Clemen, R.T. (2008) Improving and measuring the effectiveness of decision analysis: Linking decision analysis and behavioral decision research. In: T. Kugler, J. C. Smith, T. Connolly, and Y.-J. Son (Eds.) Decision Modeling and Behavior in Complex and Uncertain Environments. New York: Springer, 3-31.

Improving and Measuring the Effectiveness of Decision Analysis: Linking Decision Analysis and Behavioral Decision Research

Robert T. Clemen

Fuqua School of Business, Duke University, Durham, NC 27708-1208, USA
clemen@duke.edu

Summary. Although behavioral research and decision analysis began with a close connection, that connection appears to have diminished over time. This chapter discusses how to re-establish the connection between the disciplines in two distinct ways. First, theoretical and empirical results in behavioral research in many cases provide a basis for crafting improved prescriptive decision analysis methods. Several productive applications of behavioral results to decision analysis are reviewed, and suggestions are made for additional areas in which behavioral results can be brought to bear on decision analysis methods in precise ways. Pursuing behaviorally based improvements in prescriptive techniques will go a long way toward re-establishing the link between the two fields.

The second way to reconnect behavioral research and decision analysis involves the development of new empirical methods for evaluating the effectiveness of prescriptive techniques. New techniques, including behaviorally based ones such as those proposed above, will undoubtedly be subjected to validation studies as part of the development process. However, validation studies typically focus on specific aspects of the decision-making process and do not answer a more fundamental question. Are the proposed methods effective in helping people achieve their objectives? More generally, if we use decision analysis techniques, will we do a better job of getting what we want over the long run than we would if we used some other decision-making method? In order to answer these questions, we must develop methods that will allow us to measure the effectiveness of decision-making methods. In our framework, we identify two types of effectiveness. We begin with the idea that individuals typically make choices based on their own preferences and often before all uncertainties are resolved. A decision-making method is said to be *weakly effective* if it leads to choices that can be shown to be preferred (in a way that we make precise) before consequences are experienced. In contrast, when the decision maker actually experiences his or her consequences, the question is whether decision analysis helps individuals do a better job of achieving their objectives in the long run. A decision-making method that does so is called *strongly effective*. We propose some methods for measuring effectiveness, discuss potential research paradigms, and suggest possible research projects. The chapter concludes with a discussion of the beneficial interplay between research on specific prescriptive methods and effectiveness studies.

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 (e.g., [8,11,57]). So why should we use decision analysis (DA) techniques? Typical answers include "gaining insight" and "being coherent," but the best reason to use DA would be that doing so would more likely get us what we want. Are there any results indicating that this is the case? If an individual or group uses DA, will they be more likely to get what they want? In the long run, are users of DA healthier, wealthier, safer, or in general more satisfied with the consequences of their decisions?

Certainly DA can show an individual how to be coherent in making inferences and choices. That is, adherence to DA principles can promise that your decisions will not be self-contradictory and that the inferences you make will be consistent with the laws of probability. Thus, DA tells us what one should do on the basis of logical argument. However, behavioral decision research (BDR) has shown that people do not always make coherent decisions and internally consistent inferences (e.g., [28,36,38]).

DA and BDR began with rather close ties, with BDR topics arising largely from questions associated with the appropriateness of subjective expected utility as a descriptive model of human behavior. As such, much of the early BDR literature looked at empirical questions about how well people could judge probabilities and the inconsistencies in preference judgments and decisions. In the late 1960s and early 1970s, it was not uncommon to see experiments that evaluated DA methods for assessing subjective probabilities or utilities. However, in reviewing the two literatures over the past 25 years, one comes away with the sense that BDR and DA have taken somewhat different paths. BDR has increasingly focused on psychological processes with less emphasis on helping to improve DA's prescriptive techniques. On the other hand, although DA learned many important lessons from the early BDR work, much of current practice ignores recent developments in BDR and still relies on methods developed in the 1970s and early 1980s. For example, probability assessment in practice still follows principles laid out by Spetzler and Staël Von Holstein [67] (e.g., see [39,49,50]).

In very general terms, DA helps decision makers address their decisions in careful and deliberate ways. Thus, dual-process theories from psychology, especially Sloman [63] and Kahneman [34], can provide a useful framework. Many of the behavioral phenomena studies in BDR can be explained as resulting from an intuitive or nonconscious process, which Kahneman [34] calls System I. DA can be characterized as avoiding System I's distortions and biases by careful and conscious deliberation (System II). Put simply, DA works to get decision makers past System I and into System II.

This chapter has two goals. To begin, we take the position that current best practices in DA are not always successful in avoiding cognitive biases that arise from System I, and that a reasonable strategy is to find specific ways to counteract such biases. Thus, our first goal is to show how researchers can use recently developed psychological models from BDR to develop improved and prescriptive methods for DA. Our proposal goes beyond general statements about how awareness of behavioral biases can help decision makers avoid pitfalls. Instead, an appeal is made to take BDR results and models directly into the DA domain and to develop precise prescriptive methods that, according to the proposed theory, should improve judgment and/or decision making in a specific and systematic way. In Section 2 we briefly review research on probability and preference assessment, highlighting several recent examples that have used BDR results as a basis for new DA methods.

Suppose that we are indeed able to develop new prescriptive methods. How will we know if they work? Although it may be possible to demonstrate in the lab that a method reduces a particular behavioral effect or bias, the question remains whether the method, applied in real decision settings, will genuinely be of value to decision makers. Put another way, the question is whether specific, BDR-based prescriptive methods will be more effective than current DA practice or unaided intuitive judgment in getting a decision maker what he or she wants. Thus, our second goal in the chapter is to describe research paradigms that could be used to measure the effectiveness of DA methods. Studies of effectiveness may complement the development of prescriptive DA methods by highlighting important behavioral questions about why various methods perform as they do, and behavioral research can in turn suggest specific ways to improve decision analysis.

In evaluating DA and other decision-making techniques, the basic research questions are, "Is XYZ decision-making technique effective in getting people what they want? How does XYZ compare with ABC in terms of effectiveness?" Answering such questions is somewhat problematic. One cannot, for example, compare two different alternatives that an individual might have chosen in a particular decision situation; 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 under different circumstances.

Measuring the effectiveness of decision-making techniques is a research problem fraught with challenges. For some of these challenges, satisfactory approaches are easily identified, but responding to others will require creative new methods. Nevertheless, comparative studies can be performed using readily available approaches, and we describe some ways researchers can carry out such studies.

The remainder of this chapter is organized as follows. The next section delves into the relationship between BDR and DA, with particular attention to existing examples of a productive relationship and suggestions for other areas that could yield important improvements in DA methodology. Our review is selective and meant to highlight particular streams of research. In the following two sections, we turn to the chapter's second goal of measuring the effectiveness of DA methods; we define the concepts of strong and weak effectiveness and describe in some detail a variety of possible studies and a

general research agenda. The final section concludes with a discussion of the potential benefits that BDR and DA can gain from the development of specific prescriptive methods and studies of effectiveness.

2 Using BDR to Improve Prescriptive DA Methods

DA methodology derives from the subjective expected utility paradigm, with a strong focus on subjective judgment of probabilities and assessment of personal preferences. Much of BDR has likewise focused on these issues. There are, of course, many aspects of decision making that fall outside the scope of subjective expected utility, such as generating alternatives, identifying objectives to consider, or identifying and modeling relevant risks [24]. DA has developed methods for dealing with these and other aspects of decision making, and BDR has studied some of them. For our purposes in this chapter, however, we focus on subjective assessment of probabilities and modeling and assessment associated with value and utility functions.

2.1 Probability Assessment

Early Work

Early work on subjective probability judgments by Kahneman and Tversky and others (see [36]) emphasized how heuristic judgment processes can lead to biases. This work was important for the development of standard DA procedures for eliciting subjective probabilities [67]. For example, the anchor-and-adjust heuristic in particular has played an important role in DA, because overconfidence is one of the most persistent biases that decision analysts face. For the assessment of probability distributions for continuous variables, Spetzler and Staël Von Holstein advocated pushing experts to reason about scenarios that could lead to extreme events and to adjust their probabilities after such reasoning.

Much of the early behavioral work on probability judgments focused on calibration of subjective probabilities (see [41]), demonstrating the extent to which subjective probability judgments did not match the objective frequency of occurrence of corresponding events. Some efforts were made to find ways to improve calibration. For example, Staël Von Holstein [68,69] and Schaefer and Borcherding [60] reported that short and simple training procedures could improve the calibration of assessed probabilities, although their results did not show an overwhelming improvement in performance. Fischhoff [17] discussed debiasing techniques intended to improve the quality of subjective probability assessments.

More recently, specific lines of inquiry have yielded results that are important for DA; we mention two here. The first is from the work of Gigerenzer and colleagues [26,27] and includes asking the expert questions that are "ecologically consistent" with those typically encountered in his or her domain of expertise and framing assessment questions in terms of relative frequencies. Both can substantially improve calibration, but neither is a panacea. By their very nature, risk assessments often require experts to go beyond their day-to-day experience (e.g., "What is the probability of a failure of a newly redesigned system in a nuclear reactor?"). Also, not all assessment tasks can be readily reframed in frequency terms. Consider the nuclear reactor question. Given an entirely new system design, how would the analyst describe an equivalence class for which the expert could make a relative-frequency judgment?

The second is the literature on decomposition of probability judgments, the process of breaking down an assessment into smaller and presumably more manageable judgment tasks, making these simpler judgments, and then recombining them using the laws of probability to obtain the overall probability desired. For example, Hora et al. [31] showed that decomposition can improve assessment performance, and Clemen et al. [10] found similar results in the context of aggregating expert judgments. Morgan and Henrion [50] reviewed the empirical support for decomposition in probability judgment.

Recent Directions: Underlying Processes

More recent BDR work on probability judgment has turned to understanding the processes underlying observed biases. One example is the notion of dual processing systems. For example, Sloman [63] makes the case for associative and rule-based reasoning systems and their impact on judgments of uncertainty. Kahneman and Frederick [35] and Kahneman [34] propose "System I," the quick, intuitive processing system, and "System II," the deliberative reasoning system. They argue that when an individual makes a probabilistic judgment, the process begins in System I, which is subject to a variety of nonconscious effects, one of which is the substitution of features of an object for the characteristic being judged. For example, when asked for a probability judgment, an individual may use availability, representativeness, or affect as substitutes for genuine (and more deliberative) judgments of likelihood.

Another example that emphasizes psychological process is the affect heuristic (e.g., [65]), whereby an individual makes judgments based on his affective response to the stimulus. The affect heuristic applies to both probability and preference judgments; Slovic et al. discuss the underpinnings of this heuristic, including how it ties into the fundamental workings of memory and emotion.

A third example is support theory [58,72], which provides a theoretical framework for understanding the psychological process by which an individual generates probability statements. An important feature of support theory is the support function, a modeling construct that represents how an individual summarizes the recruited evidence in favor of a *hypothesis* (a particular description of a possible event). The notation $s(A)$ is used to represent the

support for hypothesis A; that is, it represents the individual's evaluation regarding the strength of evidence in favor of A. According to support theory, $s(A)$ is not *description invariant*; different descriptions of the same event do not necessarily carry the same support. For example, suppose A is "precipitation tomorrow," which can be decomposed into "rain or frozen precipitation tomorrow." Although the two descriptions are meant to designate the same event, support theory contends that s(precipitation tomorrow) may not be equal to s(rain or frozen precipitation tomorrow). In fact, support theory typically assumes that describing an event A as a union of disjoint events (A_1 or A_2) increases support and the sum of the support for disjoint events is typically larger than the support of the union of those events. In symbols,

$$s(A) \leq s(A_1 \text{ or } A_2) \leq s(A_1) + s(A_2) .$$

In turn, differences in support due to different descriptions can lead to different stated probabilities for the same event A, depending on how A is described. Fox and Tversky [23] argue that this process is separate from and prior to the decision-making stage when an individual must make a decision under uncertainty.

More recently, Fox and his colleagues [21,22,62] show that judgments of probabilities are subject to a bias they call *partition dependence*. For example, when experimental participants were asked to assess the probability that the NASDAQ stock index would close in particular intervals, the assessed probabilities depended strongly on the specific intervals that were specified. Fox and colleagues argue that partition dependence stems from a heuristic in which the individual begins with a default prior probability distribution that assigns equal likelihood to each element in the state space; they dub this default distribution the *ignorance prior*. Because individuals tend not to adjust the ignorance prior sufficiently to account for their information, the result is that judged probabilities can depend strongly on how the state space is partitioned in the first place.

In their conclusion, Fox and Clemen [21] connect the idea of partition dependence and the ignorance prior to standard DA practices in probability assessment. They argue that probability assessment occurs in three separate stages, and that different biases are likely to operate in different stages. In particular, they argue that standard DA practice is well suited to reducing biases that occur in the first and second stages (interpretation of categories and assessment of support). For example, working with experts to define and elaborate the interpretation of each event to be judged can reduce effects due to ambiguity, and in the assessment of support the analyst can encourage an expert to fully articulate her reasoning to reduce the effects such as availability or representativeness. However, biases in the third stage, mapping of support into stated probabilities, include the tendency to anchor on the ignorance prior. Fox and Clemen argue that this bias may resist correction because it is not particularly amenable to conscious reflection. They suggest a number of ways in which the analyst can work with the expert to minimize partition

dependence, including the use of multiple partitions for use in the assessment process.

Fox and Clemen [21] make an explicit connection between their behavioral results and DA practice, however, Clemen and Ulu [13] take this connection one step further. Building on the idea of partition dependence, they construct a model of the probability judgment process that is consistent with a number of known properties of subjective probabilities, including partition dependence as well as binary complementarity and subadditivity (e.g., [72]). In addition, they show that their model is consistent with interior additivity, a property observed by Wu and Gonzalez [78]. In one of their experiments, they observed that they could calculate the revealed or "indirect" probability of event A as $P'(A) = P(A \cup B) - P(B)$ for a variety of specifications of the auxiliary event B, and their various calculations of $P'(A)$ tended to be highly consistent. Furthermore, an individual's direct assessment of $P(A)$ tended to differ substantially from the indirect probabilities $P'(A)$.

Clemen and Ulu [13] present empirical evidence in support of their model. More importantly for our purposes, they show that indirect probabilities, after being normalized to sum to one across the state space, are not biased by the ignorance prior (according to their model) and hence should not display partition dependence. Their empirical results, although preliminary, support this contention. Thus, Clemen and Ulu suggest that decision analysts can use normalized indirect probabilities as a way to counteract partition dependence.

Although early empirical BDR work was able to provide good general guidance to decision analysts, Clemen and Ulu's work shows that it is possible to build on psychological theory to develop a precise prescriptive procedure for DA. Whether the use of normalized indirect probabilities genuinely improves on standard practice will no doubt be the subject of future empirical studies.

2.2 Understanding and Assessing Preferences

Assessing Utility Functions for Risky Choices

Much of the work on preferences under risk has focused on the extent to which expressed preferences for lotteries are internally consistent, as exemplified by the Allais paradox [1,2] or Tversky and Kahneman's [70] work on framing. An especially relevant early example is the phenomenon of preference reversal as described by Lichtenstein and Slovic [42]. Stated preferences may reverse depending on response mode (choosing between two risky alternatives versus specifying a value, typically a probability that makes the two alternatives equally preferable). The result is robust, having been demonstrated in many different domains and different forms. This result was then and continues to be important for DA practice, because it shows that preference elicitation methods that are equivalent under subjective expected utility do not necessarily yield consistent responses. Ordóñez et al. [53] reviewed the preference

reversal literature and also studied whether preference reversals can be reduced by "debiasing" [17]. Having subjects perform the two assessment tasks simultaneously yielded little improvement but 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, can lead to a merging of the preference patterns displayed in the different tasks.

Hershey et al. [30] discussed biases induced by different preference-elicitation approaches in spite of their formal equivalence. One such bias is the certainty effect [70], whereby individuals tend to overweight certain or nearly certain outcomes. Understanding the certainty effect is important for DA, because standard methods for assessing utility functions under risk use reference lotteries that compare a risky lottery to a sure outcome. In order to account for the certainty effect, McCord and de Neufville [48] propose an assessment method in which the individual compares two lotteries. Wakker and Deneffe [76] took this idea one step further with their *tradeoff method*, in which the individual compares two 2-outcome lotteries at a time, each involving the same probabilities of winning and losing. The individual's task is to specify one of the four outcomes (x_i) so that she is indifferent between the two lotteries. The assessed value is then used to construct the next pair of lotteries, the assessed value (x_{i+1}) from which leads to the next lottery, and so on. Given the way in which the lotteries are constructed, it is possible to show that the assessed values $x_1, \ldots, x_n$ are equally spaced in terms of their utility values. Moreover, the tradeoff method works for assessing utility functions under uncertainty or for assessing value functions in nonexpected utility models.

Wakker and Deneffe's [76] tradeoff method is a good example of a specific DA method that was developed to account for a particular behavioral phenomenon. Another example comes from Bleichrodt et al. [6], who, like Clemen and Ulu [13], develop a prescriptive method on the basis of a specific behavioral model. Bleichrodt et al. argue that utility assessments are systematically biased in terms of loss aversion and probability weighting as specified by prospect theory [37,71] regardless of the particular assessment method used. They further show how to use the prospect theory model to remove the bias. Doing so, of course, requires estimating the model parameters. Although ideally one would estimate parameters for the particular individual making the assessment, the authors find that even using the aggregate estimates from Tversky and Kahneman [71] can improve consistency across different utility assessment methods.

Assessing Multiattribute Preferences

Although many different biases have been identified in the DR literature on multiattribute assessment, we focus here on two key issues: the scale compatibility and attribute splitting effects.

Slovic et al. [66] define the scale compatibility effect as a bias that occurs when an attribute's weight is enhanced because the scale on which that

attribute is measured is compatible with (or easily commensurable with) the scale of the response mode. For example, suppose a decision maker must judge weights for several attributes, some of which are naturally represented by dollars (e.g., profits, cost, or taxes) and some that are not (e.g., lives lost or environmental damage). A typical assessment method requires the decision maker to "price out" the various alternatives, or to identify an amount of each attribute that is consistent with a particular dollar value. Such an approach is relatively common, for example, in contingent valuation methods. However, due to the scale compatibility effect, those attributes that are already in dollar terms or can be easily converted to dollars will tend to be overweighted, whereas those that are not readily represented in dollar terms will tend to be underweighted. Tversky et al. [74] proposed scale compatibility as a key explanation of preference reversals.

If the scale compatibility effect stems from choosing a particular attribute to be the numeraire in judging weights, then a reasonable approach is not to identify a single attribute. This is the approach taken by Delquié [14], who proposes *bidimensional matching* as a prescriptive assessment method. Instead of changing only one attribute in order to identify a preferentially equivalent option, Delquié suggest changing two attributes at once. The two attributes are varied in a systematic way until indifference is found. Delquié's experiments show that this approach does reduce the scale compatibility effect.

Anderson and Hobbs [3] take a different approach. They develop a model in which the scale compatibility effect is represented by a bias parameter in the model. Using Bayesian statistical methods, they show that one can use a set of tradeoff assessments to derive a posterior distribution for the bias parameter and for the individual's weights. Thus, Anderson and Hobbs's approach is similar to that of Clemen and Ulu [13] and Bleichrodt et al. [6] in processing an individual's judgments ex post in order to adjust for anticipated biases.

Another behavioral issue in assessing multiattribute weights is the attribute splitting effect [77]. The attribute splitting effect has to do with how attributes are structured in a hierarchy. For example, consider the two hierarchies shown in Figure 1. The task would be to assess global weights w_A, w_B, and w_C for attributes A, B, and C, respectively. In the left-hand panel, the decision maker would judge these three weights directly. In the right-hand panel, the assessment is broken down. The decision maker judges the local weights v_A and $v_{A'}$ for A and A′ separately from the local weights v_B and v_C for B and C. Combining the local weights to obtain the global weights, we have $w_{A^*} = v_A$, $w_{B^*} = v_{A'}v_B$, and $w_{C^*} = v_{A'}v_C$. The problem arises from the fact that w_{A^*} obtained using the two-level hierarchy tends to be greater than w_A obtained from the one-level hierarchy. The attribute splitting effect is especially problematic for decision analysts, because any value hierarchy with more than two attributes can be represented in multiple ways (and hence may lead to different weights), and there is no canonical representation.

The attribute splitting effect is similar to the ignorance prior effect discussed above [21,22,62]. A reasonable hypothesis might be that individuals

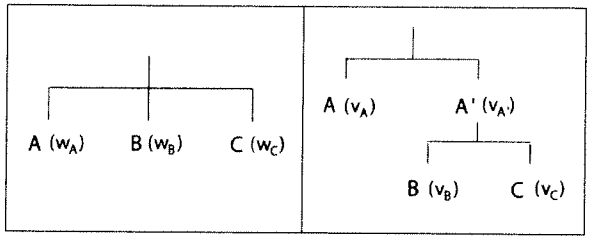

Fig. 1. Two value trees for assessing weights for attributes A, B, and C.

begin with equal weights for attributes at the same level in the hierarchy, but then adjust insufficiently. Starting from this premise, Jacobi and Hobbs [33] offer four possible models to account for attribute splitting. They test their four models in the context of an electrical utility firm evaluating environmental and economic attributes of alternative plans to expand generation capacity. The model that performs the best is similar to Clemen and Ulu's [13] linear model of probability judgment: a weighted linear combination of a default weight (equal for all attributes at the same level of the hierarchy) and the decision-makers' "true" weight performs the best by a variety of measures. Thus, like Clemen and Ulu [13], Bleichrodt et al. [6], and Anderson and Hobbs [3], Jacobi and Hobbs [33] take the approach of adjusting the decision-maker's judgments ex post to account for the attribute splitting bias.

Constructed Preferences, Emotions, and DA

Although ex post adjustment of assessed weights makes some sense as a way to correct for the attribute splitting effect, it is not clear that this is the most appropriate way to frame the problem. An alternative is to think about the effect as a result of the individual's process of constructing a response to the assessment question or, more fundamentally, from constructing preferences themselves in a situation where the issues, attributes, and tradeoffs are unfamiliar and poorly understood. In such a situation, clear preferences may not be readily articulated and may require careful thinking. The problem, of course, is that the way in which the assessment questions are asked can direct that thinking in particular ways that may affect the eventual preference judgments.

Other judgmental phenomena can be viewed in terms of constructed preferences [54,64]. One particularly intriguing example is the role that emotions play in decision making. We mentioned the affect heuristic [65] in the context of probability judgment, but affect plays perhaps a larger role in preference assessment. For example, Loewenstein et al. [43] characterize how individuals respond emotionally to risky situations, and how such responses can have an impact on both judgments and decisions, and Luce et al. [44] show that negative emotion can arise from thinking about tradeoffs in decisions, and that the emotion can affect decisions. Hsee and Rottenstreich [32] show that when

feelings predominate, judgments of value tend to reflect only the presence or absence of a stimulus, whereas when deliberation predominates, judgments reflect sensitivity to scope of the stimulus.

Exploring the interplay between emotions and decision making leads to deep psychological issues. Although the literature in this area is vast, we offer two examples that relate affect and cognitive functioning. First, in studying self-regulation, Baumeister and his colleagues have found that self-regulation, which includes suppressing one's emotions, generally appears to consume some of the brain's executive resources [5,51] and can lead to reduced cognitive functioning [61]. Second, Ashby et al. [4] observe that positive affect generally improves cognitive functioning, and they offer a neurochemical explanation. Positive affect is associated with increased levels of dopamine, which in turn has been shown to be associated with various aspects of cognitive function, including improved creative problem solving and cognitive flexibility. If such effects are occurring at the general level of cognitive functioning, it seems evident that emotions, positive or negative, and having to cope with those emotions in a complex decision situation can have substantial impacts on an individual's ability to think deliberately about tradeoffs. By extension, in a situation where preferences are not well articulated, emotions could have a profound effect on preference construction.

Is it possible to develop prescriptive DA methods that respond to the issues raised by the constructed-preference view? Although the idea of constructed preferences has been known to decision analysts for some time (e.g., [16,75]), no specific methodology has been proposed. If anything, the argument has become circular. For example, Gregory et al. [29] critique the contingent valuation methodology, adopting a constructive-preferences view. As a replacement for contingent valuation, they recommend multiattribute utility analysis, a mainstay of DA methods, arguing that the decomposition approach of multiattribute utility can help individuals think deliberately through complex and unfamiliar tradeoffs. Likewise, Payne et al. [55] appeal to many DA techniques, including multiattribute utility assessment, in describing a "building code" for constructed preferences. As we have seen above, however, even standard DA methods can affect the way in which preferences are constructed and expressed.

For the time being, finding good prescriptive ways to manage the construction of preferences appears to put BDR and DA at an impasse. Before continuing to an equally difficult topic, measuring the effectiveness of DA methods, we look for hopeful signs. Both behavioral researchers and decision analysts have a growing understanding of constructed preferences. That understanding is undoubtedly the first step. And if DA methods are viewed as a basis for avoiding many of the pitfalls in constructed preferences, it may be time for behavioral researchers and decision analysts to find productive ways to collaborate on this problem.

3 What Does "Effectiveness Mean?

3.1 Strong and Weak Effectiveness

The simple answer to the question, "What does effectiveness mean?" is that DA and other decision-making techniques are effective to the extent that they help us achieve what we want to achieve. Thus, we must measure the quality of the consequences we get—in terms of what we want—as a function of the decision-making method used. This perspective is 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 [24]. In particular, a consequentialist perspective does not include any value that might be obtained from the decision-making process itself.

In the spirit of Keeney [40], we assume that it is possible to identify a decision-maker's objectives at the time of a decision. Measuring effectiveness then requires measuring achievement of these objectives. If a decision method tends to lead to consequences that represent a high level of achievement of the decision objectives, we say that the method is *strongly effective*. In contrast, if a method tends to generate choices that informed judges generally view as preferable at the time the action is taken, then we say that the method is *weakly effective*. Although weak effectiveness may appear to be trivial (of course the decision maker must prefer the chosen alternative!), it is not when viewed more broadly. Showing weak effectiveness may be accomplished by showing that alternatives chosen by decision makers using a particular technique are judged to be preferred, or even dominant, when compared to alternatives chosen by other methods. The judgment of alternatives is made ex ante (i.e., in the context of making the decision before experiencing the consequence) by an appropriate sample of individuals. We expand on this approach below and make precise what we mean by "an appropriate sample of individuals" and how we can use their judgments to measure weak effectiveness.

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 imagine that a prescriptive technique could mislead a decision maker into thinking that the recommended alternative dominates all others, although the consequence eventually obtained from the recommended alternative would be inferior compared to the consequences from other alternatives. The issue here is not only the distinction between decision utility and experience utility in a specific decision situation, but also the extent to which a decision method itself can lead to a discrepancy between the two.

3.2 Elements of Value

Saying that a technique is effective when it helps one to achieve his objectives to a greater degree begs the question of what those objectives might be. Keeney [40] describes the process of identifying one's objectives for decision-making purposes. 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. Table 1 lists some generic objectives that may be of interest to individuals, small groups, corporations, or public-policy organizations. In what follows, we use the term "decision maker" to refer to the decision-making unit, regardless of whether that unit consists of one or more individuals.

The objectives in Table 1 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; we might call these objectives "why we govern," and the corporation's objectives could be described as "why we engage in economic activity."

Table 1 provides guidance as to what sort of objectives must be measured in order to determine effectiveness. Although specific decision contexts may have specific objectives, that objective is probably related to one of the objectives in Table 1. For example, maximize starting salary in the context of an individual's job search is related to a wealth objective. 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 method or another.

Table 1. Typical objectives of different decision makers.

- *Individual:* 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

Often, adequate measures can be found using standard DA techniques [40], and measures for many of the objectives listed in Table 1 may be found in this way. However, measuring achievement of some objectives may be straightforward whereas others are quite difficult. For example, measuring overall health

of a group of constituents may be achieved using standard epidemiological survey methods, likewise, wealth as measured in the relevant currency in principle, although there is the typical problem of obtaining truthful responses to questions about private matters. How does one measure wisdom, though, or respect? 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. Boiney [7] and Fishburn and Sarin [18,19] provide decision-theoretic procedures for evaluating fairness of allocations (including risky allocations) using the concept of envy among stakeholders. Measuring fairness of process is somewhat more problematic, but not impossible. For example, stakeholders who have a "voice" in a public-policy decision often perceive the process to be fairer than those who have no voice in the decision (e.g., [20]). Thus, one possibility for measuring process fairness would be to measure the extent to which stakeholders are given a voice in a decision (and the extent to which the stakeholders perceive themselves as having a voice).

Aside from the objectives listed in Table 1, it is also possible for a decision maker or organization to obtain value from the decision-making process itself. For example, an individual may enjoy the process of discovering her values, or may gain useful experience that can be applied to similar problems later. In an organization, improved communication among workers and heightened commitment to a path of action can result from decision making [12]. Although we acknowledge the importance of value that derives from the process itself, in this chapter we focus on consequence-oriented objectives such as those in Table 1 rather than process-oriented objectives.

4 Measuring Effectiveness

Virtually no research has been done that compares DA with other decision-making techniques in terms of strong effectiveness. Relatively little work has been done to show weak effectiveness. As mentioned above, much of the early work on behavioral decision was motivated by expected utility and implicitly asked whether DA techniques that stem directly from expected utility theory (e.g., probability and utility assessment) were weakly effective in the narrow sense of being able to provide accurate models of a decision-maker's preferences or beliefs about uncertainty. However, as argued by Frisch and Clemen [24], many elements of decision making fall outside the expected utility paradigm per se. For example, expected utility theory sheds little light on how to identify one's objectives or how to find new alternatives.

In this section, we have two goals. First, we describe some research paradigms that might be used to measure the effectiveness of DA and other decision-making methods. Second, we give examples of specific effectiveness studies.

4.1 Measuring Strong Effectiveness

Longitudinal Studies

Studies of strong effectiveness must ultimately embrace the challenge of longitudinal studies. In most important decisions for individuals, small interest groups, corporations, or public-policy decision makers, consequences are experienced over time. Thus, one obvious way to measure effectiveness is to recruit participants, subject them to manipulations regarding the use of particular decision-making methods, and to track over time the extent to which identified objectives are achieved.

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 decision situation and an adequate sample of participants who face that situation. For example, consider college graduates making decisions about careers. At a large school with a strong alumni program, it may be possible to keep track of individuals who have gone through a particular manipulation as part of making career choices. In modeling multiattribute preferences for jobs or careers, for instance, their judgments may be taken at face value or adjusted for scale compatibility effects as discussed above [3]. Another example might be upcoming retirees for a large corporation; as employees approach retirement, it may be possible to recruit some individuals as participants in a study that manipulates decision techniques for retirement planning. In this case, one might study the effectiveness of Bleichrodt et al.'s [6] approach for assessing risk-averse utility functions, which are then used for making portfolio allocation decisions. A third group might be entrepreneurs, segregated into subgroups according to decision methods used. For entrepreneurs, appropriate objectives to measure may include number of profitable ventures launched, capital attracted from outside investors, or total return on investment over a specified period of time. Here, one could imagine testing different methods for counteracting partition dependence in the entrepreneurs' probability judgments, using methods suggested in Fox and Clemen [21] or Clemen and Ulu [13].

One area that seems particularly apt for longitudinal studies is the medical arena. What are needed are clinical trials of decision-making methods; patients with the same condition and treatment options could be randomized into different groups in which individual decisions would be made based on different methods. The study would follow the progress of the patients and compare their conditions after specified periods of time depending on the particular condition being treated. The results would compare the effectiveness of the different decision methods under investigation. Some related work has been done. For example, Clancy et al. [9] and Protheroe et al. [56] showed that using decision analysis can influence individual medical decisions (screening or vaccinating for Hepatitis B in the former, treatment of atrial fibrillation in the latter). However, in neither case were patients followed in order to

track their health outcomes. Fryback and Thornbury [25] showed that the use of decision analysis, even informally, can affect physicians' diagnoses when evaluating radiological evidence of renal lesions. Their results showed that the DA-based diagnoses tended to be more accurate. The study was retrospective; the physicians examined existing patient records for which the actual outcome was known (but not to the physicians in the study). Thus, the use of DA should improve the expected outcome for a renal lesion patient. A genuine clinical trial as described above would be needed to confirm this conclusion.

Simulation Studies. The logistic difficulties of real-time longitudinal studies and clinical trials reduce their attractiveness as research methods. Simulations may provide a suitable alternative. Corporate and industry simulations, for example, are common fare in business curricula; similar games that would be amenable to manipulations in decision-making techniques could provide a testbed 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 simulated "firms," which interact as members of an industry. Different groups could use specific techniques (e.g., a particular computer decision aid versus generic use of DA modeling methods, including decision trees, Monte Carlo simulation, and optimization). A control group having no specific training or decision-making instructions would provide a benchmark. Each group's results would be measured in terms of the objectives specified in the game and could be compared across groups.

4.2 Measuring Weak Effectiveness

Comparing Expected Values

In contrast to strong effectiveness, studies of weak effectiveness need not be designed to track outcomes and consequences over time. The simplest approach, exemplified by Clemen and Kwit [12], is to compare the expected values of alternatives that are analyzed in a series of decisions. Clemen and Kwit make the comparison by calculating the difference between the expected value of the chosen alternative and the average of the other expected values for the other alternatives analyzed. If it is possible to document the strategy that would have been taken without the analysis (sometimes called the "momentum strategy"), then one can calculate the difference between the expected value of the chosen strategy and the expected value of the momentum strategy.

Regardless of the specific metric used, this approach requires substantial record keeping in a consistent way over many decision-making projects. Results that document positive value of the analysis indicate "bottom line" value added, but do not necessarily document value obtained in other ways. For example, if an organization has an objective of improving communication across functional areas, consistently using DA on projects that cut across such areas may help achieve this objective by imposing a common language for discussing decisions. However, such value is not likely to be documented in the calculation of incremental expected value added by DA.

Panel Preferences

Because it is not always possible to capture all aspects of value in a bottom-line analysis, we broaden the question to ask whether the alternatives generated by a particular decision-making method are viewed as preferable. To operationalize this notion, we propose using a panel of judges. Because we are concerned here with the notion of decision utility, it would be natural to have a panel of judges (e.g., individuals sampled from the same population as the original decision makers) 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; the decomposition approach of DA as well as other formal decision-making methods challenges the view that holistic judgments adequately capture 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 candidate alternatives.

As one possible method, consider the problem of 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. A dominated alternative is one for which another alternative can be found that improves on all of the attributes. The set of nondominated alternatives is often called the *efficient frontier*, because if one member of this set is chosen, it is not possible to switch to another without reducing achievement of at least one objective. In order to operationalize this approach, we need a way to measure an alternative's *efficiency*, its closeness to the efficient frontier.

Figure 2 presents an analytical example for measuring the relative efficiency of a set of alternatives evaluated on two attributes. A decision maker 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 C is dominated by A and that neither A nor B is dominated. Because A and B are both

on the efficient frontier with respect to this particular set of alternatives, we will set their efficiency measures E_A and E_B to 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 $aU(x) + (1-a)V(y) = t$, where a and $(1-a)$ can be thought of as weights in a two-attribute additive utility function. Using the same a, the greatest utility achievable is $t^\star$, represented by the line segment AB, parallel to the indifference curve through C. Thus, we can define E_C to be the ratio $t/t^\star$. This is equivalent to calculating the ratio of the distance DC to the distance DE in Figure 2.

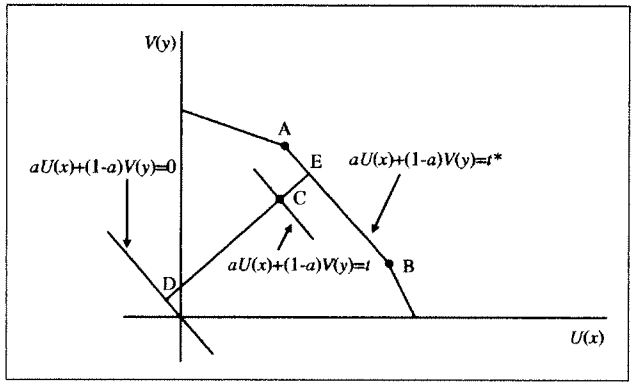

Fig. 2. Measuring the relative efficiency of an alternative. Assuming an additive utility function, alternative C lies on an indifference curve having utility t. A measure of efficiency for C relative to the alternatives included in the evaluation set is $E_C = t/t^\star$. Using the same logic, the efficiency measures for A and B would each be 100%.

We can improve this measure of efficiency slightly. Note that we could have used any of the line segments on the efficient frontier. Each one corresponds to a different set of weights for the additive utility function. To make C appear in the best light possible, we choose the efficient frontier segment that maximizes the ratio $t/t^\star$. Mathematically, suppose we have a set $\mathbf{A}$ of alternatives, and we wish to calculate the efficiency score E_i for alternative $A_i \in \mathbf{A}$, where A_i has attribute values x_i for attribute X and y_i for attribute Y. We define E_i to be

$$E_i = \max_a \left[\frac{aU(x_i) + (1-a)V(y_i)}{\sup_{A_j \in A} [aU(x_j) + (1-a)V(y_j)]} \right].$$

For a particular a, the numerator inside the brackets calculates the utility t for A_i, and the denominator finds the largest possible utility $t^\star$ over all elements of $\mathbf{A}$. The maximization finds the values for a that maximize the ratio $t/t^\star$ for alternative A_i. This 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 a set of alternatives representing choices made by decision makers in a particular decision situation and using specified decision-making methods. 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 the efficiency measures across alternatives. 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 statistical analysis can determine whether alternatives generated by one decision-making method tend to be more efficient than alternatives from another method.

Using panel preferences is not without problems. As with all decision makers, panel judges may be susceptible to many of the biases discussed in the first part of this chapter. For example, we described the problem of constructed preferences. Individual preferences can be shaped by aspects of a decision situation and the task-contingent strategies people use in judgment and choice [54,73]. If we are unsure of the quality of our elicited panel preferences, of what scientific value are the calculated efficiency scores? Although we cannot completely escape the implications of this question, we offer two mitigating considerations. First, to the extent possible, judges should be exposed to the same decision-making environment as the decision makers and should be encouraged to deliberate about the value of the alternatives on the various attributes according to standard procedures and following principles consistent with Payne et al.'s [55] "building code" for decision making. Doing so may not remove all biases, but standard procedures should minimize biases and variability across judges. Second, the approach described above calculates an efficiency score specific to each judge, and therefore agreement across judges is not required. In fact, the procedure for calculating efficiency scores for alternatives for each judge appropriately takes into account variability in preferences across judges; the objective is to identify a subset of alternatives that are preferred in the sense of having (statistically) higher efficiency scores than other alternatives.

4.3 Some Research Projects

The paragraphs above describe some general experimental paradigms that could be used to measure the effectiveness of DA and other techniques. In this section we speculate on some specific studies that might be performed.

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. Strictly proper scoring rules provide natural performance measures that can capture both calibration and skill [52]. In a probabilistic forecasting environment, enough outcomes would have to be recorded to calculate meaningful average scores for the various techniques. Competitions such as this have been run in the forecasting field [45–47].

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 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 such as 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.

Value Structuring and Creativity

In his book, *Value-Focused Thinking* (VFT), Keeney [40] recommends that decision makers should identify their objectives as a first step in the decision-making process, if possible before identifying alternatives. VFT provides many

tools and techniques for identifying, structuring, and using objectives in decision making. The overall process has become an important element in the toolkit of decision analysts who work with decision makers on complex multiobjective problems.

In VFT, Keeney stresses that his approach is valuable for many things, not least of which is its potential for generating creative alternatives. 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 textbooklike 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 in advance.

Fortunately, a procedure such as 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.

Decision-Making Tournament

Several different decision paradigms exist, among them DA, the analytic hierarchy process [59], 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 analytic hierarchy process.

Canonical Decision Problems

Researchers might want to run the tournament just described with technique proponents as participants more than once for each set of cases. After the

initial run, it would be necessary in subsequent runs to ensure that the participants were not prejudiced one way or the other by the outcomes and decisions of prior runs. 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 "naïve" 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.

Ethnological Study of Decision Making

A final example is to take 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.

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 must be refined, tested, and extended to decision situations other than multiattribute choices under certainty. We must learn how to handle the ethics and logistics of longitudinal studies. 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 may not appear attractive in and of itself unless it can directly address substantive questions

of interest. Nevertheless, the importance of such work cannot be overstated for the research program described here; these methodological problems must be solved if definitive results on effectiveness are to be obtained.

With an array of experimental methods available, research can address many aspects of the effectiveness of decision-making methods. Some general examples include the following.

- Compare specific techniques such as DA, analytic hierarchy process, naïve decision making, and so on as overall paradigms for making decisions.
- Explore different aspects of decision making: problem structuring, uncertainty modeling and assessment, preference assessment, and so on.
- Examine 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.
- Study how effectiveness of different methods varies depending on individual characteristics, such as age, education, attitudes (e.g., attitudes toward quantitative analysis or technology), or cognitive abilities (e.g., ability to work with quantitative information).

Many other, more specific studies are also possible, and they might stem from the development of new DA techniques based on BDR findings, or from the introduction of new decision methods.

5 Conclusion: The Interplay of BDR and DA

In the first part of this chapter, we discussed ways in which BDR can be brought to bear on the development of new DA methods. Recent efforts have shown how to use BDR results and theory as a basis for developing better DA methods. We discussed new directions that might prove fruitful, such as considering the role of emotions in decision making and the implications for DA methods. In the second part of the chapter, we discussed in detail how researchers can evaluate decision-making methods. The concepts of strong and weak effectiveness provide a framework for studying effectiveness of decision-making methods, and we considered some generic research approaches and specific projects.

Although we have discussed these two main topics as if they are largely separate, they actually come together in two interesting ways. First, methods from BDR may be useful in developing research paradigms and methods to study effectiveness. Second, the two topics can lead to a research cycle. We have seen that BDR can inform the development of new DA methods. When the new methods are evaluated in terms of effectiveness, knowledge from BDR can be used to explain the results of the effectiveness studies and in turn help researchers refine the methods. As the cycle continues, DA methods improve,

and we would hope that developing and evaluating new DA methods will contribute to the BDR body of knowledge.

An important dilemma that researchers will have to face is the problem of preferences that change over time. For example, in longitudinal studies of strong effectiveness, the researcher must follow participants as they mature and experience consequences over time. The reasons for making the original choice may be less compelling at a later date, leading a decision maker to regret the choice (and, perhaps, the decision to be a participant in the study!). Alternatively, the decision maker may have developed a new rationalization for the original choice based on his new preferences. Yet another possibility is that the decision-making method used led the decision maker to construct his preferences in a particular way, but experiencing the consequences leads to somewhat different preferences. Such complications must be dealt with in order to evaluate strong effectiveness. 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?

Throughout, we have implicitly made three assumptions that some readers may find controversial. First, we have implicitly assumed that deliberative, System II thinking generally leads to better decisions. Some researchers have offered evidence to the contrary. Recently, for example, Dijksterhuis et al. [15] found that, when making consumer decisions, individuals made "better decisions" when they did not process the information consciously compared to when they did. However, their conclusions are not particularly relevant to our assumption for two reasons. First, their experiments were conducted under conditions vastly different from what one would expect when using DA methods. Dijksterhuis et al. presented information about a variety of different product attributes fairly quickly, focusing on each one for a matter of seconds, and providing four minutes for an individual to think about his or her choice. In contrast, we might expect conscious processing to involve considerably more time, including time to examine claims in detail, acquire and assess new information, discuss the matter with others, and so on. Second, their measure of decision quality was subjective in three of their four experiments (either postdecision satisfaction or attitude toward the target object). In the fourth experiment (their Study 1), the experimenters identified the "best decision" as the choice with the most positive aspects, even though different individuals may weight these aspects very differently in their choices. If the "best decision" was a dominant choice—better than any of the other choices on all aspects—then their method seems reasonable, but otherwise does little to suggest either strong or weak effectiveness.

Our second implicit assumption has been that DA methods and the underlying subjective expected utility paradigm are the appropriate deliberative basis for making a decision. However, our argument does not rely on the assumption that DA is "the best" or the "only rational" way to make a decision. All of our arguments could be applied to other deliberative decision-making frameworks, such as the analytic hierarchy process [59] or other multicriteria

methods. In fact, studies of effectiveness could profitably compare different decision-making frameworks. Decision analysts who have held tenaciously to the belief that DA is the best method may be surprised!

Third, in suggesting that a decision-maker's or expert's judgments or preference statements may be improved by adjusting them ex post, we have implicitly assumed that this is a legitimate thing to do. However, it leads to situations that may appear somewhat paradoxical. For example, take the Bleichrodt et al. [6] approach to adjusting a decision-maker's stated risk preferences to account for distortions due to prospect theory. One can imagine the decision maker saying, "My risk tolerance is X," and the analyst saying, "No, your risk tolerance is really Y." Is it reasonable for the analyst to take such a bold step? We offer two justifications. First, the decision maker can be educated (on the basis of sound research) that his or her statements may indeed be biased and that correcting for the biases can improve judgments and choices. Second, in a public-policy setting, experts and stakeholders may state their judgments or tradeoff weights and certify that those statements are the best representation of their beliefs and preferences that they are capable of making. Similarly, the analyst can certify that any adjustments reflect up-to-date scientific knowledge about how to account for known biases. Such certification is already implicitly practiced, for example, by scientists who certify that a particular complex mathematical model represents up-to-date scientific understanding.

Helping individuals and organizations find paths that lead to their objectives is the ultimate goal of research in decision making. Regardless of the issues raised in the paragraphs above, BDR and effectiveness studies can lead to better decision-making methods and thus can ultimately help decision makers achieve their objectives.

Acknowledgments

I am grateful to colleagues to the Fuqua School of Business, the University of Texas at Austin, and the Workshop on Decision Modeling and Behavior in Uncertain and Complex Environments (Tucson AZ, 2006) for the opportunity to present this work and for hours of discussion. Thanks especially to Susan Brodt, Ellen Peters, and an anonymous reviewer for their comments and insights. This work was supported in part by the National Science Foundation under Grant No. SES-0317867. Any opinions, findings, conclusions, or recommendations expressed herein are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

1. M. Allais. Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americaine. *Econometrica*, 21:503–546, 1953.

2. M. Allais and J. Hagen. *Expected Utility Hypotheses and the Allais Paradox.* Reidel, Dordrecht, The Netherlands, 1979.
3. R. M. Anderson and B. F. Hobbs. Using a Bayesian approach to quantify scale compatibility bias. *Management Science*, 48:1555–1568, 2002.
4. F. G. Ashby, A. M. Eisen, and A. U. Turken. A neuropsychological theory of positive affect and its influence on cognition. *Psychological Review*, 106:529–550, 1999.
5. R. F. Baumeister and T. F. Heatherton. Self-regulation failure: An overview. *Psychological Inquiry*, 7:1–15, 1996.
6. H. Bleichrodt, J. L. Pinto, and P. P. Wakker. Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science*, 47:1498–1514, 2001.
7. L. G. Boiney. When efficient is insufficient: Fairness in decisions affecting a group. *Management Science*, 41:1523–1537, 1995.
8. D. Bunn. *Applied Decision Analysis.* McGraw-Hill, New York, 1984.
9. C. M. Clancy, R. D. Cebul, and S. V. Williams. Guiding individual decisions: A randomized, controlled trial of decision analysis. *American Journal of Medicine*, 84:283–288, 1988.
10. R. Clemen, S. K. Jones, and R. L. Winkler. Aggregating forecasts: An empirical evaluation of some Bayesian methods. In D. Berry, K. M. Chaloner, and J. K. Geweke, editors, *Bayesian Analysis in Statistics and Econometrics*, pages 3–14. Wiley, New York, 1996.
11. R. T. Clemen. *Making Hard Decisions: An Introduction to Decision Analysis.* Duxbury, Belmont, CA, second edition, 1996.
12. R. T. Clemen and R. C. Kwit. The value of decision analysis at Eastman Kodak Company, 1990–1999. *Interfaces*, 31:74–92, 2001.
13. R. T. Clemen and C. Ulu. Interior additivity and subjective probability assessment of continuous variables. Unpublished manuscript, Duke University, 2006.
14. P. Delquié. "Bimatching": A new preference assessment method to reduce compatibility effects. *Management Science*, 43:640–658, 1997.
15. A. Dijksterhuis, M. W. Bos, L. F. Nordgren, and R. B. van Baaren. On making the right choice: The deliberation-without-attention effect. *Science*, 311:1005–1007, 2006.
16. G. W. Fischer. Utility models for multiple objective decisions: Do they accurately represent human preferences? *Decision Sciences*, 10:451–479, 1979.
17. B. Fischhoff. Debiasing. In D. Kahneman, P. Slovic, and A. Tversky, editors, *Judgment Under Uncertainty: Heuristics and Biases*, pages 422–444. Cambridge University Press, Cambridge, UK, 1982.
18. P. C. Fishburn and R. K. Sarin. Fairness and social risk I: Unaggregated analyses. *Management Science*, 40:1174–1188, 1994.
19. P. C. Fishburn and R. K. Sarin. Fairness and social risk II: Aggregated analyses. *Management Science*, 43:115–126, 1997.
20. R. Folger. Distributive and procedural justice: Combined impact of "voice" and improvement on experienced inequity. *Journal of Personality and Social Psychology*, 35:108–119, 1977.
21. C. R. Fox and R. T. Clemen. Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science*, 51:1417–1432, 2005.

22. C. R. Fox and Y. Rottenstreich. Partition priming in judgment under uncertainty. *Psychological Science*, 14:195–200, 2003.
23. C. R. Fox and A. Tversky. A belief-based account of decision under uncertainty. *Management Science*, 44:879–895, 1998.
24. D. Frisch and R. T. Clemen. Beyond expected utility: Rethinking behavioral decision research. *Psychological Bulletin*, 116:46–54, 1994.
25. D. G. Fryback and J. R. Thornbury. Informal use of decision theory to improve radiological patient management. *Radiology*, 129:385–388, 1978.
26. G. Gigerenzer. How to make cognitive illusions disappear: Beyond heuristics and biases. *European Review of Social Psychology*, 2:83–115, 1991.
27. G. Gigerenzer, U. Hoffrage, and H. Kleinbölting. Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98:506–528, 1991.
28. T. Gilovich, D. Griffin, and D. Kahneman, editors. *Heuristics and Biases: The Psychology of Intuitive Judgment.* Cambridge University Press, Cambridge, UK, 2002.
29. R. Gregory, S. Lichtenstein, and P. Slovic. Valuing environmental resources: A constructive approach. *Journal of Risk and Uncertainty*, 7:177–197, 1993.
30. J. Hershey, H. C. Kunreuther, and P. J. Schoemaker. Sources of bias in assessment of utility functions. *Management Science*, 28:936–954, 1982.
31. S. C. Hora, N. G. Dodd, and J. A. Hora. The use of decomposition in probability assessments on continuous variables. *Journal of Behavioral Decision Making*, 6:133–147, 1993.
32. C. K. Hsee and Y. Rottenstreich. Music, pandas, and muggers: On the affective psychology of value. *Journal of Experimental Psychology: General*, 133:23–30, 2004.
33. S. K. Jacobi and B. F. Hobbs. Quantifying and mitigating splitting biases in value trees. Unpublished manuscript, Johns Hopkins University, Baltimore, MD, 2006.
34. D. Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 93:1449–1475, 2003.
35. D. Kahneman and S. Frederick. Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman, editors, *Heuristics and Biases: The Psychology of Intuitive Judgment*, pages 49–81. Cambridge University Press, New York, 2002.
36. D. Kahneman, P. Slovic, and A. Tversky, editors. *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge University Press, Cambridge, UK, 1982.
37. D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47:263–291, 1979.
38. D. Kahneman and A. Tversky. *Choices, Values, and Frames.* Cambridge University Press, Cambridge, UK, 2000.
39. R. Keeney and D. von Winterfeldt. Eliciting probabilities from experts in complex technical problems. *IEEE Transactions on Engineering Management*, 38:191–201, 1991.
40. R. L. Keeney. *Value-Focused Thinking: A Path to Creative Decision Making.* Harvard University Press, Cambridge, MA, 1992.
41. S. Lichtenstein, B. Fischhoff, and L. D. Phillips. Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, and A. Tversky, editors, *Judgment Under Uncertainty: Heuristics and Biases*, pages 306–334. Cambridge University Press, Cambridge, UK, 1982.

42. S. Lichtenstein and P. Slovic. Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, 89:46–55, 1971.
43. G. F. Loewenstein, C. K. Hsee, E. U. Weber, and N. Welch. Risk as feelings. *Psychological Bulletin*, 127:267–286, 2001.
44. M. F. Luce, J. R. Bettman, and J. W. Payne. Choice processing in emotionally difficult decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23:384–405, 1997.
45. S. Makridakis, A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1:111–153, 1982.
46. S. Makridakis, C. Chatfield, M. Hibon, M. Lawrence, T. Mills, K. Ord, and L. Simmons. The M-2 competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, 9:5–22, 1993.
47. S. Makridakis and M. Hibon. The M3-competition. *International Journal of Forecasting*, 16:451–476, 2000.
48. M. McCord and R. de Neufville. Lottery equivalents: Reduction of the certainty effect problem in utility assessment. *Management Science*, 32:56–60, 1986.
49. M. W. Merkhofer. Quantifying judgmental uncertainty: Methodology, experiences, and insights. *IEEE Transactions on Systems, Man, and Cybernetics*, 17:741–752, 1987.
50. M. G. Morgan and M. Henrion. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge, UK, 1990.
51. M. Muraven and R. F. Baumeister. Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin*, 126:247–259, 2000.
52. A. H. Murphy and R. L. Winkler. Scoring rules in probability assessment and evaluation. *Acta Psychologica*, 34:273–286, 1970.
53. L. D. Ordóñez, B. A. Mellers, S.-J. Chang, and J. Roberts. Are preference reversals reduced when made explicit? *Journal of Behavioral Decision Making*, 8:265–277, 1995.
54. J. W. Payne, J. R. Bettman, and E. J. Johnson. *The Adaptive Decision Maker*. Cambridge University Press, Cambridge, UK, 1993.
55. J. W. Payne, J. R. Bettman, and D. A. Schkade. Measuring constructed preferences: Towards a building code. *Journal of Risk and Uncertainty*, 19:243–270, 1999.
56. J. Protheroe, T. Fahey, A. A. Montgomery, and T. J. Peters. The impact of patients' preferences on the treatment of atrial fibrillation: Observational study of patient based decision analysis. *British Medical Journal*, 320:1380–1384, 2000.
57. H. Raiffa. *Decision Analysis*. Addison-Wesley, Reading, MA, 1968.
58. Y. Rottenstreich and A. Tversky. Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review*, 2:406–415, 1997.
59. T. Saaty. *The Analytic Hierarchy Process*. McGraw-Hill, New York, 1980.
60. R. E. Schaefer and K. Borcherding. The assessment of subjective probability distributions: A training experiment. *Acta Psychologica*, 37:117–129, 1973.
61. B. J. Schmeichel, K. D. Vohs, and R. F. Baumeister. Intellectual performance and ego depletion: Role of the self in logical reasoning and other information processing. *Journal of Personality and Social Psychology*, 85:33–46, 2003.

62. K. E. See, C. R. Fox, and Y. Rottenstreich. Between ignorance and truth: Partition dependence and learning in judgment under uncertainty. Unpublished manuscript, University of Pennsylvania, 2006.
63. S. Sloman. The empirical case for two systems of reasoning. *Psychological Bulletin*, 119:3–22, 1996.
64. P. Slovic. The construction of preferences. *American Psychologist*, 50:364–371, 1995.
65. P. Slovic, M. Finucane, E. Peters, and D. G. MacGregor. The affect heuristic. In T. Gilovich, D. Griffin, and D. Kahneman, editors, *Heuristics and Biases: The Psychology of Intuitive Judgment*, pages 397–420. Cambridge University Press, Cambridge, UK, 2002.
66. P. Slovic, D. Griffin, and A. Tversky. Compatibility effects in judgment and choice. In R. Hogarth, editor, *Insights in Decision Making: A Tribute to Hillel J. Einhorn*, pages 5–27. University of Chicago Press, IL, 1990.
67. C. S. Spetzler and C.-A. S. Staël Von Holstein. Probability encoding in decision analysis. *Management Science*, 22:340–352, 1975.
68. C.-A. S. Staël Von Holstein. The effect of learning on the assessment of subjective probability distributions. *Organizational Behavior and Human Decision Processes*, 6:304–315, 1971.
69. C.-A. S. Staël Von Holstein. Two techniques for assessment of subjective probability distributions: An experimental study. *Acta Psychologica*, 35:478–494, 1971.
70. A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211:453–458, 1981.
71. A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 26:297–323, 1992.
72. A. Tversky and D. J. Koehler. Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101:547–567, 1994.
73. A. Tversky, S. Sattath, and P. Slovic. Contingent weighting in judgment and choice. *Psychological Review*, 95:371–84, 1988.
74. A. Tversky, P. Slovic, and D. Kahneman. The causes of preference reversal. *The American Economic Review*, 80:204–217, 1990.
75. D. von Winterfeldt and W. Edwards. *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, UK, 1986.
76. P. Wakker and D. Deneffe. Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management Science*, 42:1131–1150, 1996.
77. M. Weber, F. Eisenführ, and D. von Winterfeldt. The effects of splitting attributes on weights in multiattribute utility measurement. *Management Science*, 34:431–445, 1988.
78. G. Wu and R. Gonzalez. Nonlinear decision weights in choice under uncertainty. *Management Science*, 45:74–85, 1999.