later participants were not prejudiced one way or the other by the outcomes and decisions of the prior study. An interesting twist on the tournament idea would be to develop a set of canonical decision problems and an experimental procedure that could be used as a way to test new techniques as they are developed. A parallel can be found in the field of mathematical programming, which has adopted a few computational problems that are commonly used as benchmarks for comparing algorithm performance.

Although no similar set of canonical problems exists for decision making, the creation of such a collection would facilitate the comparison of decision-making techniques. A strict procedure must be established, however, in which “naive” subjects would be instructed in the use of a particular technique prior to applying it to the canonical problem. The prior knowledge of the subjects, regarding both the problems they would face and previous results on other decision techniques, must be carefully controlled.

Finally, because decisions would not necessarily be made contemporaneously, the panel approach to judging the quality of alternatives is not appropriate. Instead, the decision problems must have appropriate built-in measures for determining the quality of the decisions, and those built-in measures must be directly related to the objectives of the roles adopted by the participants, whether those roles are in personal, corporate, public-policy contexts.

4.2.5. Ethnological Study of Decision Making. A final example of a possible study is to pursue an ethnological approach. In this case, one would collect accounts of decision-making styles and techniques in different contexts. For example, a database might be developed that contained accounts of individual decision making, including decision context, framing, techniques used, and choices made. Similar databases could be created for corporate or public-policy decisions, although gag rules may render collection of such data quite difficult. To be able to compare effectiveness, such a database must be augmented with either judgments of quality of the alternative chosen (possibly done by a panel reviewing all decisions in the database) or a measure of the consequences to the decision maker (via later reports of performance). A large database could be analyzed to determine the characteristics of the most effective decision makers.

5. A RESEARCH AGENDA

The suggestions above regarding types of decision-effectiveness studies are brought together in this section with a brief discussion of an overall research agenda. It is clear that a program to study the effectiveness of DA techniques has many facets and issues to which the researcher must attend.

The first order of business is to create appropriate experimental and analytical paradigms. For example, the analysis of panel judgments as described above in Section 4.1.2 must be more fully developed and refined. We must learn how to handle the logistics of longitudinal studies; other social-science disciplines with experience in longitudinal studies may be of assistance here. Good case studies must be developed and refined for use in simulation studies or as canonical decision problems for ongoing research on decision effectiveness. Methodological work of this nature is not terribly attractive in and

of itself because it does not directly address substantive questions of interest. Nevertheless, the importance of such work cannot be overstated for the research program described here; as discussed above, studying decision effectiveness poses important methodological problems that must be solved if definitive results are to be obtained.

With an array of experimental methods available, the research can take several approaches:

- Compare specific techniques such as DA, AHP, naive decision making, and so on as overall paradigms for making decisions.
- Study different aspects of decision making: problem structuring, uncertainty modeling and assessment, preference assessment, etc.
- Study different types of decision-making contexts, such as personal decisions, corporate strategy, public policy, or multiple-stakeholder decisions.
- Investigate decisions in specific domains such as environmental risk assessment, college or career choice, consumer product marketing strategy, research and development, or municipal waste facility location.

These four bullets define four different dimensions along which research on decision effectiveness can proceed, and presumably the initial experimental and analytical paradigms will be refined and supplemented as experience with such studies increases.

Finally, it is important to acknowledge the interdisciplinary nature of decision making and the many different areas in which decision-making research can take place. Based on textbook adoptions, for example, courses in DA are taught in professional schools such as business, engineering, law, public policy, and medicine. Within the traditional disciplines, economics, political science, sociology, psychology, mathematics, and statistics have all taught DA. Research on decision effectiveness will be most useful in practice if effectiveness can be demonstrated in specific applications and contexts rather than in abstract, general terms. In other words, we should be taking DA and research on the effectiveness of DA techniques to areas in which it can contribute directly -- and the results of the research should be published in the corresponding journals -- rather than

doing general studies, publishing in traditional DA outlets, and hoping that other areas will find our results.

6. CONCLUSION: DILEMMAS FOR RESEARCHERS

The foregoing text has laid out the principle issues that must drive a research program aimed at determining whether DA techniques are effective. Such research will be challenging, but the technical challenges appear not to be insurmountable. As with many research programs, though, this one faces some specific and unique problems. In this section we turn to the dilemmas that researchers in this area must face.

6.1. *Incorporating Behavioral Decision Theory.* As described, research to determine effectiveness of techniques appears theoretically sterile. A cynical description of the proposed work might be, “Identify a couple of techniques, stick them in an experiment, and measure the performance of the techniques.” This makes the work sound similar to data dredging; keep looking until you find something that works with no attention paid to theoretical issues. Clearly, this is not the intention of the work! We hope that decision techniques used and tested will be sensitive to results from behavioral decision theory. Moreover, behavioral results may provide specific hypotheses to guide decision-effectiveness research, and results from research may in turn provide insight regarding human decision-making phenomena, especially in conjunction with specific decision tasks.

On the other hand, it is not clear that the experimental work required to test the effectiveness of entire DA systems and techniques will be helpful for understanding details of human behavior. Behavioral decision research has largely succeeded in elucidating details of human behavior by using a microscopic approach that controls almost everything except for some small element of human decision making and studying that element in an environment meant to highlight its effect. The relatively coarse experimental methods necessary to compare decision-making techniques may preclude such finesse.

6.2. *What if DA Doesn't Work?* Asking whether DA works is risky. What if the answer is negative? The contribution will clearly be scientifically valuable, but many individuals -- consultants, academics, instructors -- with a vested interest in DA could lose

standing, clients, or even jobs. Pursuing decision-effectiveness research and being faithful to the science will clearly require maintaining an open mind regarding potential results! At the same time, results from the research may elucidate reasons for poor performance or yield insight on how to combine the best of different techniques to create even more effective decision-making methods. It is also completely possible that we learn that the real difference in performance is due not to the specific technique used, but to the particular analyst constructing the model!

6.3. *Publication Risk.* For academics interested in establishing a reputation, pursuing research on decision effectiveness is itself a risky prospect, and both of the dilemmas posed above contribute to that risk. Because of the potential threat to existing paradigms it may be difficult to publish research results. Add to this the nontraditional orientation in terms of behavioral decision theory, and publication may become quite problematic. And longitudinal studies, which may take years to complete, are certainly not for the weak of heart or the untenured! All of these issues will pose difficulties for researchers in this area, although they are clearly not insurmountable. The presence of such barriers suggests that researchers must be careful to maintain impeccable standards in their work, both in execution of experiments and communication of results.

6.4. *Changing Preferences.* To show strong effectiveness of decision techniques will inevitably require researchers to engage in longitudinal studies. The logistical issues of such studies suggest the need to do the work within the context of a DA research center with an established long-term program to study decision making. Beyond the logistical issues, however, are some more fundamental issues. Following participants as they mature and experience consequences over time means that researchers will inevitably be faced with changes in participant preferences. The reasons for making a choice at one time may no longer be compelling later; this divergence between decision and experience utility may challenge long-term studies. What does it mean to say that a decision technique is effective at getting us what we want if what we want has changed substantially by the time we experience the consequences?

Still more troublesome is the issue of habituation. As individuals make changes in their lives, they become accustomed to their new environments and surroundings, and

their level of satisfaction adjusts to the new situation. Preliminary results (Schkade, personal communication) suggest that after moving from one part of the country to another, US residents typically suffer a transient decrease in satisfaction, but that their self-reported level of satisfaction reverts to the pre-move level after a year or so, and that this result is consistent regardless of where people move or their expressed preferences for locations. For longitudinal studies of decision effectiveness, this may pose a genuine dilemma. What does it mean to determine whether a decision technique is effective if we become accustomed to our circumstances regardless of the path we follow? Put more bluntly, how can any decision technique make me happier if my level of happiness always tends to revert to the same level?

Of course, one answer to the habituation dilemma is to measure non-subjective quantities. Thus, even though we may not be able to determine whether an individual is happier, we may be able to tell whether he or she is wealthier or safer, or whether his or her firm has a greater market share, more profits, or a higher stock price. But the value of these is ultimately subjective, and the real question lies in the subjective experience of an individual who follows a particular path.

Regardless of the dilemmas posed in the last few paragraphs, the many research techniques and studies proposed in this paper can help to determine what works and what does not in helping decision makers to obtain desired consequences, to achieve their objectives, in short to get what they want. Helping individuals and organizations find paths that lead to their objectives is the ultimate goal of decision analysis. Whether DA is truly effective in accomplishing this has yet to be seen, but the research agenda outlined here can help us answer the title question, "Does decision analysis work?"

REFERENCES

- Abramson, B., Brown, J., Edwards, W., Murphy, A., & Winkler, R.L. (1994). Hailfinder: A Bayesian System for Forecasting Sever Weather. Draft.
- Aldag, R.J., & Power, D.J. (1986). An empirical assessment of computer-assisted decision analysis. *Decision Sciences*, 17, 572-588.

- Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica*, 21, 503-546.
- Allais, M., & Hagen, J. [Eds.] (1979). *Expected utility hypotheses and the Allais paradox*. Dordrecht, The Netherlands: Reidel.
- Argote, L., Seabright, M. A., & Dyer, L. (1986). Individual versus group use of base-rate and individuating information. *Organizational Behavior and Human Decision Processes*, 38, 65-75.
- Armstrong, J.S. (1984). Forecasting by extrapolation: Conclusions from 25 years of research. *Interfaces*, 14, 52-61.
- Boiney, L.G. (1995). When efficient is insufficient: Fairness in decisions affecting a group. *Management Science*, 41, 1523-1537.
- Clemen, R.T. (1996). *Making hard decisions: An introduction to decision analysis*. Belmont, CA: Duxbury.
- Clemen, R.T., Jones, S.K., & Winkler, R.L. (1996). Aggregating forecasts: An empirical evaluation of some Bayesian methods. In D.A. Berry, K.M. Chaloner, & J.K. Geweke [Eds.] *Bayesian analysis in statistics and econometrics* (pp. 3-14). New York: Wiley.
- Farquhar, P.H. (1984). Utility assessment methods. *Management Science*, 30, 1283-1300.
- Fischer, G. (1975). An experimental study of four procedures for aggregating subjective probability assessments. Technical report 75-7. Decisions and Designs, Inc.
- Fischer, G.W. (1975). Experimental applications of multiattribute utility models. In D. Wendt & C. Vlek [Eds.], *Utility, probability, and human decision making* (pp. 7-46). Dordrecht, The Netherlands: Reidel.
- Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky, [Eds.] *Judgment under uncertainty: Heuristics and Biases* (422-444). Cambridge: Cambridge University Press.
- Fishburn, P.C., & Sarin, R.K. (1994). Fairness and social risk I: Unaggregated analyses. *Management Science*, 40, 1174-1188.
- Fishburn, P.C., & Sarin, R.K. (in press). Fairness and social risk II: Aggregated analyses. *Management Science*, forthcoming.
- Folger, R. (1977). Distributive and procedural justice: Combined impact of "voice" and improvement on experienced inequity. *Journal of Personality and Social Psychology*, 35, 108-119.
- Frisch, D., & Clemen, R.T. (1994). Beyond expected utility: Rethinking behavioral decision research. *Psychological Bulletin*, 116, 46-54.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102, 684-704.
- Gough, R. (1975). The effects of group format on aggregate subjective probability distributions. In *Utility, Probability, and Human Decision Making*. Dordrecht, Netherlands: Reidel.
- Gustafson, D. H., Shukla, R. U., Delbecq, A., & Walster, G. W. (1973). A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups, and nominal groups. *Organizational Behavior and Human Performance*, 9, 280-291.
- Hershey, J.C., Kunreuther, H. C., & Schoemaker, P.J.H. (1982). Sources of bias in assessment procedures for utility functions. *Management Science*, 28, 936-954.
- Hora, S. C., Dodd, N. G., & Hora, J. A. (1993). The use of decomposition in probability assessments on continuous variables. *Journal of Behavioral Decision Making*, 6, 133-147.
- Humphreys, P.C., & McFadden, W. (1980). Experiences with MAUD: Aiding decision structuring versus bootstrapping the decision maker. *Acta Psychologica*, 45, 51-69.

- John, R.S., von Winterfeldt, D., & Edwards, W. (1983). The quality and user acceptance of multiattribute utility analysis performed by computer and analyst. In P. Humphreys, D. von Winterfeldt, & W. Edwards [Eds.], *Analysing and aiding decision processes* (pp. 301-319). Amsterdam: North-Holland.
- Kahneman, D. (1994). New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics*, 150/1, 18-36.
- Kahneman, D., & Snell, J. (1990). Predicting utility. In R.M. Hogarth (Ed.), *Insights in decision making*, Chicago: University of Chicago Press.
- Kahneman, D., & Snell, J. (1992). Predicting a changing taste: Do people know what they will like? *Journal of Behavioral Decision Making*, 5, 187-200.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.
- Keeney, R.L. (1992). *Value-focused thinking: A path to creative decision making*. Cambridge, MA: Harvard University Press.
- Kimbrough, S.O. & Weber, M. (1994). An empirical comparison of utility assessment programs. *European Journal of Operational Research*, 75, 617-633.
- Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, 89, 46-55.
- Lichtenstein, S., Fischhoff, B., & Phillips, L.D. (1982). Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky, [Eds.] *Judgment under uncertainty: Heuristics and Biases* (306-334). Cambridge: Cambridge University Press.
- Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1, 111-153.
- Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. (1993). The M-2 competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, 9, 5-22.
- McCord, M., & de Neufville, R. (1987). Lottery equivalent: Reduction of the certainty effect problem in utility assessment. *Management Science*, 32, 56-60.
- Morgan, M. G., & Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge: Cambridge University Press.
- Murphy, A.H., & Winkler, R.L. (1970). Scoring rules in probability assessment and evaluation. *Acta Psychologica*, 34, 273-286.
- Myers, D. G., & Lamm, H. (1975). The polarizing effect of group discussion. *American Scientist*, 63, 297-303.
- Newbold, P., & Granger, C.W.J. (1974). Experience with forecasting univariate time series and the combination of forecasts (with discussion). *Journal of the Royal Statistical Society, Series A*, 137, 131-149.
- Ordonez, L.D., Mellers, B.A., Chang, S.-J., & Roberts, J. (1995). Are preference reversals reduced when made explicit? *Journal of Behavioral Decision Making*, 8, 265-277.
- Reilly, R.J. (1982). Preference reversal: Further evidence and some suggested modifications in experimental design. *The American Economic Review*, 72, 576-584.
- Saaty, T. (1980). *The Analytic Hierarchy Process*. New York: McGraw-Hill.
- Savage, L.J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66, 783-801.

- Schaefer, R.E., & Borcharding, K. (1973). The assessment of subjective probability distributions: A training experiment. *Acta Psychologica*, 37, 117-129.
- Seaver, D. A. (1978). Assessing probability with multiple individuals: Group interaction versus mathematical aggregation (Report No. 78-3). Social Science Research Institute, University of Southern California.
- Stael Von Holstein, C.A.S. (1971a). The effect of learning on the assessment of subjective probability distributions. *Organizational Behavior and Human Decision Processes*, 6, 304-315.
- Stael Von Holstein, C.A.S. (1971b). Two techniques for assessment of subjective probability distributions: An experimental study. *Acta Psychologica*, 35, 478-494.
- Tindale, R. S., Sheffey, S., & Filkins, J. (1990). Conjunction errors by individuals and groups. Paper presented at the annual meeting of the Society for Judgment and Decision Making, New Orleans, LA.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453-458.
- von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge: Cambridge University Press.
- Wallsten, T.S., & Budescu, D.V. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, 29, 151-173.