Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Innovations in the Science and Practice of Decision Analysis: The Role of Management Science

James S. Dyer , James E. Smith

To cite this article:
James S. Dyer , James E. Smith (2021) Innovations in the Science and Practice of Decision Analysis: The Role of Management Science . Management Science 67(9):5364-5378. https://doi.org/10.1287/mnsc.2020.3652

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 67, No. 9, September 2021, pp. 5364–5378
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Innovations in the Science and Practice of Decision Analysis: The Role of *Management Science*

James S. Dyer,[a] James E. Smith[b]

[a] McCombs School of Business, University of Texas at Austin, Austin, Texas 78705; [b] Tuck School of Business, Dartmouth College, Hanover, New Hampshire 03755

Contact: jim.dyer@mccombs.utexas.edu (JSD); jim.smith@dartmouth.edu, https://orcid.org/0000-0002-9429-7567 (JES)

Received: November 1, 2019
Revised: February 12, 2020
Accepted: February 19, 2020
Published Online in Articles in Advance: August 4, 2020

https://doi.org/10.1287/mnsc.2020.3652

Abstract. In this paper, we reflect on research in decision analysis that has appeared in *Management Science* and its impact on decision analysis practice and applications. We consider professional applications of decision analysis in business and government settings as well as everyday conversational applications.

History: Accepted by David Simchi-Levi, Special Section of *Management Science*: 65th Anniversary.

Keywords: decision analysis

1. Introduction

In the 65 years since its founding in 1954, *Management Science* (*MS*) has played a major role in shaping the field of decision analysis (DA) in both research and practice. Though the term *decision analysis* would not become popular until the 1960s (Howard 1966, Raiffa 1968), from the very beginning, those who founded *MS* recognized the practical potential of what would later be called *decision analysis*. For example, Merrill Flood (1955, *MS*),[1] a professor at Columbia University who was instrumental in founding The Institute of Management Sciences (TIMS) and who would serve as its second president, said at the first TIMS meeting in 1954,

> L. J. Savage, in his *Foundations of Statistics,* offers a probability-utility type theory of decision that shows the close logical connection between any such theory and a very few plausible assumptions about rational behavior. In fact, if the over-all normative problem is in some sense necessarily one requiring probabilistic considerations of valuations leading to conscious choices among known classes of alternatives, then it seems likely that a good many of these interestingly complex mathematical findings will have practical importance. (pp. 167–168)

Flood also highlighted psychological research on decision making, citing Ward Edwards' (1954) "The Theory of Decision Making." Flood (1955, p. 168) concludes, "I confidently hope and expect that this Institute will be of help both in unifying and extending scientific efforts toward an acceptable normative theory of decision making and in hastening effective applications to practical management problems such as those met in organizational design and in production control."

The practical potential of DA was highlighted in a 62-page paper by Maurice Allais (1957, *MS*). In that paper, Allais describes a probabilistic forecasting model used to determine "the best and economically optimal strategy to be used in prospecting for metal deposits in the Sahara" (pp. 285–286). Allais notes, "Mining exploration is per excellence a field to which methods of operations research, economic theory of risk as well as those of the games theory can be applied. Mining exploration is in fact a lottery where the tickets cost hundreds of millions and billions can be won." Allais concludes by saying,

> Whatever the precise value of the estimates may be, two arguments cannot be disputed: a) these estimates provide a precise and useful synthesis of all available informations; b) the research and the analysis required by these estimates are extremely indicative, and makes us think about problems which we otherwise would certainly overlook. (p. 319).

Of course, Allais is famous for the *Allais paradox* (Allais 1953), which challenged the expected utility model of rational choice, and he won the Nobel Prize for economics in 1988 for his contributions to the theory of markets and efficient utilization of resources. With this paper in *MS*, Allais (1957) published one of the first—*perhaps the first*—practical, large-scale applications of DA.

MS has continued to be a key outlet for research in DA over the years with a focus on theoretical and methodological research. Although Flood (1955, *MS*) and others have long distinguished between normative and descriptive perspectives on the study of decision making, Bell et al. (1988) consider a prescriptive perspective as well:

- The *normative* perspective builds on theories of rational choice developed by Savage (1954), von Neumann

and Morgenstern (1944), de Finetti (1937), and Ramsey (1926), among others. This work established expected utility theory as the dominant normative model for decision making under uncertainty.

- The *prescriptive* perspective of decision analysis is focused on helping people make decisions by implementing the normative theories and addressing the practical realities of real-world limitations on identifying and valuing outcomes, estimating or assessing probabilities, and performing the calculations necessary to obtain expected utilities (e.g., through dynamic programming or simulation).
- Finally, the *descriptive* perspective is concerned with how people actually make decisions without "rational" decision aids. Leading examples of descriptive research on decision making include Kahneman and Tversky's work identifying heuristics and biases displayed by intuitive statisticians and their development of prospect theory (Kahneman and Tversky 1979, Tversky and Kahneman 1992).

Bell et al. (1988) also note that researchers in these areas tend to have different disciplinary backgrounds—with normative researchers coming from statistics, mathematics, and economics; prescriptive researchers coming from operations research and management science; and descriptive researchers coming from psychology and other behavioral sciences.

As we discuss, a number of significant papers on the theory and applications of decision analysis were published in *MS* in the 1970s and 1980s, particularly in the areas of multiattribute utility theory and probability elicitation. Though theoretical and methodological research in DA continues to have a significant impact on practice, practical applications of DA have been rarely published in *MS* in recent years. We can illustrate this thesis with a personal example involving one of the authors. Smith published a theory/methodology paper in *MS* (Smith and Nau 1995, *MS*) that integrates option pricing techniques with decision analysis methods for valuing risky projects. Subsequent papers applying the approach in a model of developing an oil and gas field (Smith and McCardle 1998, *OR*) and describing actual applications in the oil and gas industry (Smith and McCardle 1999, *OR*) were published in *Operations Research* (*OR*).

This example is illustrative of what we perceive to be a general trend: Theory and methodological work in DA is often published in *MS*, but applications are usually published elsewhere, including other INFORMS journals (*OR*, *The INFORMS Journal on Applied Analytics*, *Decision Analysis*) or field journals (e.g., in medicine, risk analysis, or petroleum engineering). Consulting or corporate applications of decision analysis are rarely published, sometimes because of confidentiality concerns or a lack of incentives to publish or, perhaps, because of editors' and reviewers' beliefs that the methodological approaches are not sufficiently novel or innovative to merit publication in *MS*. This trend reflects the founding vision for *MS*. For example, Churchman (1994, p. 107) recalls,

> My hope was that *MS* would be quite different from *OR*, because *MS*, the journal, the meetings, and the research would be the attempt to create and design a science of management that lived up to the standards of good science, whereas *OR* would be the practical application of that science.

MS's focus on theoretical and methodological research in DA and the small number of recent papers on applications make it difficult to identify a direct impact of DA research on practice. Our belief is that the impact of this research is real and significant, but one has to "pull the thread" a bit to reveal the connections and influence of this research on practice. In tracing these connections from theory to practice, we find it helpful to consider Martin Shubik's (1987, *MS*) classification of game theories that he put forward when reflecting on the first 32 years of *MS* in a plenary address at the 1987 TIMS conference. Shubik (1987, pp. 1515–1516) distinguishes three kinds of game theories:

- *High-church* game theory focuses on mathematics, axioms, and formal solution concepts with much of it verging on "art for art's sake." Such work "moves one step closer to operating concerns but without direct sponsorship or immediate application."
- *Low-church* game theory involves "work on a specific application producing, if only for illustrative purposes, actual calculations and possibly parametric sensitivity analysis."
- *Conversational* game theory consists of "advice, suggestions and counsel as to how to think strategically." Shubik (1987) highlights the concepts of zero-sum games, non-zero-sum games, and the prisoner's dilemma as examples that are useful for framing problems in a competitive environment.

Shubik (1987) believed that low-church applications of game theory "have been of some, but nevertheless relatively modest, worth, but nowhere near the applied value of linear programming" (p. 1516). By contrast, "in terms of application and value to managers at the highest level," conversational game theory is of "considerable worth," but, he argues, "without high-church game theory, the concepts, illustrations and stories of conversational game theory would hardly exist and certainly would not have a coherent intellectual basis" (p. 1517).

DA research appearing in *MS* is typically of the high-church variety: Regardless of whether the work is normative, prescriptive, or descriptive in nature, the research is often motivated by potential applications but without a direct application described in the paper. In contrast to game theory and like linear programming,

we believe that low-church applications of DA have been of considerable value. Applications are now routine in some industries—such as the oil and gas industry (or energy industry, more broadly) and the pharmaceutical industry—and have had important contributions in many public-sector settings as well. And, like Shubik (1987), we also believe that conversational applications of DA are of considerable worth, bringing clarity to everyday discussions about decisions and the risks and trade-offs involved.

In the remainder of this paper, we discuss the connections between high-church research appearing in *MS* and low-church and conversational applications in two broad research areas: (1) multiattribute utility theory in Section 2 and (2) probability assessment in Section 3. We chose these two areas because research in these areas is well represented in *MS* and because these research areas lie at the heart of the expected utility paradigm that underlies DA. Moreover, research in these areas highlights the interplay between normative, prescriptive, and descriptive perspectives. In Section 4, we conclude with some comments on trends in DA research in *MS* and implications for applications.

2. Multiattribute Utility Theory

Many important decisions involve multiple objectives and multiple attributes. These decision problems are often associated with a hierarchy of objectives, in which some overall objective is decomposed into subobjectives, and each subobjective may be decomposed into lower-level objectives that are the basis for evaluating the alternatives. These lower-level objectives are associated with one or more attributes, which are measurable on some scale. For example, the objective of finding the "best automobile" might be decomposed into an evaluation based on cost, performance, and appearance subobjectives. The cost subobjective might be evaluated on attributes related to the purchase price and maintenance costs.

2.1. Early Research on Multiattribute Utility Theory in *MS*

MS has been a leading outlet for research and applications of multiattribute utility. Early research in multiattribute utility theory in *MS* was motivated by both low-church practical concerns and high-church theoretical interests. On the practical side, in the 1950s and 1960s, there was great interest in *systems analysis*, popularized and championed by U.S. Secretary of Defense Robert McNamara. For example, Black (1967, *MS*) describes the adoption of tools from systems engineering, decision theory, and optimization to evaluate government programs and initiatives. A major concern that Black (1967, *MS*, p. B-51) highlights is the need to quantify program benefits that may include multiple dimensions: "A benefit function is the embodiment of the structure of benefits, and the manner in which one can be substituted for another without changing the combined benefit from all system outputs. Great care must be exercised that various benefits are incorporated into the function in order to obtain a true measure of the extent to which the complex of objectives is achieved."

An important goal of research in multiattribute utility and value theory is to identify independence conditions for preferences over multiple attributes that allow the utility function (for settings with uncertainty) or value function (for settings without uncertainty) to be decomposed into forms that are easier to assess in practical applications, such as additive or multiplicative forms. Fishburn (1968, *MS*) provides an early review of the foundations of utility theory, spanning 44 pages in *MS*. There Fishburn describes an independence condition—now called *mutual preferential independence*—that must be satisfied to justify the existence of an additive value function for the case with no uncertainty: the condition requires the preferences for any subset of the attributes, say x_1 and x_2, given the others $(x_3, \ldots, x_n)$, not to depend on the specific values for $(x_3, \ldots, x_n)$. This additive representation result is based on algebraic developments by Scott and Suppes (1958), on the topological results of Debreu (1960), and on the work of Luce and Tukey (1964) and Krantz (1964) and others in mathematical psychology. Krantz et al. (1971) provides a summary of this and related work.

Fishburn (1968, *MS*) also describes conditions leading to an additive representation for multiattribute utility functions under uncertainty. Here the independence condition—from Fishburn (1965, *OR*; 1966; 1967a, b, *OR*)—is called *additive independence* and requires preferences for multiattribute alternatives with uncertain outcomes to depend only on the marginal distributions for each attribute. The advantage of the additive representation is its simplicity. The assessment of the n-dimensional multiattribute utility function $u(x_1, x_2, \ldots, x_n)$ is reduced to the assessment of n one-dimensional utility functions $u_i(x_i)$, one for each attribute x_i, and the associated scaling constants k_i:

$$u(x_1, x_2, \ldots, x_n) = k_1 u_1(x_1) + k_2 u_2(x_2) + \cdots + k_n u_n(x_n), \tag{1}$$

where the $u_i(\cdot)$ are scaled to range from zero to one for the worst to best outcomes for the given attribute. Fishburn (1967c, *MS*) reviews 24 methods for estimating such additive preference functions for settings with certainty and uncertainty.

The disadvantage of this additive representation (1) is the restrictiveness of the necessary conditions: in

many cases, we would expect preferences to depend on the joint probability distribution over attributes. Keeney (1972, *MS*) studies this more general case and provides assumptions about decision maker (DM) preferences for multiattribute consequences that still greatly simplify the assessment of multiattribute utility functions. Here the key assumption is a *utility independence* condition that is satisfied if the DM's preference for lotteries involving two alternatives that differ on any one of the attributes is not affected by common values of the other attributes. In this case, Keeney (1972) shows that a multiattribute utility function can be determined by assessing conditional utility functions $u_i(x_i)$ for each attribute, holding the other attributes constant. The resulting multiattribute utility function can then be written in a *multilinear* form, that is, as the sum of these conditional utility functions plus the sum of all combinations of cross-product terms involving these same conditional utility functions, subject to the appropriate scaling. For example, with three attributes, the multilinear form is

$$\begin{aligned} u(x_1, x_2, x_3) = {} & k_0 + k_1 u_1(x_1) + k_2 u_2(x_2) + k_3 u_3(x_3) \\ & + k_4 u_1(x_1) u_2(x_2) + k_5 u_1(x_1) u_3(x_3) \\ & + k_6 u_2(x_2) u_3(x_3) + k_7 u_1(x_1) u_2(x_2) u_3(x_3), \end{aligned} \tag{2}$$

where the k_i are scaling constants. Note that both the additive and multiplicative forms of a multiattribute utility function are special cases of (2). Later, Bell (1979, *MS*) describes techniques for assessing multiattribute utility functions that cannot be decomposed into additive or multilinear combinations of univariate utility functions.

Keeney and Raiffa's (1976) classic book, *Decisions with Multiple Objectives*, provides a rigorous summary of multiattribute utility theory that is supported by examples of assessment techniques associated with the different models and with discussions of illustrative real-world applications. This book served—and still serves—to make the high-church research on multiattribute utility theory more accessible and provided the foundation for the many applications that would follow.

2.2. Applications of Multiattribute Utility Theory in *MS*

After publication of the book by Keeney and Raiffa (1976), the focus of papers related to multiattribute utility theory in *MS* shifted to applications of the methodology. Many of these applications address problems in the public sector, continuing the focus on systems analysis from the 1960s. For example, Bodily (1978, *MS*) describes a model to allocate police mobile units in an urban area. The allocations are based on a multiattribute utility model involving two attributes—efficiency and equality of service—that are assumed to be mutually utility independent. The weights on the corresponding utility terms are determined by consulting representatives of interest groups, including citizens, police, and city administrators.

Golabi et al. (1981, *MS*) consider the problem of selecting a portfolio of solar energy experiments, assisting the U.S. Department of Energy. In the model, the technical quality of each experiment is evaluated using a multiattribute utility function. Then the overall value of a portfolio is estimated using a measurable value function defined on these levels of technical quality. The measurable multiattribute value function (Dyer and Sarin 1979, *OR*) used here was an additive form—like Equation (1)—but involves comparisons of differences in levels that an attribute achieves when an experiment is funded versus when it is not funded. The portfolio selection problem is formulated as an integer programming problem with budgetary and programmatic issues represented as constraints.

There have been a number of other public-sector applications of multiattribute utility theory published in *MS*. For example, Crawford et al. (1978, *MS*) describe the evaluation of alternative designs for a transmission line to be installed in the lower peninsula of Michigan; they consider four attributes representing three different cost estimates and another representing the noise associated with the system. Ford et al. (1979, *MS*) evaluate methodologies for choosing locations for nuclear power plants. Keeney et al. (1990, *MS*) describe the elicitation of information from members of the West German public for the evaluation of long-term energy strategies. Gregory and Keeney (1994, *MS*) consider a decision to locate a coal mine in a sensitive environmental area in Malaysia. Parnell et al. (1998, *MS*) use an additive value function to evaluate future air and space forces with attributes and weights specified by 200 military experts; this model includes 134 attributes. In a more recent paper, Grushka-Cockayne et al. (2008, *MS*) describe a major study of the European Air Traffic Management System that involves complex alternatives, multiple stakeholders, and multiple attributes.

2.3. Later Research on Multiattribute Utility Theory in *MS*

The interest in applications of multiattribute utility theory stimulated research on methods to improve the assessment of utility functions. A major focus of applications-related research appearing in *MS* was on the estimation of the utility weights or scaling constants assessed for each attribute, the k_i in Equation (1). For example, Weber et al. (1988, *MS*) identify the so-called splitting effect: "When objectives are split into detailed attributes, the sum of the attribute weights is typically larger than the weight directly attached

to the objective" (p. 432–433). This finding suggests that the structure of the objectives hierarchy may bias the assessed multiattribute utility function. In a related study, von Nitzsch and Weber (1993, *MS*) find that subjects do not consistently adjust the weights that they assign to attributes in response to changes in the ranges over which those attributes are measured: as the range increases for measuring an attribute, the weight should increase also. Schoemaker and Waid (1982, *MS*), Stillwell et al. (1987, *MS*), and Borcherding et al. (1991, *MS*) also evaluate different methods for assessing attribute weights.

Another important issue related to multiattribute utility theory is the use of *proxy* attributes rather than the fundamental objectives about which the DM cares. For example, a DM may wish to evaluate alternative medical emergency response systems based on a fundamental objective of increasing the number of lives saved. However, the number of lives saved may be difficult to relate directly to the proposed emergency response system alternatives, so the DM may use proxy attributes, such as the speed of response and the quality of medical equipment, instead. Fischer et al. (1987, *MS*) find that subjects tend to overestimate the weights on proxy attributes and suggest that efforts to use fundamental objectives to characterize outcomes should be encouraged whenever possible.

The interest of decision analysts in problems with multiple objectives is also shared by scholars associated with the International Society on Multiple Criteria Decision Making (MCDM), which was established in 1978. The MCDM society has continued to thrive with its own conferences and journal, the *Journal of Multi-Criteria Decision Analysis*, which was first published in July 1992. Recognizing the overlap between the interests of many members of the MCDM society and INFORMS, the INFORMS section on MCDM was established to enhance the visibility and participation of MCDM society members in INFORMS. A recent paper that surveys the state of the art in this field and its connections to DA is provided by Wallenius et al. (2008, *MS*).

2.4. Research on Single-Attribute Utility Theory in *MS*

Though our focus is on multiattribute utility theory, papers published in *MS* have also played an important role in developing methods for assessing single-attribute utility functions. These methods are useful when there is a single objective in the problem (e.g., money) and also when assessing conditional utility functions as part of an additive or multilinear utility function, as discussed in Section 2.1. Farquhar (1984, *MS*) provides an early comprehensive review of many different assessment techniques, focusing on potential biases. Hershey et al. (1982, *MS*) and present experimental evidence that different assessment methods (e.g., varying the form of the questions) and context and framing effects "make it difficult to speak of *the* utility function for a given person" (Hershey and Schoemaker 1982, *MS*, p. 936). The biases observed in this work are generally consistent with the predictions of prospect theory (Kahneman and Tversky 1979, Tversky and Kahneman 1992). Subsequent research in this vein, including McCord and de Neufville (1986, *MS*), Johnson and Schkade (1989, *MS*), and Wakker and Deneffe (1996, *MS*), further studies these biases and introduces methods to reduce their effects. Bleichrodt et al. (2001, *MS*), by contrast, use the biases predicted by prospect theory to correct assessed utilities.

In addition to this work relating utility assessment to the biases predicted by prospect theory, there is a burgeoning literature in *MS* on estimating the prospect theory model. This work includes Wu and Gonzales (1996, *MS*; 1999, *MS*), who study the shape of the probability weighting function in prospect theory; Kilka and Weber (2001, *MS*), who study source dependence in the probability weighting function; and numerous papers by Abdellaoui and various coauthors (e.g., Abdellaoui 2000, *MS*; Abdellaoui et al. 2005, *MS*; 2007, *MS*; 2011, *MS*; Abdellaoui and Kemel 2014, *MS*), who study a number of assessment issues associated with the prospect theory model. Baucells et al. (2011, *MS*) consider the formation of reference points, and Murphy and ten Brincke (2018, *MS*) discuss methods for improving the reliability of individual parameter estimates for the prospect theory model.

2.5. Applications of Multiattribute Utility Theory in Medical Decision Making

Multiattribute utility theory has greatly influenced the study of medical decision making. A central issue in medical decision making is finding a metric that allows comparisons of medical outcomes associated with different health policies or medical treatments. Early work in *MS* included Stimson (1969, *MS*), who reviews the use of a multiattribute value model based on the intuitively appealing approach of Churchman and Ackoff (1954, *OR*), and Torrance (1976, *MS*), who provides a unified review of different health status indices. Pliskin and Beck (1976, *MS*) used preferential independence concepts—such as those discussed in Section 2.1—to develop an additive-value function for prioritizing patients with chronic renal failure for treatment with dialysis or a kidney transplant.

A seminal contribution of this early work on utility models of health status indices is the development of the *quality-adjusted life-year* (QALY) measure, which has become a widely used measure of health

improvement to guide healthcare resource-allocation decisions. The U.S. Panel on Cost-Effectiveness in Health and Medicine (Gold et al. 1996) and the National Institute of Health and Clinical Excellence in Britain have endorsed the use of QALYs for assessing the cost-effectiveness of different healthcare interventions. Pliskin et al. (1980, *OR*) is widely credited with providing a preference theory foundation for the QALY measure. Pliskin et al. (1980, *OR*) focus on the case with a health state q over a lifespan of y years and show that if preferences satisfy the following assumptions:

1. Utility independence, as developed in Keeney (1972, *MS*) and discussed in Section 2.1, between life years y and health status q;

2. Constant proportional trade-offs, which mean that the proportion of remaining life years one is willing to give up for the same improvement in health status does not depend on the number of life years remaining y; and

3. Risk neutrality for life years for a given health state q, then the utility function must have the form

$$U(y, q) = yH(q), \tag{3}$$

where $H(q)$ is a utility function for the quality of life associated with health state q. Thus, Equation (3) can be interpreted as quality-adjusted life years. Miyamoto et al. (1998, *MS*) provide a simpler set of assumptions that lead to this same form: in their analysis, assumptions (1) and (2) are replaced by a *zero condition* that says that for a duration of zero life years, all quality-of-life levels are equivalent.

MS has continued to publish work on the theory of QALYs (e.g., Hazen 2000, *MS*; Smith and Keeney 2005, *MS*) and occasionally publishes research related to medical decision making using QALYs as a measure of effectiveness. For example, Zaric et al. (2000, *MS*) develop a dynamic model of the spread of human immunodeficiency virus (HIV) to evaluate a methadone maintenance program based on its cost-effectiveness. Zenios (2002, *MS*) and Su and Zenios (2006, *MS*) evaluate the effectiveness of kidney-exchange programs using the QALY framework. More recently, Chan et al. (2016, *MS*) use QALYs to analyze the optimal deployment of public-access defibrillators in Toronto, and Ayer et al. (2016, *MS*) use a partially observable Markov decision process to evaluate optimal breast-screening policies, taking into account the fact that many women do not adhere to the recommended screening guidelines.

However, most applications of the QALY model are published in medical journals. According to Google Scholar, from 2009 to 2019, there have been about 27,000 papers that mention "quality-adjusted life years." For example, a perspective piece published in the *New England Journal of Medicine* (Neumann et al. 2014) discusses the origins of the popular benchmark of $50,000-per-QALY gained as a standard for cost-effectiveness and argues that in the United States, one should now use a benchmark of $100,000 or $150,000 instead. Although there is some debate around the use of QALYs in cost-effectiveness analyses, there is no debate that high-church research on multiattribute utility is having a tremendous impact in the healthcare arena.

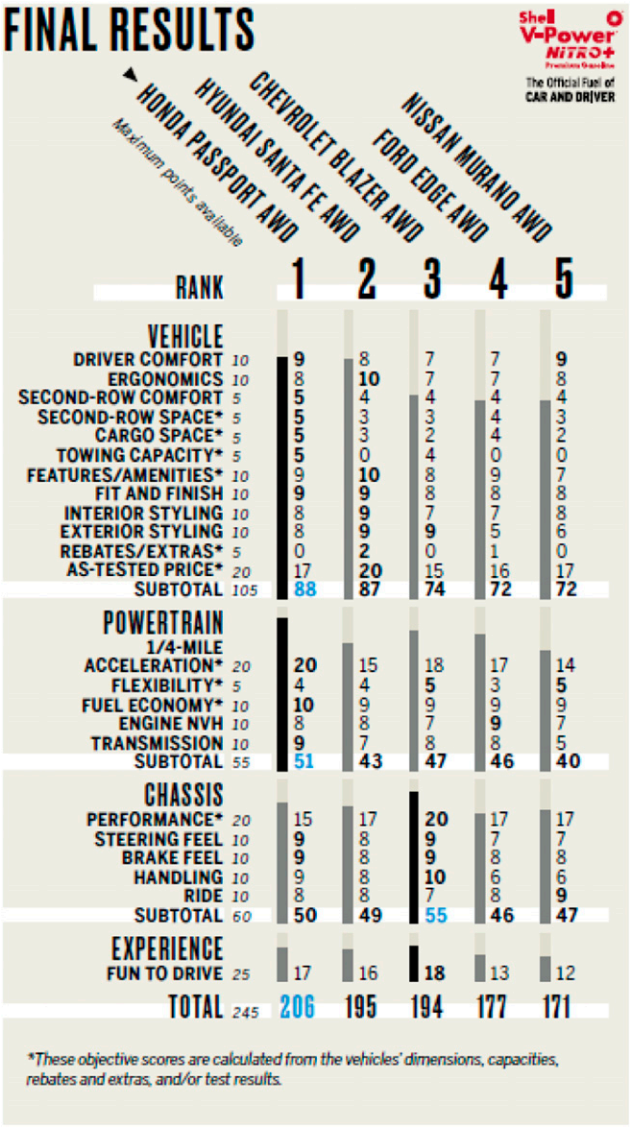
The Society for Medical Decision Making (SMDM, https://smdm.org) was formed in 1979 to foster the application of DA models and methods to medical decision making. SMDM "is the leading society for studying and advancing decision sciences in health, including incorporation of patients' values and preferences" (SMDM 2020). The society currently has more than 1,000 members and publishes its own journal, *Medical Decision Making*. One of the goals of the SMDM is to develop curricula and to provide training to people who are involved in medical decision making and in medical policy analysis.

2.6. Multiattribute Product Rankings

Many published applications of multiattribute utility theory involve major resource-allocation decisions that may cost millions of dollars and may involve multiple stakeholders. These applications are often implemented with the support of trained decision analysts who assist with the tasks of identifying objective hierarchies and attributes and with the evaluation of complex alternatives. Considerable time may be devoted to communication between analysts, DMs, and other stakeholders to ensure that the problem is framed correctly. These applications are typically carried out with care and with attention to research on assessing multiattribute utility functions.

However, the additive-value functions that appear in many major resource-allocation applications have a simple rate-and-weight form that also appears in many conversational applications. For example, consider the evaluation of five small sport utility vehicles (SUVs) in *Car and Driver* magazine, as shown in Figure 1 (Jacquot 2019). Here we see a hierarchy of objectives with four high-level objectives: vehicle (with comfort and styling as well as price as attributes), powertrain (including acceleration and fuel-economy attributes), chassis (with performance and handling attributes), and experience (fun to drive is the sole attribute). Each attribute is assessed on a numerical scale with a range from zero to the maximum possible points for the attribute, which implicitly reflects the scaling constants or weights associated with the attributes. The total score for each SUV is simply the sum of the individual scores.

Figure 1. (Color online) Ranking of Small Sport Utility Vehicles from *Car and Driver*



This rating system can be interpreted as an additive-value function, and its appropriateness rests on the mutual preferential independence condition discussed in Section 2.1. Here the preferential independence assumptions mean that, for example, the DM's trade-offs between "as-tested price" and "performance" would not be affected by changes in common values of styling or other attributes. Of course, even if assessed properly, the objectives, attributes, ratings, and weights here reflect the preferences of the staff of *Car and Driver*, a magazine that appeals to automobile enthusiasts who would likely place large weights on attributes that relate to handling, performance, and being fun to drive. In contrast, an evaluation of these cars in *Consumer Reports* may place more weight on the objectives of reliability and safety; individuals whose preferences are more aligned with these objectives might find *Consumer Reports*' rankings to be more helpful. Alternatively, one could use a personalized recommendation system, such as MyProductAdvisor.com, to specify preferences for car styles, brands, and other attributes to develop one's own ranking.

We believe that these kinds of product evaluation guides can add significant value to major consumer purchase decisions. As discussed in Bond et al. (2008, *MS*), these tools can help make consumers aware of important product features. In their research, Bond et al. (2008, *MS*, p. 56) found that individuals asked to list their objectives "consistently omitted nearly half of the objectives that they later identified as personally relevant." However, one should be clear about the assumptions underlying the recommendations and the interpretation of the weights in the additive model. For example, the weights should be influenced by the ranges over which the attributes are rated (von Nitzsch and Weber 1993, *MS*). Similarly, we should be aware of the splitting effect (Weber et al. 1988, *MS*) that suggests rating systems that decompose attributes to different degrees may lead users to place more or less weight on different objectives. For example, even if consumers were to assign their own weights to attributes, *Car and Driver*'s choice of attributes may nudge consumers to place more weight on performance attributes and *Consumer Reports* may nudge consumers to put more weight on safety and reliability.

In addition to multiattribute tools being useful for major purchase decisions, we could imagine similar tools being used for repeated everyday decisions, such as the choice of a restaurant, hotel, or movie. However, such multiattribute, utility-based recommendation systems are currently not common. One of the authors (Dyer) surveyed the product-recommendation websites that were available to consumers in 2008 and identified several multiattribute, utility-based systems (Butler et al. 2008). Of these, MyProductAdvisor.com is the only one that is currently still active. Meanwhile, collaborative filtering-based recommendation systems have become quite common. These systems predict the interests of a user by collecting preference information or evaluations from many users and identifying those with similar tastes. Nevertheless, for unique, high-cost purchases (such as cars) or repeated purchases (e.g., restaurants, hotels, or movies), we would argue that there is a role for recommender systems that explicitly assess consumers' preferences to generate recommendations. We hope to see more such systems in the future.

3. Research and Practice in Probability Assessment

To apply the expected utility framework, one needs probabilities. The decision analysis perspective is

closely aligned with that of Bayesian statistics in that the probabilities are taken to represent the beliefs of the DM. Savage's axioms (Savage 1954) imply that a DM has probabilities, and de Finetti (1937) shows that if a DM does not accept bets that result in a certain loss (i.e., is *coherent*), then the DM has probabilities. Other axiomatizations of decision theory (e.g., von Neumann and Morgenstern 1944) simply assume that probabilities exist.

In practice, of course, these probabilities must somehow be assessed. In Allais's (1957, *MS*) early study of mining in the Sahara, these probabilities reflect expert judgment, with the assessments for the Sahara being informed by data from other regions. For example, Allais (1957, *MS*, p. 298) writes that probabilities of success may be "estimated from the information available to those who have had long experience in the field of mining exploration." This approach remains standard practice in decision analysis applications today.

3.1. Assessment and Calibration Research in *MS*

Early research on probability assessment began in the 1960s (see, e.g., Winkler 1967) and has been a central research theme in *MS* since that time. A key paper on probability assessment is that by Spetzler and Stael von Holstein (1975, *MS*). Spetzler and Stael von Holstein (1975, *MS*) were affiliated with the Decision Analysis Group at the Stanford Research Institute, which focused on consulting applications of decision analysis. Their 1975 paper represents a summary of best practices for probability assessment in applications based on their consulting experience. In this paper, Spetzler and Stael von Holstein emphasize a number of things that are often overlooked or taken for granted in academic studies on probability assessment: for example, the choice of uncertainties to assess (model more or assess directly?) and the need to motivate and establish rapport with the subjects. Before assessing a probability, Spetzler and Stael von Holstein (1975, *MS*) suggest asking experts to generate several scenarios that may lead to the occurrence or nonoccurrence of the event in question or a high or low value for an uncertain quantity. The authors also recommend assessing probabilities using a probability wheel and asking subjects if they would rather bet on the spinner landing on orange or the event in question happening. This focus on betting when assessing probabilities reflects high-church research in decision theory. For example, Ramsey (1926) and de Finetti (1937) define probabilities in terms of bets, and Anscombe and Aumann (1963) give a definition of subjective probabilities that include "roulette lotteries" that serve as reference gambles, like the probability wheel recommended by Spetzler and Stael von Holstein (1975, *MS*).

Wallsten and Budescu (1983, *MS*) provide an early review of the literature on probability assessment from a psychological perspective. A key issue in this early work is calibration. To assess calibration, one looks at the set of events to which a subject assigns probability p; if the subject is well calibrated, these events should actually occur with frequency p. For example, if you look at the cases in which a subject says $p = 0.80$ (or 0.20), the subject is well calibrated if these events occur 80% (20%) of the time. The subject is overconfident if the events occur, say, 60% (or 40%) of the time. Intuitively, overconfidence means that subjects overestimate the degree to which they "know" what is true. Early research showed that meteorologists are very well calibrated when providing probability-of-precipitation forecasts (e.g., Murphy and Winkler 1977). However, many other experiments show that people are generally quite poorly calibrated (e.g., Lichtenstein et al. 1982). After reviewing the literature involving experts (mostly meteorologists and physicians) and nonexperts (e.g., experimental subjects), Wallsten and Budescu (1983, *MS*, p. 166) conclude that "when encoding subjective probabilities about events with which they are familiar, experts can be exceedingly well-calibrated, whereas a similar degree of goodness has rarely been demonstrated by nonexperts in laboratory contexts."

Calibration research continues to be prominent in *MS*. For example, Tannenbaum et al. (2017, *MS*) study how the epistemic versus aleatory nature of uncertainty affects the extremity and calibration of forecasts. An *epistemic* uncertainty is one that is knowable in principle (such as the answer to a trivia question), whereas an *aleatory* uncertainty is inherently unpredictable (like a coin flip); of course, there is a range of "epistemicness" between these two extreme examples. In a series of experiments, Tannenbaum et al. (2017, *MS*) find that subjects tend to assign more extreme probabilities (e.g., closer to zero or one) and be more poorly calibrated when considering events that are more epistemic in nature. As a simple demonstration of the phenomenon, consider two different ways to assess the probability that a team wins a basketball game:

- *Singular presentation:* The Chicago Bulls play the Detroit Pistons on March 21. What is the probability that the Bulls win?
- *Distributional presentation:* The Chicago Bulls play the Detroit Pistons on February 20, March 21, and April 3. What is the probability that the Bulls win on March 21?

Of course, these are two ways of asking the same question, but the distributional presentation encourages subjects to think about a series of games and adopt a more aleatory perspective, that is, to recall that good basketball teams lose some games. Tannenbaum et al. (2017, *MS*)

find that with the singular presentation, the mean response to the question was 0.72 versus 0.63 for the distributional presentation. In related work, Walters et al. (2017, *MS*) show that overconfidence is often driven by the neglect of unknowns and that subjects who explicitly consider the unknowns (spontaneously or with prompting) are much better calibrated than those who do not.

The work of Tannenbaum et al. (2017, *MS*) and Walters et al. (2017, *MS*) is also related to what Kahneman and Lovallo (1993, *MS*, p. 25) call the "inside" and "outside" views for assessing probabilities. In the inside view, a forecast is generated by focusing on the "case at hand, by considering the plan, and the obstacles to its completion, by constructing scenarios of future progress, and by extrapolating current trends." In the outside view, one "essentially ignores the details of the case at hand" and focuses on "the statistics of a class of cases chosen to be similar in relevant respects to the present one." Expertise can be applied in both modes of thought, for example, by using expertise to elaborate on a particular scenario or, alternatively, to think of many possible future scenarios. Kahneman and Lovallo (1993, *MS*) tell a story of a curriculum expert whose forecasts about project completion times changed dramatically when he was forced to adopt an outside rather than an inside view. Kahneman and Lovallo (1993, *MS*, p. 25) say that "it should be obvious that when both methods are applied with equal intelligence and skill, the outside view is much more likely to yield a realistic estimate." But Kahneman and Lovallo (1993, *MS*, p. 26) caution that

> the inside view is overwhelmingly preferred in intuitive forecasting. The natural way to think about a problem is to bring to bear all one knows about it, with special attention to its unique features. The intellectual detour into the statistics of related cases is seldom chosen spontaneously. Indeed, the relevance of the outside view is sometimes explicitly denied: physicians and lawyers often argue against the application of statistical reasoning to particular cases. In these instances, the preference for the inside view almost bears a moral character. The inside view is valued as a serious attempt to come to grips with the complexities of the unique case at hand, and the outside view is rejected for relying on crude analogy from superficially similar instances. This attitude can be costly in the coin of predictive accuracy.

Although we have focused on calibration research in our brief discussion here, there are equally important streams of research on scoring rules and combining probabilistic forecasts that have been featured in *MS* over the years. Calibration is important because we want to be able to take probabilities at face value and know that probabilities should reflect actual frequencies. But calibration is clearly not everything: a weather forecaster who says that there is a 30% chance of rain *every day* may be well calibrated if it, in fact, rains 30% of the time, but such forecasts are not very helpful. A *scoring rule* is a measure that considers the agreement between a probability forecast and the outcome of the predicted event. In an ex ante sense, a strictly proper scoring rule provides an incentive for honest forecasting by the forecaster or forecasting system. In an ex post sense, a scoring rule rewards accurate forecasts. The literature on evaluating forecasts using scoring rules in *MS* includes Matheson and Winkler (1976, *MS*), Winkler (1994, *MS*), Lichtendahl et al. (2013, *MS*), Jose et al. (2013, *MS*), Grushka-Cockayne et al. (2017, *MS*), and Regnier (2018, *MS*). The role of *MS* in the early literature on combining probabilistic forecasts is discussed in Smith and von Winterfeldt (2004, *MS*); Winkler et al. (2019) provides a recent discussion of research on combining forecasts.

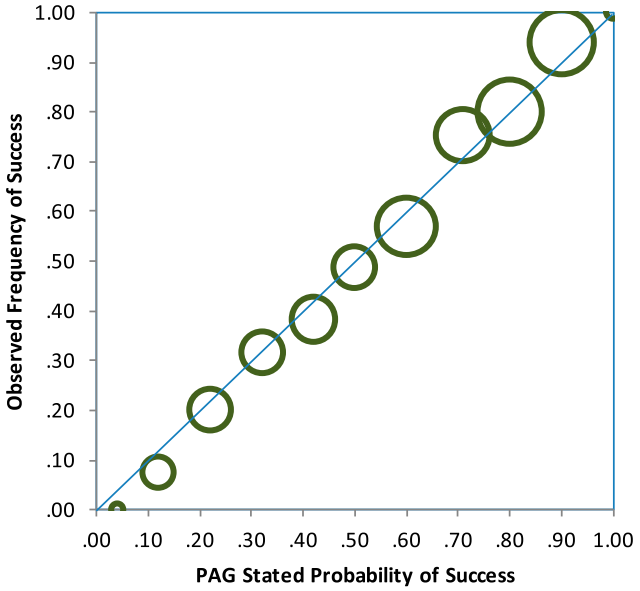
In the remainder of this section, we briefly review three recent applications of probability assessment and consider the calibration of these assessments.

3.2. Research and Development Probability Forecasting at Eli Lilly

Eli Lilly and Company is a research-based pharmaceutical company located in Indianapolis. Since 1997, Lilly has had an independent review board with 10–15 members—called the *Portfolio Analysts Group* (PAG)—that assesses the probability of success for most of Lilly's research and development (R&D) projects. (Most people at Lilly think that PAG stands for "Probability Assessment Group" because that is its main function.) The PAG consists of people with deep expertise in drug development and is responsible for providing assessments for Lilly's entire R&D portfolio and for all stages of the drug-development process from the preclinical stage through the three phases of clinical research and registration success (i.e., approval by the U.S. Food and Drug Administration). The assessment process is led by a facilitator, and most PAG members have received training on probability assessment (e.g., emphasizing how people have a tendency to be overconfident). When there are differences in probabilities across members of the PAG, the official PAG forecast is taken to be the average of the individual probabilities. Lilly uses these probabilities to set expectations and prioritize its R&D investments.

How does the PAG do? Jay Andersen and Charles Persinger at Eli Lilly have done a retrospective study comparing these probability assessments with actual project outcomes for 1,274 PAG estimates assessed over the years 1997–2019 (Persinger 2019).[2] Figure 2 shows a calibration plot summarizing their findings.

Figure 2. (Color online) Calibration Plot for the Eli Lilly PAG Forecasts



In this calibration plot, the x-coordinates represent a probability of success stated by the PAG, and the y-coordinates represent the actual success rate (meaning the fraction of projects that actually succeeded) when the PAG gave the stated probability. Nearby stated probabilities are grouped into bins, with one bin for each decile. In the plot, the x-coordinates represent the average of the stated probabilities in a given bin. The sizes of the circles in the plot are proportional to the number of forecasts in the bin. If the PAG forecasts were perfectly calibrated, the centers of the circles would all fall on the 45-degree line.

In Figure 2, we see that the PAG's forecasts are remarkably well calibrated. As discussed earlier, Wallsten and Budescu (1983, *MS*) note that experts can be exceedingly well calibrated. Here we think it is essential that Lilly's experts take this task seriously (Eli Lilly really is placing bets in accordance with these probabilities) and that the experts are experienced and trained as probability assessors (i.e., aware of overconfidence biases). In the assessment process, the PAG is explicitly reminded of benchmark data (e.g., the base rate of success for phase II clinical trials) and then adjusts for the specifics of a given case; thus, the experts are prompted to integrate the inside and outside views of the problem. Moreover, the fact that they are averaging forecasts across a number of experts is consistent with best practices, reflected in the growing literature on the "wisdom of the crowd" (see, e.g., Winkler et al. 2019 for more discussion). Andersen and Persinger's data (Persinger 2019) also show that the PAG assessments significantly improve on forecasts based on the benchmark data alone.

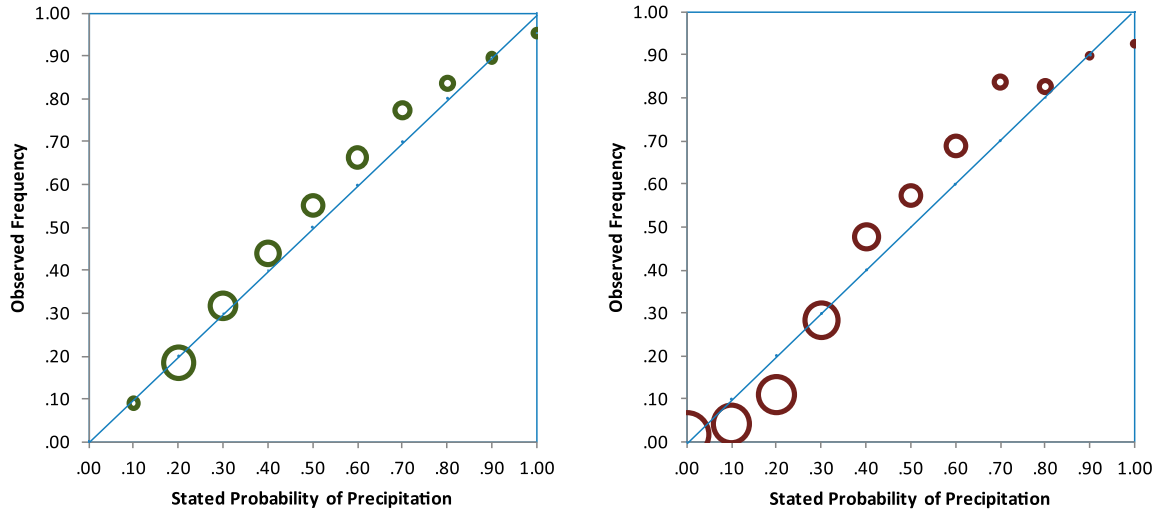
3.3. Probability-of-Precipitation Forecasts at the National Weather Service and The Weather Channel

As discussed in Section 3.1, professional weather forecasters have long provided probability-of-precipitation (PoP) forecasts and have been held up as a positive example of well-calibrated probability assessors (e.g., Murphy and Winkler 1977). A more recent study by Bickel et al. (2011) compared the PoP forecasts provided by National Weather Service (NWS) forecasters with those provided by The Weather Channel (TWC). TWC's PoP forecasts are widely distributed on cable television and on the internet, including many popular cell phone applications (e.g., the iPhone's Weather app). Though Bickel et al. (2011) considered several different data sets, we focus on the day-ahead PoP forecast for the "warm season" (April–September) in 2009 and 2010, combining the results across all regions in the United States. There are about 250,000 forecasts observed, spanning 365 days and 734 different locations.

The left panel of Figure 3 shows a calibration plot for NWS forecasters in this time frame. Here we see that the probability forecasts are quite well calibrated. Note that there are few PoPs of 0.10 and no PoPs less than 0.10: This is because the NWS issues PoP forecasts as part of storm advisories and does not typically issue storm advisories when storms are unlikely. These forecasts appear to be very well calibrated, with the possible exception of the probabilities in the 50%–70% range slightly underestimating the actual frequency of precipitation.

The right panel of Figure 3 shows a calibration plot for TWC forecasts over this same time period. Here we see that the probabilities in the 10%–20% range significantly understate the actual frequency of precipitation: When TWC says that there is a 10% chance of rain, it actually rains only 4% of the time. In an earlier study by Bickel and Kim (2008), this underestimate is even worse: When TWC said that there was a 20% chance of rain, it actually rained only 5% of the time. In the weather forecasting business, this is referred to as the *wet bias*. And perhaps this is no surprise, but TWC does this on purpose: Silver (2012, p. 135) quotes Bruce Rose, a former TWC vice president, as saying, "If the forecast was objective, if it has zero bias in precipitation, we'd probably be in trouble." Thus, it is not that TWC lacks the expertise to provide well-calibrated probability forecasts; it lacks incentives to do so. Here the high-church concern about having proper incentives—for example, subjects being willing to bet in accordance with their stated probabilities—leads to poor calibration because of conversational concerns, namely the concern that people will be mad at TWC if it rains unexpectedly.

Figure 3. (Color online) Calibration Plots for NWS Forecasters (Left) and TWC Forecasters (Right)



Source. Data from Bickel et al. (2011).

Whereas the Eli Lilly forecasts typify subjective probability forecasts provided by human experts, the weather forecasters rely heavily on computer-generated forecasts. However, a lot of human expertise is still involved in refining and adjusting the computer forecasts for known model weaknesses and peculiarities of certain locations. Silver (2012, p. 125) quotes a veteran meteorologist, Jim Hoke, as saying, "The best forecasters need to think visually and abstractly while at the same time being able to sort through the abundance of information the computer provides them with. Moreover, they must understand the dynamic and nonlinear nature of the system they are trying to study. It is not an easy task, requiring vigorous use of both the left and right brain." Or, as Kahneman and Lovallo (1993, *MS*) might say, it requires reconciling the inside and outside perspectives on the problem. Silver (2012) cites NWS statistics showing that the meteorologists improve the accuracy of the forecasts by 25% over computer-generated forecasts and notes that this improvement has been relatively constant over time even as the computer models have improved.

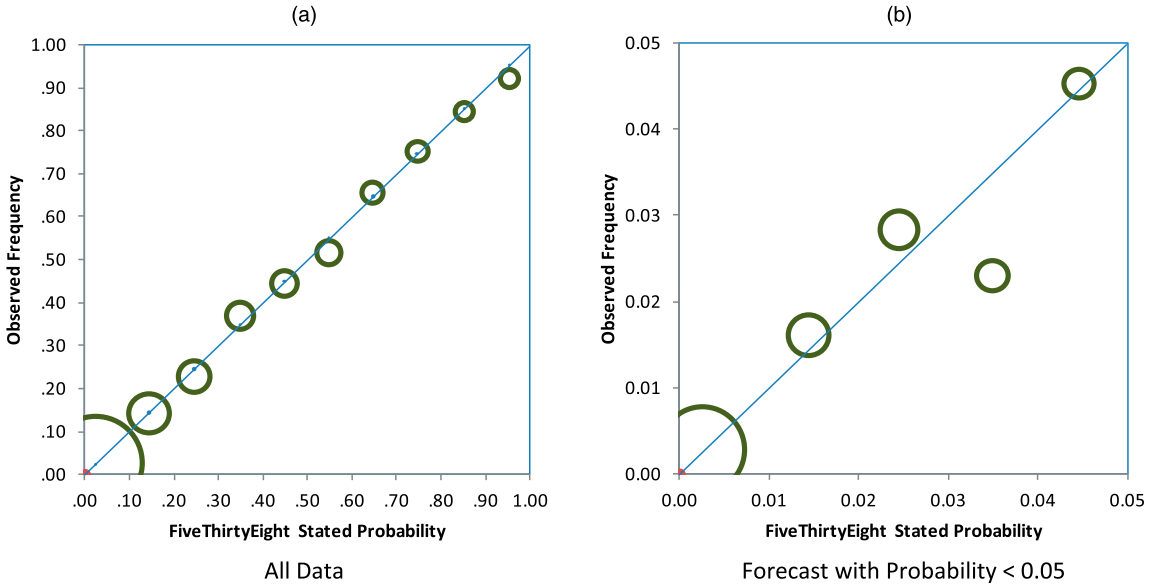
3.4. FiveThirtyEight Probability Forecasts

Although the use of PoP forecasts has a long history, probability forecasts are a more recent phenomenon in sports and politics. FiveThirtyEight is a leading purveyor of such forecasts. Founded in 2008 by Nate Silver as a website for aggregating political polls, FiveThirtyEight now publishes probabilistic forecasts for elections and sporting events. The forecasts are based on models that integrate polling data for election results and various team and player ratings for sporting events. FiveThirtyEight then uses simulation models to convert these poll results and team and player ratings into probabilistic forecasts, for example, tracking how likely a team is to make the playoffs or win a championship or how likely a U.S. presidential candidate is to come out ahead in the Electoral College.

How good are FiveThirtyEight's forecasts? In April 2019, FiveThirtyEight published a retrospective analysis considering the calibration of its forecasts dating back to 2008 (FiveThirtyEight 2019). Although there is not enough data to carefully study the calibration of many of FiveThirtyEight's election forecasts (e.g., for presidential elections), there is a significant amount of data for sporting events. As an example, Figure 4(a) shows a calibration plot for predictions associated with the men's National Collegiate Athletic Association (NCAA) basketball tournament for the years 2014–2019. For each of the 68 teams in the tournament, FiveThirtyEight estimated the probability of the team reaching each round of the tournament; there are a total of 11,133 forecasts in this data set. Note that many of these probabilities are quite small because teams with low seeds are unlikely to advance into the later rounds of the tournament; the median probability forecast is 0.057. However, highly ranked teams usually have a high probability of advancing beyond the first round. In the calibration plot of Figure 4(a), we see that FiveThirtyEight's forecasts are very well calibrated. In Figure 4(b), we focus on probability forecasts that are less than 5%—representing 48% of the 11,133 forecasts in the data set—and again see that these forecasts are also well calibrated.

As with the weather forecasts, FiveThirtyEight's forecasts are based on models. FiveThirtyEight's model for the NCAA tournament combines six different computer ratings and two human rankings—NCAA seedings and *pre*season polls—and adjusts for injuries

Figure 4. (Color online) Calibration Plots for FiveThirtyEight's Forecasts for the NCAA Men's Basketball Tournament, 2014–2019



Notes. (a) All data. (b) Forecast with probability < 0.05.

and travel time, among other things. The result is a power rating in which the difference in power ratings between teams is a forecast of the point difference if these two teams play. These forecasted point differences are translated to win probabilities by assuming that the actual point differential follows a logistic distribution with mean equal to the forecast and a standard deviation (~10.3) that was chosen to ensure good calibration. As with the weather forecasts, these probabilistic forecasts thus represent a mixture of models and human judgment integrated in a way that generates well-calibrated probability forecasts.

4. Conclusions

As we said in the Introduction, DA research appearing in *MS* in recent years has typically been of the high-church variety regardless of whether the work is normative, prescriptive, or descriptive in nature. Although there are numerous applications of DA that have a significant impact in practice, these are usually published elsewhere, including other INFORMS journals or field journals (such as medical, risk analysis, or petroleum engineering journals) if they are published at all. For example, there are numerous applications of decision analysis in the oil and gas area, but most of this work is not published because of confidentiality concerns. The same is true in the pharmaceutical industry, in which research-based pharmaceutical companies (such as Eli Lilly) routinely build decision-analytic models. These kinds of applications are very impactful but, by now, somewhat routine from a methodological standpoint. The practitioners conducting these studies may have little connection to *MS* or the management science community even if their practice is informed by research that appeared in *MS*.

This phenomenon is not surprising. Indeed, Merrill Flood (1956, p. 180), one of the founders of TIMS who we quoted in the Introduction, suggests "that progress in scientific management consists in the creation and development of a sequence of new professional groups, each specializing in techniques for handling an old management problem in a new manner grounded in a central concept of basic scientific validity." Flood described the statistical quality control movement as an example of this evolutionary process. DA itself is an example of such a new professional group grounded in the central concept of the subjective expected utility paradigm. The Society for Medical Decision Making, as discussed in Section 2.3, represents further evolution along these lines.

In preparing this paper, we have been impressed by just how much DA research has appeared in *MS*. In particular, there has been an explosion in descriptively oriented research on decision making in *MS* in recent years. This is evident, for example, with the recent focus on work in prospect theory, as discussed in Section 2.4. This growth led to the creation of the behavioral economics (BE) and judgment and decision making (JDM) departments at *MS* in 2011 and 2012. The BE department rapidly grew into one of the largest departments at *MS* in terms of the number of papers published. The BE and JDM departments have recently been folded back into a single DA

department that again considers normative, prescriptive, and descriptive research (Simchi-Levi 2018).

This descriptive research has led to new kinds of applications. Traditionally, applications of DA have been aimed at helping firms, governments, or individuals make decisions, with the analysis providing advice and guidance to the DMs. In these new applications, epitomized in Thaler and Sunstein's (2008) book, *Nudge*, analysts serve as "choice architects" who use known biases to "nudge" individuals to improve decision making. For example, Benartzi and Thaler (1999, *MS*) look at the effects of myopic loss aversion on retirement investment decisions: They show that investors who were presented with historical returns for stocks over a 30-year horizon invested much more in stocks than those who were given historical returns over a one-year horizon. If we believe that individuals are overly cautious in their investment decisions (as Benartzi and Thaler seem to), then the choice architect should present investors with returns over longer time horizons. These nudging applications have gathered significant attention in recent years and the British government even created a behavioral research team (known unofficially as the "nudge unit") to try to use the results of behavioral economics to improve government policies and services.

What is exciting to us is the apparent growth and potential for further growth in conversational applications of DA, either by explicit use of DA concepts or by nudges in this direction. For example, it is great to see the explicit use of probabilities in everyday conversations: What is the probability of rain today? What are my team's chances in the tournament this year? And it is good to know that popular forecasts (e.g., from the NWS and FiveThirtyEight) are well informed and well calibrated. Of course, we could do better: For example, as discussed earlier, we would like to see more multiattribute, utility-based product recommendation systems. Such conversational applications of DA are of considerable value in themselves because they help bring clarity to discussions about risks and trade-offs and can help improve decision making.

We also believe that conversational applications of DA may create good habits and lead to more serious applications. For example, somebody who is used to thinking about probabilities for basketball games may ask doctors for probabilities when facing a personal medical decision or may adopt an outside view when talking about the risks associated with a business or government decision. Similarly, somebody who is used to seeing rate-and-weight evaluations of cars in *Car and Driver* may want a similar analysis for other personal and professional applications. And, of course, these new applications may motivate additional research ideas and challenges. The path from high-church research to low-church and conversational applications in DA is not a one-way street; it is a virtuous circle and is sure to keep DA research in journals such as *MS* vibrant for years to come.

Endnotes

[1] We distinguish papers appearing in *MS* when referring to them. We also distinguish papers appearing in *Operations Research* (*OR*).

[2] We are grateful to Andersen and Persinger for sharing these data with us.

References

Abdellaoui M (2000) Parameter-free elicitation of utility and probability weighting functions. *Management Sci.* 46(11):1497–1512.

Abdellaoui M, Kemel E (2014) Eliciting prospect theory when consequences are measured in time units: "Time is not money. *Management Sci.* 60(7):1844–1859.

Abdellaoui M, Bleichrodt H, Paraschiv C (2007) Loss aversion under prospect theory: A parameter-free measurement. *Management Sci.* 53(10):1659–1674.

Abdellaoui M, L'Haridon O, Paraschiv C (2011) Experienced vs. described uncertainty: Do we need two prospect theory specifications? *Management Sci.* 57(10):1879–1895.

Abdellaoui M, Vossmann F, Weber M (2005) Choice-based elicitation and decomposition of decision weights for gains and losses under uncertainty. *Management Sci.* 51(9):1384–1399.

Allais M (1953) Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. *Econometrica* 21(4):503–546.

Allais M (1957) Method of appraising economic prospects of mining exploration over large territories: Algerian Sahara case study. *Management Sci.* 3(4):285–347.

Anscombe F, Aumann R (1963) A definition of subjective probability. *Ann. Math. Statist.* 34(1):199–205.

Ayer T, Alagoz O, Stout N, Burnside E (2016) Heterogeneity in women's adherence and its role in optimal breast cancer screening policies. *Management Sci.* 62(5):1339–1362.

Baucells M, Weber M, Welfens F (2011) Reference-point formation and updating. *Management Sci.* 57(3):506–519.

Bell D (1979) Multiattribute utility functions: Decompositions using interpolation. *Management Sci.* 25(8):744–753.

Bell D, Raiffa H, Tversky A (1988) Descriptive, normative, and prescriptive interactions in decision making. Bell D, Raiffa H, Tversky A, eds. *Decision Making: Descriptive, Normative, and Prescriptive Interactions* (Cambridge University Press, New York), 9–30.

Benartzi S, Thaler R (1999) Risk aversion or myopia? Choices in repeated gambles and retirement investments. *Management Sci.* 45(3):364–381.

Bickel J, Kim S (2008) Verification of the weather channel probability of precipitation forecasts. *Monthly Weather Rev.* 136(12):4867–4881.

Bickel J, Floehr E, Kim S (2011) Comparing NWS PoP forecasts to third-party providers. *Monthly Weather Rev.* 139(10):3304–3321.

Black G (1967) Systems analysis in government operations. *Management Sci.* 14(2): B41–B58.

Bleichrodt H, Pinto J, Wakker P (2001) Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Sci.* 47(11):1498–1514.

Bodily S (1978) Police sector design incorporating preferences of interest groups for equality and efficiency. *Management Sci.* 24(12):1301–1313.

Bond S, Carlson K, Keeney R (2008) Improving the generation of decision objectives. *Decision Anal.* 13(3):235–326.

Borcherding K, Eppel T, von Winterfeldt D (1991) Comparison of weighting judgments in multiattribute utility measurement. *Management Sci.* 37(12):1603–1619.

Butler J, Dyer J, Jia J, Tomak K (2008) Enabling e-transactions with multi-attribute preference models. *Eur. J. Oper. Res.* 186(2):748–765.

Chan T, Demirtas D, Kwon R (2016) Optimizing the deployment of public access defibrillators. *Management Sci.* 62(12): 3617–3635.

Churchman C (1994) Management science: Science of managing and managing of science. *Interfaces* 24(4):99–110.

Churchman C, Ackoff R (1954) An approximate measure of value. *J. Oper. Res. Soc. Amer.* 2(2):172–187.

Crawford D, Huntzinger B, Kirkwood C (1978) Multiobjective decision analysis for transmission conductor selection. *Management Sci.* 24(16):1700–1709.

de Finetti B (1937) La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré* 7(1):1–68.

Debreu G (1960) Topological methods in cardinal utility theory. Arrow K, Karlin S, Suppes P, eds. *Mathematical Methods in the Social Sciences* (Stanford University Press, Stanford, CA), 16–26.

Dyer J, Sarin R (1979) Measurable multiattribute value functions. *Oper. Res.* 27(4):810–822.

Edwards W (1954) The theory of decision making. *Psych. Bull.* 51(4):380–417.

Farquhar P (1984) Utility assessment methods. *Management Sci.* 30(11):1283–1300.

Fischer G, Damodaran N, Laskey K, Lincoln D (1987) Preferences for proxy attributes. *Management Sci.* 33(2):198–214.

Fishburn P (1965) Independence in utility theory with whole product sets. *Oper. Res.* 13(1):28–45.

Fishburn P (1966) Additivity in utility theory with denumerable product sets. *Econometrica* 34(2):500–503.

Fishburn P (1967a) Additive utilities with incomplete product sets: Applications to priorities and assignments. *Oper. Res.* 15(3): 537–542.

Fishburn P (1967b) Conjoint measurement in utility theory with incomplete product sets. *J. Math. Psych.* 4(1):104–119.

Fishburn P (1967c) Methods of estimating additive utilities. *Management Sci.* 13(7):435–453.

Fishburn P (1968) Utility theory. *Management Sci.* 14(5): 335–378.

FiveThirtyEight (2019) How good are FiveThirtyEight forecasts? Accessed October 28, 2018, https://projects.fivethirtyeight.com/checking-our-work/.

Flood M (1955) Decision making. *Management Sci.* 1(2):167–169.

Flood M (1956) The objectives of TIMS. *Management Sci.* 2(2):107–195.

Ford C, Keeney R, Kirkwood C (1979) Evaluating methodologies: A procedure and application to nuclear power plant siting methodologies. *Management Sci.* 25(1):1–10.

Golabi K, Kirkwood C, Sicherman A (1981) Selecting a portfolio of solar energy projects using multiattribute preference theory. *Management Sci.* 27(2):174–189.

Gold M, Siegel J, Russell L, Weinstein M, eds. (1996) *Cost-Effectiveness in Health and Medicine* (Oxford University Press, New York).

Gregory R, Keeney R (1994) Creating policy alternatives using stakeholder values. *Management Sci.* 30(8):1035–1048.

Grushka-Cockayne Y, De Reyck B, Degraeve Z (2008) An integrated decision-making approach for improving European air traffic management. *Management Sci.* 54(8):1395–1409.

Grushka-Cockayne Y, Jose V, Lichtendahl K (2017) Ensembles of overfit and overconfident forecasts. *Management Sci.* 63(4):1110–1130.

Hazen G (2000) Preference factoring for stochastic trees. *Management Sci.* 46(3):389–403.

Hershey J, Schoemaker P (1985) Probability vs. certainty equivalence methods in utility measurement: Are they equivalent? *Management Sci.* 31(10):1213–1231.

Hershey J, Kunreuther H, Schoemaker P (1982) Sources of bias in assessment procedures for utility functions. *Management Sci.* 28(8):936–954.

Howard R (1966) Decision analysis: Applied decision theory. Hertz DB, Melese J, eds. *Proc. 4th Internat. Conf. Oper. Res.* (Wiley-Interscience, New York), 55–71.

Jacquot J (2019) The 2019 Honda Passport and Chevrolet Blazer vs. the Ford Edge, Nissan Murano, and Hyundai Santa Fe, *Car and Driver* (April 5), https://www.caranddriver.com/reviews/comparison-test/a27046083/2019-ford-edge-vs-hyundai-santa-fe-honda-passport/.

Johnson E, Schkade D (1989) Bias in utility assessments: Further evidence and explanations. *Management Sci.* 35(4):406–424.

Jose V, Grushka-Cockayne Y, Lichtendahl K (2013) Trimmed opinion pools and the crowd's calibration problem. *Management Sci.* 60(2):463–475.

Kahneman D, Lovallo D (1993) Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Sci.* 39(1):17–31.

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263–291.

Keeney R (1972) Utility functions for multiattributed consequences. *Management Sci.* 18(5):276–287.

Keeney R, Raiffa H (1976) *Decisions with Multiple Objectives: Preferences and Value Tradeoffs* (Wiley, New York).

Keeney R, von Winterfeldt D, Eppel T (1990) Eliciting public values for complex policy decisions. *Management Sci.* 36(9):1011–1030.

Kilka M, Weber M (2001) What determines the shape of the probability weighting function under uncertainty? *Management Sci.* 47(12):1712–1726.

Krantz D (1964) Conjoint measurement: The Luce-Tukey axiomatization and some extensions. *J. Math. Psych.* 1(2):248–277.

Krantz D, Luce D, Suppes P, Tversky A (1971) *Foundations of Measurement*, vol. I: *Additive and Polynomial Representations* (Academic Press, New York).

Lichtendahl K, Grushka-Cockayne Y, Winkler R (2013) Is it better to average probabilities or quantiles? *Management Sci.* 59(7):1594–1611.

Lichtenstein S, Fischhoff B, Phillips L (1982) Calibration of probabilities: The state of the art to 1980. Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge, UK), 306–334.

Luce R, Tukey J (1964) Simultaneous conjoint measurement: A new type of fundamental measurement. *J. Math. Psych.* 1(1):1–27.

Matheson J, Winkler R (1976) Scoring rules for continuous probability distributions. *Management Sci.* 22(10):1087–1096.

McCord M, de Neufville R (1986) "Lottery equivalents": Reduction of the certainty effect problem in utility assessment. *Management Sci.* 32(1):56–60.

Miyamoto J, Wakker P, Bleichrodt H, Peters H (1998) The zero-condition: A simplifying assumption in QALY measurement and multiattribute utility. *Management Sci.* 44(6):839–849.

Murphy A, Winkler R (1977) Reliability of subjective probability forecasts of precipitation and temperature. *J. Royal Statist. Soc. Ser. C: Appl. Statist.* 26(1):41–47.

Murphy R, ten Brincke R (2018) Hierarchical maximum likelihood parameter estimation for cumulative prospect theory: Improving the reliability of individual risk parameter estimates. *Management Sci.* 64(1):308–326.

Neumann P, Cohen J, Weinstein M (2014) Updating cost-effectiveness—The curious resilience of the $50,000-per-QALY threshold. *New England J. Medicine* 371(9):796–797.

Parnell G, Conley H, Jackson J, Lehmkuhl L, Andrew J (1998) Foundations 2025: A value model for evaluating future air and space forces. *Management Sci.* 44(10):1336–1350.

Persinger C (2019) 20 years of expert-elicited probability assessments at Eli Lilly: Approach and analysis of performance. Presentation at the INFORMS Annual Meeting, October 22, Seattle, WA.

Pliskin J, Beck C (1976) A health index for patient selection: A value function approach with application to chronic renal failure patients. *Management Sci.* 22(9):1009–1021.

Pliskin J, Shepard D, Weinstein M (1980) Utility functions for life years and health status. *Oper. Res.* 28(1):206–224.

Raiffa H (1968) *Decision Analysis: Introductory Lectures on Choices Under Uncertainty* (Addison-Wesley, Oxford, UK).

Ramsey F (1926) Truth and probability. R Braithwaite, ed. *The Foundations of Mathematics and Other Logical Essays* (Harcourt, Brace and Company, New York), 156–198.

Regnier E (2018) Probability forecasts made at multiple lead times. *Management Sci.* 64(5):2407–2426.

Savage L (1954) *The Foundations of Statistics* (Wiley, New York).

Schoemaker P, Waid C (1982) An experimental comparison of different approaches to determining weights in additive utility models. *Management Sci.* 28(2):182–196.

Scott P, Suppes P (1958) Foundational aspects of theories of measurement. *J. Symbolic Logic* 23(2):113–128.

Shubik M (1987) What is an application and when is theory a waste of time. *Management Sci.* 33(12):1511–1522.

Silver N (2012) *The Signal and the Noise: Why So Many Predictions Fail—But Some Don't* (Penguin Books, New York).

Simchi-Levi D (2018) From the editor. *Management Sci.* 64(1):1–4.

SMDM (Society for Medical Decision Making) (2020) About SMDM. Last accessed June 5, 2020, https://smdm.org/hub/page/about-smdm/about.

Smith J, Keeney R (2005) Your money or your life: A prescriptive model for health, safety, and consumption decisions. *Management Sci.* 51(9):1309–1325.

Smith J, McCardle K (1998) Valuing oil properties: Integrating option pricing and decision analysis approaches. *Oper. Res.* 46(2):198–217.

Smith J, McCardle K (1999) Options in the real world: Lessons learned in evaluating oil and gas investments. *Oper. Res.* 47(1):1–15.

Smith J, Nau R (1995) Valuing risky projects: Option pricing theory and decision analysis. *Management Sci.* 41(5):795–816.

Smith J, von Winterfeldt D (2004) Decision analysis in *Management Science*. *Management Sci.* 50(5):561–574.

Spetzler C, Stael von Holstein C (1975) Probability encoding in decision analysis. *Management Sci.* 22(3):340–358.

Stillwell W, von Winterfeldt D, John R (1987) Comparing hierarchical and nonhierarchical weighting methods for eliciting multiattribute value models. *Management Sci.* 33(4):442–450.

Stimson D (1969) Utility measurement in public health decision making. *Management Sci.* 16(2):17–30.

Su X, Zenios S (2006) Recipient choice can address the efficiency-equity trade-off in kidney transplantation: A mechanism design model. *Management Sci.* 52(11):1647–1660.

Tannenbaum D, Fox C, Ülkümen G (2017) Judgment extremity and accuracy under epistemic vs. aleatory uncertainty. *Management Sci.* 63(2):497–518.

Thaler R, Sunstein C (2009) *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Yale University Press, New Haven, CT).

Torrance G (1976) Health status index models: A unified mathematical view. *Management Sci.* 22(9):990–1001.

Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertainty* 5(4):297–323.

von Neumann J, Morgenstern O (1944) *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, NJ).

von Nitzsch R, Weber M (1993) The effect of attribute ranges on weights in multiattribute utility measurements. *Management Sci.* 39(8):937–943.

Wakker P, Deneffe D (1996) Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management Sci.* 42(8):1131–1150.

Wallenius J, Dyer J, Fishburn P, Steuer R, Zionts S, Deb K (2008) Multiple criteria decision making, multiattribute utility theory: Recent accomplishments and what lies ahead. *Management Sci.* 54(7):1336–1349.

Wallsten T, Budescu D (1983) Encoding subjective probabilities: A psychological and psychometric review. *Management Sci.* 29(2):151–173.

Walters D, Fernbach P, Fox C, Sloman S (2017) Known unknowns: A critical determinant of confidence and calibration. *Management Sci.* 63(12):4298–4307.

Weber M, Eisenfuhr F, von Winterfeld D (1988) The effects of splitting weights on multiattribute utility measurement. *Management Sci.* 34(3):431–445.

Winkler R (1967) The assessment of prior distributions in Bayesian analysis. *J. Amer. Statist. Assoc.* 62(319):776–800.

Winkler R (1994) Evaluating probabilities: Asymmetric scoring rules. *Management Sci.* 40(11):1395–1405.

Winkler R, Grushka-Cockayne Y, Lichtendahl K, Jose V (2019) Probability forecasts and their combination: A research perspective. *Decision Anal.* 16(4):239–260.

Wu G, Gonzalez R (1996) Curvature of the probability weighting function. *Management Sci.* 42(12):1676–1690.

Wu G, Gonzalez R (1999) Nonlinear decision weights in choice under uncertainty. *Management Sci.* 45(1):74–85.

Zaric G, Brandeau M, Barnett P (2000) Methadone maintenance and HIV prevention: A cost-effectiveness analysis. *Management Sci.* 46(8):1013–1031.

Zenios S (2002) Optimal control of a paired-kidney exchange program. *Management Sci.* 48(3):328–342.