CHAPTER

2

MODELING PREFERENCES

Farrokh Alemi and David H. Gustafson

This chapter introduces methods for modeling decision makers' values: multi-attribute value (MAV) and multi-attribute utility models. These models are useful in decisions where more than one thing is considered important.

In this chapter, a model is developed for one decision maker. (For more information on modeling a group decision, consult Chapter 6.) Although the model-building effort focuses on the interaction between an analyst and a decision maker, the same process can be used for self-analysis: A decision maker can build a model of her own decisions without the help of an analyst.

Value models are based on Bernoulli's (1738) recognition that money's value does not always equal its amount. He postulated that increasing the amount of income has decreasing value to the wage earner. A comprehensive and rather mathematical introduction to constructing value models was written by Von Winterfeldt and Edwards (1986): This chapter ignores this mathematical foundation, however, to focus on behavioral instructions for making value models.

Value models quantify a person's priorities and preferences. Value models assign numbers to options so that higher numbers reflect more preferred options. These models assume that the decision maker must select from several options and that the selection should depend on grading the preferences for the options. These preferences are quantified by examining the various attributes (i.e., characteristics, dimensions, or features) of the options. For example, if a decision maker were choosing among different electronic health record (EHR) systems, the value of the different EHR systems could be scored by examining such attributes as compatibility with legacy systems, potential effect on practice patterns, and cost. First, the effect of each EHR system on each attribute would be scored—this is often called *single-attribute value function*. Second, scores would be weighted by the relative importance of each attribute. Third, the scores for all attributes would be aggregated, often by using a weighted sum. Fourth, the EHR with the highest weighted score would be chosen.

This book has a companion web site that features narrated presentations, animated examples, PowerPoint slides, online tools, web links, additional readings, and examples of students' work. To access this chapter's learning tools, go to ache.org/DecisionAnalysis and select Chapter 2.

If each option were described in terms of n attributes $A_1, A_2, \ldots, A_n$, then each option would be assigned a score on each attribute: $V(A_1)$, $V(A_2), \ldots, V(A_n)$. The overall value of an option is

$$\text{Value} = \text{Function}\,[V(A_1), V(A_2), \ldots, V(A_n)].$$

In other words, the overall value of an option is a function of the value of the option on each attribute.

Why Model Values?

Values (e.g., attitudes, preferences) play major roles in making management decisions. As mentioned in the first chapter, a *value* is a principle or quality that is intrinsically desirable. It refers to the relative worth, utility, or importance of something. In organizations, decision making is often very complex and a product of collective action. Frequently, decisions must be made concerning issues on which little data exist, forcing managers to make decisions on the basis of opinions rather than fact. Often, there is no correct resolution to a problem because all options are equally legitimate and values play major roles in the final choice.

Many everyday decisions involve value trade-offs. Often, a decision entails finding a way to balance a set of factors that are not all attainable at the same time. Thus, some factors must be given up in exchange for others. Decisions that can benefit from MAV modeling include the following:

- Purchasing software and equipment,
- Contracting with vendors,
- Adding a new clinic or program,
- Initiating float staffing,
- Hiring new staff or paying overtime,
- Balancing missions (e.g., providing service with revenue generating activities), and
- Pursuing quality improvement projects.

In all these decisions, the manager has to trade gains in one area against losses in other areas.

For example, initiating a quality improvement project in a stroke unit means you might not have resources to do the same in a trauma unit, or hiring a technically savvy person may mean you will have to put up with social ineptness. In business, difficult decisions usually involve giving up something to attain other benefits.

Most people acknowledge that a manager's decisions involve consideration of value trade-offs. This is not a revelation. What is unusual is that decision analysts model these values. Some may wonder why the analyst needs to model and quantify value trade-offs. The reasons for modeling decision maker's values include the following:

1. *To clarify and communicate decision makers' perspectives.* Modeling values helps managers communicate their positions by explicitly showing their priorities. These models clarify the basis of decisions so others can see the logic behind the decision and ideally agree with it. For example, Cline, Alemi, and Bosworth (1982) constructed a value model to determine the eligibility of nursing home residents for a higher level of reimbursement (i.e., the "super-skilled" level of nursing care). This model showed which attributes of an applicant affected eligibility and how much weight each attribute deserved. Because of this effort, the regulator, the industry, the care providers, and the patients became more aware of how eligibility decisions were made.
2. *To aid decision making in complex situations.* In complicated situations, decision makers face uncertain events as well as ill-expressed values. In these circumstances, modeling the values adds to the decision maker's understanding of the underlying problem. It helps the decision maker break the problem into its parts and manage the decision more effectively. In short, models help decision makers divide and conquer. Because of the modeling, decision makers may arrive at insight into their own values.
3. *To repeatedly consult the mathematical model instead of the decision maker.* Consider screening a large number of job applicants. If the analyst models the decision maker's values, then he could go through thousands of applicants and select a few that the manager needs to interview. Because the model reflects the manager's values, the analyst is reassured that he has not erroneously screened out applicants that the manager would have liked to interview.
4. *To quantify hard-to-measure concepts.* Concepts such as the severity of illness (Krahn et al. 2000), the medically underserved area (Fos and Zuniga 1999), or the quality of the remaining years of life (Chiou et al. 2005) are difficult concepts to define or measure. These hard-

to-measure concepts are similar to preferences because they are subjective and open to disagreement. Modeling describes these hard-to-measure concepts in terms of several objective attributes that are easier to measure.

Chatburn and Primiano (2001) used value models to examine large capital purchases, such as the decision to purchase a ventilator. Value models have also been used to evaluate drug therapy options (Eriksen and Keller 1993), to measure nurse practice patterns (Anthony et al. 2004), and to evaluate a benefit manager's preferences for smoking cessation programs (Spoth 1990).

Misleading Numbers

Though value models allow you to quantify subjective concepts, the resulting numbers are rough estimates that should not be mistaken for precise measurements. It is important that managers do not read more into the numbers than they mean. Analysts must stress that the numbers in value models are intended to offer a consistent method of tracking, comparing, and communicating rough, subjective concepts and not to claim a false sense of precision.

An important distinction is whether the model is to be used for rank ordering (ordinal scale) or for rating the worth of options (interval scale). Some value models produce numbers that are only useful for rank-ordering options. For example, some severity indexes indicate whether one patient is sicker than another, not how much sicker. In these circumstances, a patient with a severity score of four may not be twice as ill as a patient with a severity score of two. Averaging such ordinal scores is meaningless. In contrast, value models that score on an interval scale show how much more preferable one option is than another. For example, a severity index can be created to show how much more severe one patient's condition is than another's. A patient scoring four can be considered twice as ill as one scoring two. Further, averaging interval scores is meaningful.

Numbers can also be used as a means of classification, such as the nominal scale. Nominal scales produce numbers that are neither ordinal nor interval—for example, the numbers assigned to diseases in the international classification of diseases.

In modeling decision makers' values, single-attribute value functions must be interval scales. If single attributes are measured on an interval scale, these numbers can be added or multiplied to produce the overall score. If measured as an ordinal scale or a nominal scale, one cannot calculate the

overall severity from the single-attribute values. In contrast, overall scores for options need only have an ordinal property. When choosing one option over another, most decision makers care about which option has the highest rating, not how much higher that rating is.

Keep in mind that the purpose of quantification is not to be precise in numerical assessment. The analyst quantifies values of various attributes so that the calculus of mathematics can be used to keep track of them and to produce an overall score that reflects the decision maker's preferences. Quantification allows the use of logic embedded in numbers in aggregation of values across attributes. In the end, model scores are a rough approximation of preferences. They are helpful not because they are precise but because they adequately track contributions of each attribute.

Examples of the Use of Value Models

There are many occasions in which value models can be used to model a decision. A common example is in hiring decisions. In choosing among candidates, the attributes shown in Table 2.1 might be used to screen applicants for subsequent interviews.

In Table 2.1, each attribute has an assigned weight. Each attribute level has an assigned value score. By common convention, attribute levels are set to range from zero to 100. Attribute levels are set so that only one level can be assigned to each applicant. Attribute weights are set so that all weights add up to one. The overall value of an applicant can be measured as the weighted sum of attribute-level scores. In this example, the model assigns to each applicant a score between zero and 100, where 100 is the most preferred applicant. Note that the way the decision maker has rated these attributes suggests that internal promotion is less important than appropriate educational degrees and computer experience. The model can be used to focus interviews on a handful of applicants.

Consider another example about organizing a health fair. Assume that a decision needs to be made about which of the following screenings should be included in the fair:

- Blood pressure
- Peak air flow
- Lack of exercise
- Smoking habits
- Knowledge of breast self-examination
- Depression
- Poor food habits
- Access to a primary care clinician
- Blood sugar levels

TABLE 2.1
A Model for Hiring Decisions

Attribute Weight	*Attribute*	*Attribute Level*	*Value of the Level*
.40	Applicant's education	No college degree	0
		Bachelor of Science or Bachelor of Arts	60
		Master of Science in healthcare field	70
		Master of Science in healthcare-related field	100
		Ph.D. or higher degree	90
.30	Computer skills	None	0
		Data entry	10
		Experience with a database or a worksheet program	80
		Experience with both databases and worksheet programs	100
.20	Internal promotion	No	0
		Yes	100
.10	People skills	Not a strength of the applicant	0
		Contributes to teams effectively	50
		Organizes and leads teams	100

The decision maker is concerned about cost but is willing to underwrite the cost of the fair if it leads to a significant number of referrals. Discussions with the decision maker led to the specification of the attributes shown in Table 2.2.

This simple model will score each screening based on three attributes: (1) the cost of providing the service, (2) the needs of the target group, and (3) whether the screening may generate a visit to the clinic. Once all screening options have been scored and the available funds considered, the top-scoring screening activities can be chosen and offered in the health fair.

A third example concerns constructing practice profiles. Practice profiles are helpful for hiring, firing, disciplining, and paying physicians (Vibbert 1992; McNeil, Pedersen, and Gatsonis 1992). A practice profile compares cost and outcomes of individual physicians to each other. Because patients differ in their severity of illness, it is important to adjust outcomes by the provider's mix of patients. Only then can one compare apples to apples. If there is a severity score, managers can examine patient outcomes to see if

TABLE 2.2 A Model for Health Fair Composition Decisions

Attribute Weight	*Attribute*	*Attribute Level*	*Value of the Level*
.45	Cost of providing the service	Interview cost	0
		Interview and nonintrusive test costs	60
		Interview and intrusive test costs	100
.35	Need in target group	Unknown	0
		Less than 1% are likely to be positive	10
		1% to 5% are likely to be positive	80
		More than 5% likely to be positive	100
.20	Generates a likely visit	No	0
		Yes	100

they are within expectations. Managers can compare two clinicians to see which one had better outcomes for patients with the same severity of illness. Armed with a severity index, managers can compare cost of care for different clinicians to see which one is more efficient. Value models can be used to create severity indexes—for example, a severity index for acquired immunodeficiency syndrome (AIDS).

After the diagnosis of human immunodeficiency virus (HIV) infection, patients often suffer a complex set of different diseases. The cost of treatment for each patient is heavily dependent on the course of their illness. For example, patients with skin cancer, Kaposi's sarcoma, have significantly lower first-year costs than patients with more serious infections (Freedberg et al. 1998). Thus, if a manager wants to compare two clinicians in their ability to care for AIDS patients, it is important to measure the severity of AIDS among their patients. Alemi and colleagues (1990) used a value model to create a severity index for AIDS. Even though much time has elapsed since the creation of this index, and care of AIDS patients has progressed, the method of developing the severity index is still relevant. The development of this index will be referred to at length throughout this chapter.

Steps in Modeling Values

Using the example of the AIDS severity index (Alemi et al. 1990), this section shows how to examine the need for a value model and how to create such a model.

Step 1: Determine if a Model Would Help

The first and most obvious question is whether constructing a value model will help resolve the problem faced by the manager. Defining the problem is the most significant step of the analysis, yet surprisingly little literature is available for guidance. To define a problem, the analyst must answer several related questions: Who is the decision maker? What are the objectives this person wishes to achieve? What role do subjective judgments play in these goals? Should a value model be used? How will it be used?

Who Is the Decision Maker?

In organizations, there are often many decision makers. No single person's viewpoint is sufficient, and the analyst needs a multidisciplinary consensus instead. Chapter 6 discusses how the values of a group of people can be modeled.

The core of the problem in the example was that AIDS patients need different amounts of resources depending on the severity of their illness. The federal administrators of the Medicaid program wanted to measure the severity of AIDS patients because the federal government paid for part of their care. The state administrators were likewise interested because state funds paid for another portion of their care. Individual hospital administrators were interested in analyzing a clinician's practice patterns and recruiting the most efficient. No single decision maker was involved. In short, the model focused on how a clinician makes severity judgments and thus brought together a group of physicians involved with care of and research on AIDS patients. For simplicity, the following discussion assumes that only one person is involved in the decision-making process.

What Are the Objectives?

Problem solving starts by recognizing a gap between the present situation and the desired future. Typically, at least one decision maker has noticed a difference between what is and what should be and begins to share this awareness with the relevant levels of the organization. Gradually, a motivation is created to change, informal social ties are established to promote the change, and an individual or group receives a mandate to find a solution.

Often, a perceived problem may not be the real issue. Occasionally, the decision maker has a solution in mind before fully understanding the problem, which shows the need for examining the decision maker's circumstances in greater depth. When solutions are proposed prematurely, it

is important to sit back and gain a greater perspective on the problem. In these situations, it is the analyst's responsibility to redefine the problem to make it relevant to the real issues. An analyst can use tools that help the decision maker better define the problem. There are many ways to encourage creativity (Sutton 2001), including structured techniques such as brainstorming (Fields 1995) and less structured techniques such as analogies.

What Role Do Subjective Judgments Play?

After the problem has been defined, the analyst must examine the role subjective judgments will play in its resolution. One can do this by asking questions such as the following: What plans would change if the judgment were different? How are things being done now? If no one makes a judgment about the underlying concept, would it really matter, and who would complain? Would it be useful to tell how the judgment was made, or is it better to leave matters rather ambiguous? Must the decision maker choose among options, or should the decision maker let things unfold on their own? Is a subjective component critical to the judgment, or can it be based on objective standards?

In the severity index example, the administrators needed to budget for the coming years, and they knew judgments of severity would help them anticipate utilization rates and overall costs. Programs caring for low-severity patients would receive a smaller allocation than programs caring for high-severity patients. But no objective measures of severity were available, so clinicians' judgments concerning severity were used instead.

Should a Value Model Be Used?

Experts seem to intuitively know the prognosis of a patient and can easily recognize a very sick patient. Although in the AIDS example it was theoretically possible to have an expert panel review each case and estimate severity, it was clear from the outset that a model was needed because of the high cost of case-by-case review. Moreover, the large number of cases would require the use of several expert panels, each judging a subset of cases, and the panels might disagree. Further, judgments within a panel can be quite inconsistent over time. In contrast, the model provided a quick and consistent way of rating the severity of patients. It also explained the rationale behind the ratings, which allowed skeptics to examine the fairness of judgments, thus increasing the acceptance of those judgments.

How Will the Value Model Be Used?

In understanding what judgments must be made, it was crucial to attend to the limitations of circumstances in which these judgments are going to be made. The use of existing AIDS severity indexes was limited because they relied on physiological variables that were unavailable in many databases. Alemi and his colleagues (1990) were asked to predict prognoses from existing data. The only information widely available on AIDS patients was

diagnoses, which were routinely collected after every encounter. Because these data did not include any known physiological predictors of survival (such as number of T4 cells), the manager needed an alternative way to predict survival. The severity index was created to serve as this alternative.

Step 2: Soliciting Attributes

After determining whether a model would be useful, the second step is to identify the attributes needed for making the judgment. For example, Alemi and his colleagues (1990) needed to understand and identify the patient attributes that should be used to predict AIDS severity. For the study, six experts known for clinical work with AIDS patients or for research on the survival of AIDS patients were assembled. Physicians came from several programs located in states with high rates of HIV/AIDS. The experts were interviewed to identify the attributes used in creating the severity index. When interviewing an expert to determine the attributes needed for a model, the analyst should keep introductions brief, use tangible examples, arrange the attributes in a hierarchy, take notes, and refrain from interrupting.

Keep Introductions Brief

Being as brief as possible, the analyst should introduce herself and explain the expert's role, the model's purpose, and how the model will be developed. An interview is going well if the analyst is listening and the expert is talking. If it takes five minutes just to describe the purpose of the interview, then something is amiss. Probably, the analyst does not understand the problem well, or possibly the expert is not familiar with the problem.

Be assertive in setting the interview's pace and agenda. Because the expert is likely to talk whenever the analyst pauses, the analyst should be judicious about pausing. For example, if one pauses after saying, "Our purpose is to construct a severity index to work with existing databases," the expert will likely use that opportunity for an in-depth discussion about the purpose. But if the analyst immediately follows the previous sentence with a question about the expert's experience in assessing severity, the expert is more likely to begin describing his background. The analyst sets the agenda and should pause in such a way as to make progress in the interview.

Use Tangible Examples

Concrete examples help the analyst understand which patient attributes should be used in the model and how they can be measured. Ask the expert to recall an actual situation and to contrast it with other occasions to discern the key discriminators. For example, the analyst might ask the expert to describe a severely ill patient in detail to ensure that the expert is referring to a particular patient rather than a hypothetical one. Then, the analyst asks for a description of a patient who was not severely ill and tries to elicit the key differences

between the two patients; these differences are attributes the analyst can use to judge severity. The following is a sample dialog:

Analyst: Can you recall a specific patient with a very poor prognosis?
Expert: I work in a referral center, and we see a lot of severely ill patients. They seem to have many illnesses and are unable to recover completely, so they continue to worsen.

Analyst: Tell me about a recent patient who was severely ill.
Expert: A 28-year-old homosexual male patient deteriorated rapidly. He kept fighting recurrent influenza and died from gastrointestinal (GI) cancer. The real problem was that he couldn't tolerate AZT, so we couldn't help him much. Once a person has cancer, we can do little to maintain him.

Analyst: Tell me about a patient with a good prognosis—say, close to five years.
Expert: Well, let me think. A year ago, we had a 32-year-old male patient diagnosed with AIDS who has not had serious disease since—a few skin infections, but nothing serious. His spirit is up, he continues working, and we have every reason to expect he will survive four or five years.

Analyst: What key difference between the two patients made you realize that the first patient had a poorer prognosis than the second?
Expert: That's a difficult question. Patients are so different from each other that it's tough to point to one characteristic. But if you really push me, I would say two characteristics: the history of illness and the ability to tolerate AZT.

Analyst: What about the history is relevant?
Expert: If I must predict a prognosis, I want to know whether the patient has had serious illness in vital organs.

Analyst: Which organs?

Expert: Brain, heart, and lungs are more important than, say, skin.

In this dialog, the analyst started with tangible examples and used the terminology and words introduced by the expert to discuss concrete examples. There are two advantages to this process. First, it helps the expert recall the details without the analyst introducing unfamiliar words, such as "attributes." Second, soliciting attributes by contrasting patients helps

single out those attributes that truly affect prognosis. Thus, it does not produce a wish list of information that is loosely tied to severity—an extravagance one cannot afford in model building.

After the analyst has identified some attributes, the analyst can ask directly for additional attributes that indicate prognosis. One might ask if there are other markers of prognosis, if the expert has used the word "marker." If necessary, the analyst might say, "In our terminology, we refer to the kinds of things you have mentioned as markers of prognosis. Are there other markers?" The following is an example dialog:

Analyst: Are there other markers for poor prognosis?
Expert: Comorbidities are important. Perhaps advanced age suggests poorer prognosis. Sex may matter.

Analyst: Does the age or sex really matter in predicting prognosis?
Expert: Sex does not matter, but age does. But there are many exceptions. You cannot predict the prognosis of a patient based on age alone.

Analyst: What are some other markers of poor prognosis?

As you can see in the dialog, the analyst might even express her own ideas without pushing them on the expert. In general, analysts are not there to express their own ideas; they are there to listen. However, they can ask questions to clarify things or even to mention things overlooked by the expert, as long as it does not change the nature of the relationship between the analyst and the expert.

The analyst should always use the expert's terminology, even if a reformulation might help. Thus, if the expert refers to "sex," the analyst should not substitute "gender." Such new terminology may confuse the conversation and create an environment where the analyst acts more like an expert, which can undermine the expert's confidence that she is being heard. It is reasonable, however, to ask for clarification—"sex" could refer to gender or to sex practices, and the intended meaning is important.

In general, less esoteric prompts are more likely to produce the best responses, so formulate a few prompts and use the prompts that feel most natural for the task. Avoid jargon, including the use of terminology from decision analysis (e.g., attribute, value function, aggregation rules).

Arrange the Attributes in a Hierarchy

An attribute hierarchy should move from broad to specific attributes (Keeney 1996). Some analysts suggest using a hierarchy to solicit and structure the attributes. For example, an expert may suggest that a patient's prognosis depends on medical history and demographics, such as age and sex. Medical history involves the nature of the illness, comorbidities, and tolerance of

AZT. The nature of illness breaks down into body systems involved (e.g., skin, nerves, blood). Within each body system, some diagnoses are minor and other diagnoses are more threatening. The expert then lists, within each system, a range of diseases. The hierarchical structure promotes completeness and simplifies tracking many attributes. A detailed example of arranging attributes in hierarchical structure is presented later in this chapter.

Take Notes and Refrain from Interrupting

The analyst should take notes and not interrupt. He should have paper and a pencil available, and write down the important points. Not only does this help the expert's recall, but it also helps the analyst review matters while the expert is still available. Experts tend to list a few attributes, then focus attention on one or two. The analyst should actively listen to these areas of focus. When the expert is finished, the analyst should review the notes for items that need elaboration. If certain points are vague, the analyst should ask for examples, which are an excellent means of clarification. For instance, after the expert has described attributes of vital organ involvement, the analyst may ask the expert to elaborate on something mentioned earlier, such as "acceptance of AZT." If the expert mentions other topics in the process, return to them after completing the discussion of AZT acceptance. This ensures that no loose ends are left when the interview is finished and reassures the expert that the analyst is indeed listening.

Other Approaches

Other, more statistical approaches to soliciting attributes are available, such as multidimensional scaling and factor analysis. However, the behavioral approach to soliciting attributes (i.e., the approach of asking the expert to specify the attributes) is preferred because it involves the expert more in the process and leads to greater acceptance of the model.

Step 3: Examine and Revise the Attributes

After soliciting a set of attributes, it is important to examine and, if necessary, revise them. Psychological research suggests that changing the framing of a question alters the response (Kahneman 2003). Consider the following two questions:

1. What are the markers for survival?
2. What are the markers for poor prognosis?

One question emphasizes survival, the other mortality. One would expect that patient attributes indicating survival would also indicate mortality, but researchers have found this to be untrue (see Chow, Haddad, Wong-Boren 1991; Nisbett and Ross 1980). Experts may identify entirely different attributes for survival and mortality. This research suggests that value-laden prompts tap different parts of the memory and can evoke recall

of different pieces of information. Evidence about the impact of questions on recall and judgment is substantial. How questions are framed affects what answers are provided (Kim et al. 2005). Such studies suggest that analysts should ask their questions in two ways, once in positive terms and again in negative terms.

Several tests should be conducted to ensure that the solicitation process succeeded. The first test ensures that the listed attributes are exhaustive by using them to describe several hypothetical patients and asking the expert to rate their prognosis. If the expert needs additional information for a judgment, solicit new attributes until the expert has enough information to make the judgment.

A second test checks that the attributes are not redundant by examining whether knowledge of one attribute implies knowledge of another. For example, the expert may consider "inability to administer AZT" and "cancer of GI tract" redundant if no patient with GI cancer can accept AZT. In such cases, either the two attributes should be combined into one, or one must be dropped from the analysis.

A third test ensures that each attribute is important to the decision maker's judgment. The analyst can test this by asking the decision maker to judge two hypothetical situations: one with the attribute at its lowest level and another with the attribute at peak level. If the judgments are similar, the attribute may be ignored. For example, gender may be unimportant if male and female AIDS patients with the same history of illness have identical prognoses.

Fourth, a series of tests examines whether the attributes are related or are independent (Goodwin and Wright 2004; Keeney 1996).

In the AIDS severity study (Alemi et al. 1990), discussions with the expert and later revisions led to the following set of 18 patient attributes for judging the severity of AIDS:

1. Age
2. Race
3. Transmission mode
4. Defining diagnosis
5. Time since defining diagnosis
6. Diseases of nervous system
7. Disseminated diseases
8. GI diseases
9. Skin diseases
10. Lung diseases
11. Heart diseases
12. Recurrence of a disease
13. Functioning of the organs
14. Comorbidity
15. Psychiatric comorbidity
16. Nutritional status
17. Drug markers
18. Functional impairment

As the number of attributes in a model increases, the chances for preferential dependence also increases. The rule of thumb is that

preferential dependencies are much more likely in value models with more than nine attributes.

Step 4: Set Attribute Levels

Once the attributes have been examined and revised, the possible levels of each attribute can be identified. The analyst starts by deciding if the attributes are discrete or continuous. Attributes such as age are continuous; attributes such as diseases of the nervous system are discrete. However, continuous attributes may be expressed in terms of a few discrete levels, so that age can be described in decades, not individual years. The four steps in identifying the levels of an attribute are to (1) define the range, (2) define the best and worst levels, (3) define some intermediate levels, and (4) fill in the other possible levels so that the listing of the levels is exhaustive and capable of covering all possible situations.

To define the range, the analyst must select a target population and ask the expert to describe the possible range of the attributes in it. Thus, for the AIDS severity index, the analyst asked the experts to focus on adult AIDS patients and, for each attribute, suggest the possible ranges. To assess the range of nervous system diseases, the analyst asked the following question:

> Analyst: In adult AIDS patients, what is a disease that suggests the most extensive involvement of the nervous system?

Next, the analyst asked the expert to specify the best and the worst possible levels of each attribute. In the AIDS severity index, one could easily identify the level with the best possible prognosis: the normal finding within each attribute—in common language, the healthy condition. The analyst accomplished the more difficult task of identifying the level with the worst possible prognosis by asking the expert the following question:

> Analyst: What would be the gravest disease of the central nervous system, in terms of prognosis?

A typical error in obtaining the best and the worst levels is failing to describe these levels in detail. For example, in assessing the value of nutritional status, it is not helpful to define the levels as simply the best nutritional status or the worst nutritional status. Nor does it help to define the worst level as "severely nutritionally deficient" because the adjective "severe" is not defined. Analysts should avoid using adjectives in describing levels, as experts perceive words like "severely" or "best" in different ways. The levels must be defined in terms of the underlying physical process measured in each attribute, and the descriptions must be connected to the nature of the attribute. Thus, a good level for the worst nutritional status might

be "patients on total parenteral nutrition," and the best status might be "nutritional treatment not needed."

Next, the analyst should ask the expert to define intermediate levels. These levels are often defined by asking for a level between the best and worst levels. In the severity index example, this dialog might occur as follows:

Analyst: I understand that patients on total parenteral nutrition have the worst prognosis. Can you think of other relatively common conditions with a slightly better prognosis?

Expert: Well, a host of things can happen. Pick up any book on nutritional diseases, and you find all kinds of things.

Analyst: Right, but can you give me three or four examples?

Expert: Sure. The patient may be on antiemetics or nutritional supplements.

Analyst: Do these levels include a level with a moderately poor prognosis and one with a relatively good prognosis?

Expert: Not really. If you want a level indicative of moderately poor prognosis, then you should include whether the patient is receiving Lomotil or Imodium.

It is not always possible to solicit all possible levels of an attribute from the expert interviews. In these circumstances, the analyst can fill in the gaps afterward by reading the literature or interviewing other experts. The levels specified by the first expert are used as markers for placing the remaining levels, so that the levels range from best to worst. In the example, a clinician on the project team reviewed the expert's suggestions and filled in a long list of intermediate levels.

Step 5: Assign Values to Single Attributes

The analysis proceeds with the evaluation of single-attribute value function (i.e., a scoring procedure that assigns the relative value of each level in a single attribute). The procedure recommended here is called double-anchored estimation. In this method, the attribute levels are first ranked, or, if the attribute is continuous, the most and least preferred levels are specified and assigned scores of 0 and 100. Finally, the best and the worst levels are used as "anchors" for assessing the other levels.

For example, skin infections have the following levels:

- No skin disorder
- Kaposi's sarcoma

- Shingles
- Herpes complex
- *Candidiasis*
- Thrush

The following interaction typifies the questioning for the double-anchored estimation method:

Analyst: Which skin disorder has the worst prognosis?
Expert: None is really that serious.

Analyst: Yes, I understand that, but which is the most serious?
Expert: Patients with thrush perhaps have a worse prognosis than patients with other skin infections.

Analyst: Let's rate the severity of thrush at 100 and place the severity of no skin disorder at zero. How would you rate shingles?
Expert: Shingles is almost as serious as thrush.

Analyst: This tells me that you might rate the severity of shingles nearer 100 than zero. Where exactly would you rate it?
Expert: Maybe 90.

Analyst: Can you now rate the remaining levels?

Several psychologists have questioned whether experts are systematically biased in assessing value because using different anchors produces different value functions (Chapman and Johnson 1999). For example, in assessing the value of money, gains are judged differently than losses; furthermore, the value of money is judged according to the decision maker's current assets (Kahneman 2003). Because value may depend on the anchors used, it is important to use different anchors besides just the best or worst levels. Thus, if the value of skin infections is assessed by anchoring on shingles and no skin infections, then it is important to verify the ratings relative to other levels. Assume the expert rated skin infections as follows:

Attribute level	*Rating*
No skin disorder	0
Kaposi's sarcoma	10
Shingles	90
Herpes complex	95
Candidiasis	100
Thrush	100

The analyst might then ask the following:

Analyst: You have rated herpes complex halfway between shingles and *candidiasis*. Is this correct?
Expert: Not really. Prognosis of patients with herpes is closer to patients with *candidiasis*.

Analyst: How would you change the ratings?
Expert: Maybe we should rate herpes 98.

It is occasionally useful to change not only the anchors but also the assessment method. A later section describes several alternative methods of assessing single-attribute value functions. When a value is measured by two different methods, there would be inadvertent discrepancies; the analyst must ask the expert to resolve these differences.

By convention, the single-attribute value function must range from zero to 100. Sometimes, experts and decision makers refuse to assign the zero value. In these circumstances, their estimated values should be revised to range from zero to 100. The following formula shows how to obtain standardized value functions from estimates that do not range from zero to 100:

$$\text{Standardized value for level } X = 100 \times \frac{\text{Value assigned to level } X - \text{Value of least important level}}{\text{Value of most important level} - \text{Value of least important level}}$$

For example, assume that the skin diseases attributes are rated as follows:

Attribute level	*Rating*
No skin disorder	10
Kaposi's sarcoma	20
Thrush	90

Then, the maximum value is 90 and the minimum value is 10, and standardized values can be assigned to each level using the formula above. For example, for Kaposi's sarcoma the value is

$$\text{Standardized value for Kaposi's sarcoma} = 100 \times \frac{20 - 10}{90 - 10} = 12.5.$$

Step 6: Choose an Aggregation Rule

In this step, the analysis proceeds when one finds a way to aggregate single-attribute functions into an overall score evaluated across all attributes.

Note that the scoring convention has produced a situation in which the value of each attribute is somewhere between zero and 100. Thus, the prognosis of patients with skin infection and the prognosis of patients with various GI diseases have the same range. Adding these scores will be misleading because skin infections are less serious than GI problems, so the analyst must find an aggregation rule that differentially weights the various attributes.

The most obvious rule is the *additive value model.* Assume that S represents the severity of AIDS. If a patient is described by a series of n attributes of $(A_1, A_2, \ldots, A_i, \ldots, A_n)$, then, using the additive rule, the overall severity is

$$S = \sum_i W_i \times V_i (A_j),$$

where

- $V_i (A_j)$ is the value of the jth level in the ith patient attribute,
- W_i is the weight associated with the ith attribute in predicting prognosis, and
- $\sum_i W_i = 1$.

Several other models are possible in addition to the additive model. The multiplicative model form is described in a later section of this chapter.

Step 7: Estimate Weights

The analyst can estimate the weights for an additive value model in a number of ways. It is often useful to mix several approaches. Some analysts estimate weights by assessing how many times one attribute is more important than the other (Edwards and Barron 1994; Salo and Hämäläinen 2001). The attributes are rank ordered, and the least important is assigned ten points. The expert is then asked to estimate the relative importance of the other attributes by estimating how many times the next attribute is more important. There is no upper limit to the number of points other attributes can be assigned. For example, in estimating the weights for the three attributes of skin infections, lung infections, and GI diseases, the analyst and the expert might have the following discussion:

Analyst: Which of the three attributes is most important?

Expert: Well, they are all important, but patients with either lung infections or GI diseases have worse prognoses than patients with skin infections.

Analyst: Do lung infections have a worse prognosis than GI diseases?

Expert: That's more difficult to answer. No, I would say that for all practical purposes, they have the same prognosis. Well, now that I think about it, perhaps patients with GI diseases have a slightly worse prognosis.

Having obtained the rank ordering of the attributes, the analyst can proceed to estimating the importance weights as follows:

Analyst: Let's say that we arbitrarily rate the importance of skin infection in determining prognosis at ten points. GI diseases are how many times more important than skin infections?

Expert: Quite a bit. Maybe three times.

Analyst: That is, if we assign 10 points to skin infections, we should assign 30 points to the importance of GI diseases?

Expert: Yes, that sounds right.

Analyst: How about lung infections? How many more times important are they than GI diseases?

Expert: I would say about the same.

Analyst: (Checking for consistency in the subjective judgments.) Would you consider lung infections three times more serious than skin infections?

Expert: Yes, I think that should be about right.

In the dialog above, the analyst first found the order of the attributes and then asked for the ratio of the weights of the attributes. Knowing the ratio of attributes allows the analyst to estimate the attribute weights. If the model has only three attributes, the weights for the attributes can be obtained by solving the following three equations:

$$\frac{W(\text{GI diseases})}{W(\text{skin infection})} = 3,$$

$$\frac{W(\text{lung diseases})}{W(\text{skin infection})} = 3,$$

$$W(\text{lung diseases}) + W(\text{skin infection}) + W(\text{GI diseases}) = 1.$$

One characteristic of this estimation method is that its emphasis on the ratio of the importance of the attributes leads to relatively extreme weighting

compared to other approaches. Thus, some attributes may be judged critical, and others rather trivial. Other approaches, especially the direct magnitude process, may judge all attributes as almost equally important.

In choosing a method to estimate weights, the analyst should consider several trade-offs, such as ease of use and accuracy of estimates. The analyst can introduce errors by asking experts awkward and partially understood questions. It is best to estimate weights in several ways and use the resulting differences to help experts think more carefully about their real beliefs. In doing so, the analyst usually starts with a rank-order technique, then moves on to assess ratios, obtain a direct magnitude estimate, identify discrepancies, and finally ask the expert to resolve them.

One note of caution: Some scientists have questioned whether experts can describe how they weight attributes. Nisbett and Miyamoto (2005) argue that directly assessed weight may not reflect an expert's true beliefs. Other investigators find that directly assessing the relative importance of attributes is accurate (Naglie et al. 1997). In the end, what matters is not the weight of individual attributes but the accuracy of the entire model, which is discussed in the next section.

Step 8: Evaluate the Accuracy of the Model

Although researchers know the importance of carefully evaluating value models, analysts often lack the time and resources to do this. Because of the importance of having confidence in the models and being able to defend the analytical methodology, this section presents several ways of testing the adequacy of value models.

Most value models are devised to apply to a particular context, and they are not portable to other settings or uses. This is called *context dependence*. In general, it is viewed as a liability, but this is not always the case. For example, the AIDS severity index may be intended for evaluating practice patterns, and its use for evaluating prognosis of individual patients is inappropriate and possibly misleading.

The value model should require only available data for input. Relying on obscure data may increase the model's accuracy at the expense of practicality. Thus, the severity index should rely on reasonable sources of data, usually from existing databases. A physiologically based database, for instance, would predict prognosis of AIDS patients quite accurately. However, such an index would be useless if physiological information is generally unavailable and routine collection of this information would take considerable time and money. While the issue of data availability may seem obvious, it is a very common error in the development of value models. Experts used to working in organizations with superlative data systems may want data

that are unavailable at average institutions, and they may produce a value model with limited usefulness. If there are no plans to compare scores across organizations, one can tailor indexes to each institution's capabilities and allow each institution to decide whether the cost of collecting new data is justified by the expected increase in accuracy. However, if scores will be used to compare institutions or allocate resources among institutions, then a single-value model based on data available to all organizations is needed.

The model should be simple to use. The index of medical underservice areas is a good example of the importance of simplicity (Health Services Research Group 1975). This index, developed to help the federal government set priorities for funding HMOs, community health centers, and health-facility development programs, originally had nine attributes; the director of the sponsoring federal agency rejected the index because of the number of variables. Because he wanted to be able to "calculate the score on the back of an envelope," the index was reduced to four attributes. The simplified version performed as well as one with a larger model; it was used for eight years to help set nationwide funding priorities. This example shows that simplicity does not always equal incompetence. Simplicity nearly always makes an index easy to understand and use.

When different people apply the value model to the same situation, they must arrive at the same scores; this is referred to as *interrater reliability*. In the example of the severity index (Alemi et al. 1990), different registered record abstractors who use the model to rate the severity of a patient should produce the same score. If a model relies on hard-to-observe patient attributes, the abstractors will disagree about the condition of patients. If reasonable people using a value model reach different conclusions, then one loses confidence in the model's usefulness as a systematic method of evaluation. Interrater reliability is tested by having different abstractors rate the severity of randomly selected patients.

The value model should also seem reasonable to experts—this is coined *face validity*. Thus, the severity index should seem reasonable to clinicians and managers. Otherwise, even if it accurate, one may experience problems with its acceptance. Clinicians who are unfamiliar with statistics will likely rely on their experience to judge the index, meaning that the variables, weights, and value scores must seem reasonable and practical to them. Face validity is tested by showing the model to a new set of experts and asking if they understand it and whether it is conceptually reasonable.

One way to establish the validity of a model is to show that it simulates the judgment of the experts; then, if the experts' acumen is believed, the model should be considered valid. In this approach, the expert is asked to score several (perhaps 100) hypothetical case profiles described only by

attributes included in the model. If the model accurately predicts the expert's judgments, confidence in the model increases; but this measure has the drawback of producing optimistic results. After all, if the expert who developed the model cannot get the model to predict her judgments, who can? It is far better to ask a separate panel of experts to rate the patient profiles. In the AIDS severity project, the analyst collected the expert's estimate of survival time for 97 hypothetical patients and examined whether the value model could predict these ratings. The correlation between the additive model and the rating of survival was –0.53. (The negative correlation means that high severity scores indicate shorter survival; the magnitude of the correlation ranges between 1.0 and –1.0.) The correlation of –0.53 suggests low to moderate agreement between the model and the expert's intuitions; correlations closer to 1.0 or –1.0 imply greater agreement. A correlation of zero suggests no agreement. One can judge the adequacy of the correlations by comparing them with agreement among the experts themselves. The correlation between several pairs of experts rating the same 97 hypothetical patients was also in a similar range. The value model agreed with the average of the experts as much as the experts agreed with each other. Thus, the value model may be a reasonable approach to measuring severity of AIDS.

A model is considered valid if several different ways of measuring it lead to the same finding. This method of establishing validity is referred to as *construct validity.* For example, the AIDS severity model should be correlated with other measures of AIDS severity. If the analyst has access to other severity indexes, such as physiologically based indexes, the predictions of the different approaches can be compared using a sample of patients. One such study was done for the index described in this section. In a follow-up article about the severity index, Alemi and his colleagues (1999) reported that the index did not correlate well against physiological markers. If it had, confidence in the severity index would have been increased because physiological markers and the index were measuring the same thing. Given that the two did not have a high correlation, clearly they were measuring different aspects of severity, and the real question was which one was more accurate. As it turns out, the severity index presented in this chapter was more accurate in predicting survival than physiological markers.

In some situations, one can validate a value model by comparing the model's predictions against observable behavior. This method of establishing validity is referred to as *predictive validity.* If a model is used to measure a subjective concept, its accuracy can be evaluated by comparing predictions to an observed and objective standard, which is often called

the *gold standard*, to emphasize its status as being beyond debate. In practice, gold standards are rarely available for judging the accuracy of subjective concepts (otherwise, one would not need the models in the first place). For example, the accuracy of a severity index can be examined by comparing it to observed outcomes of patients' care. When the severity index accurately predicts outcomes, there is evidence favoring the model. The model developed in this section was tested by comparing it to patients' survival rates. The medical histories of patients were analyzed using the model, and the ability of the severity score to predict patients' prognoses was examined. The index was more accurate than physiological markers in predicting patients' survival.

Other Methods for Assessing Single-Attribute Value Functions

Single-attribute value functions can be assessed in a number of different ways aside from the double-anchored method (Torrance et al. 1995). The *midvalue splitting technique* sets the best and worst levels of the attributes at 100 and zero. Then the decision maker finds a level of the attribute that psychologically seems halfway between the best and the worst levels. The value for this level is set to 50. Using the best, worst, and midvalue points, the decision maker continues finding points that psychologically seem halfway between any two points. After several points are identified, the values of other points are assessed by linear extrapolation from existing points. The following conversation illustrates how the midvalue splitting technique could be used to assess the value of age in assessing AIDS severity.

Analyst: What is the age with the best prognosis?
Expert: A 20-year-old has the best chance of survival.

Analyst: What is the age with the worst prognosis?
Expert: AIDS patients over 70 years old are more susceptible to opportunistic infections and have the worst prognosis. Of course, infants with AIDS have an even worse prognosis, but I understand we are focusing on adults.

Analyst: Which age has a prognosis half as bad as a 70-year-old?
Expert: I am going to say about 40, though I am not really sure.

Analyst: I understand. We do not need exact answers. Perhaps it may help to ask the question differently. Do you think an

increase in age from 40 to 70 causes as much of a deterioration in prognosis as an increase from 20 to 40 years?

Expert: If you are asking roughly, yes.

Analyst: If 20 years is rated as zero and 70 years as 100, do you think it would be reasonable to rate 40 years as 50?

Expert: I suppose my previous answers imply that I should say yes.

Analyst: Yes, but this is not binding—you can revise your answers.

Expert: A rating of 50 for the age of 40 seems fine as a first approximation.

Analyst: Can you tell me what age would have a prognosis halfway between 20 and 40 years old?

Using the midvalue splitting technique, the analyst chooses a value score, and the expert specifies the particular attribute level that matches it. This is opposite to the double-anchored estimation, in which the analyst specifies an attribute level and asks for its value. The choice between the two methods should depend on whether the attribute is discrete or continuous. Often with discrete attributes, there are no levels to correspond to particular value scores, so analysts have no choice but to select the double-anchored method.

Another method for assessing a value function is to draw a curve in the following fashion: The levels of the attributes are sorted and set in the x-axis. The y-axis is the value associated with each attribute level. The best attribute level is assigned 100 and drawn on the curve. The worst attribute level is assigned zero. The expert is asked to draw a curve between these two points showing the value of remaining attribute levels. Once the graph is drawn, the analyst and the expert review its implications. For example, a graph can be constructed with age (20 to 70 years) on the x-axis and value (0 to 100) on the y-axis. Two points are marked on the graph (age 20 at zero value and age 70 at 100 value). The analyst asks the expert to draw a line between these two points showing the prognosis for intermediate ages.

Finally, an extremely easy method, which requires no numerical assessment at all, is to assume a linear value function over the attribute. This arbitrary assumption introduces some errors, but they will be small if an ordinal value scale is being constructed and if the single-attribute value function is monotonic (meaning that an increase in the attribute level will cause either no change or an increase in value).

For example, one cannot assume that increasing age will cause a proportionate decline in prognosis. In other words, the relationship between

the variables is not monotonic: The prognosis for infants is especially poor, while 20-year-old patients have the best prognosis and 70-year-old patients have a poor outlook. Because increasing age does not consistently lead to increasing severity—and in fact it can also reduce severity—an assumption of linear value is misleading.

Other Methods for Estimating Weights

In the *direct magnitude estimate*, the expert is asked to rank order the attributes and then rate their importance by assigning each a number between zero and 100. Once the ratings are obtained, they are scaled to range between zero and one by dividing each weight by the sum of the ratings. Subjects rarely rate the importance of an attribute near zero, so the direct magnitude estimation has the characteristic of producing weights that are close together, but the process has the advantage of simplicity and comprehensibility.

Weights can also be estimated by having the expert distribute a fixed number of points, typically 100, among the attributes. The main advantage of this method is simplicity, as it is only slightly more difficult than the ranking method. But if there are a large number of attributes, experts will have difficulty assigning numbers that total 100.

One approach to estimating weights is to ask the expert to rate "corner" cases. A corner case is a description of a patient with one attribute at its most extreme level and the remainder at minimum levels. The expert's score for the corner case shows the weight of the attribute that was set at its maximum level. The process is continued until all possible corner cases have been rated, each indicating the weight for a different attribute. In multiplicative models (described later), the analyst can estimate other parameters by presenting corner cases with two or more attributes at peak levels. After the expert rates several cases, a set of parameters is estimated that optimizes the fit between model predictions and expert's ratings.

Another approach is to mix and match methods. Several empirical comparisons of assessment methods have shown that different weight-estimation methods lead to similar assessments. A study that compared seven methods for obtaining subjective weights, including 100-point distribution, ranking, and ratio methods, found no differences in their results (Jia, Fischer, and Dyer 1998; Cook and Stewart 1975). Such insensitivity to assessment procedures is encouraging because it shows that the estimates are not by-products of the method and thus are more likely to reflect the expert's true opinions. This allows the substitution of one method for another.

Other Aggregation Rules: Multiplicative MAV Models

The additive value model assumes that single-attribute value scores are weighted for importance and then added together. In essence, it calculates a weighted average of single-attribute value functions.

The *multiplicative model* is another common aggregation rule. In the AIDS severity study, discussions with physicians suggested that a high score in any single-attribute value function was sufficient ground for judging the patient severely ill. Using a multiplicative model, overall severity would be calculated as

$$S = \frac{-1 + \Pi_i [1 + k \times k_i \times U(A_i)]}{k},$$

where ki and k are constants chosen so that $k = -1 + \Pi_i (1 + k \times k_i)$.

In a multiplicative model when the constant k is close to -1, high scores in one category are sufficient to produce an overall severity score even if other categories are normal. This model better resembled the expert's intuitions. The additive MAV model would have led to less severe scores due to having numerous attributes at the normal level. To construct the multiplicative value model, the expert must estimate "n + 1" parameters: the n constants k_i; and one additional parameter, the constant k.

In the AIDS severity project, the analyst constructed a multiplicative value model. On 97 hypothetical patients, the severity ratings of the multiplicative and the additive models were compared to the expert's intuitive ratings. The multiplicative model was more accurate (correlation between additive model and experts' judgment was 0.53, while the correlation between multiplicative model and expert judgment was 0.60). The difference in the accuracy of the two models was statistically significant. Therefore, the multiplicative severity model was chosen.

Resulting Multiplicative Severity Index

Appendix 2.1 is an example of a multiplicative value model. Experts on HIV/AIDS were interviewed by Alemi and his colleagues (1990), and an index was built based on their judgments. This index is intended for assessing the severity of the course of AIDS based on diagnosis and without access to physiological markers. As such, it is best suited for analysis of data from regions of the world where physiological markers are not readily available or for analysis of data from large administrative databases where

diagnoses are widely available. Kinzbrunner and Pratt (1994), as well as Alemi and his colleagues in a later article (1999), provide evaluations of this index. This index is in the public domain and can be used without royalty payments. Please note that advances in HIV/AIDS treatment may have changed the relative severity of various levels in the index.

In the multiplicative MAV model used in the Severity of the Course of AIDS index, the k value was set to −1 and all other parameters (single-attribute value functions and k_i constants) were estimated by querying a panel of experts. The scores presented in the index are the result of multiplying the single-attribute value function by its k_i coefficient. The index is scored by selecting a level within each attribute, finding the score associated with that level, multiplying all selected scores, and calculating the difference between one and the resulting multiplication.

Model Evaluation

In evaluating MAV models, it is sometimes necessary to compare model scores against experts' ratings of cases. For example, the analyst might want to see if a model makes a similar prediction on applicants for a job as a decision maker. Or the analyst might want to test if a model's score is similar to a clinician rating of severity of illness. This section describes how a model can be validated by comparing it to the expert or decision maker's judgments.

Models should be evaluated against objective data, but objective data do not always exist. In these circumstances, one can evaluate a model by comparing it against consensus among experts. A model is considered valid if it replicates the average rating of the experts and if there is consensus among experts about the ratings.

The steps in testing the ability of a model to predict an expert's rating are as follows:

1. Generate or identify cases that will be used to test the model.
2. Ask the experts to rate each case individually, discuss their differences, and rate the case again.
3. Compare the experts to each other and establish that there is consensus in ratings.
4. Compare the model scores against the average of the experts' ratings. If there is more agreement between the model and the average of the experts than among the experts, consider the model effective in simulating the experts' consensus.

Generate Cases

The first step in comparing a model to experts' rating is to have access to a large number of cases. A *case* is defined as a collection of the levels of the attributes in the model. For each attribute, one level is chosen; a case is the combination of the chosen levels. For example, a case can be constructed for judging the severity of AIDS patients by selecting a particular level for each attribute in the severity index. There are two ways for constructing cases. The first is to rely on real cases, which are organized by using the model to abstract patients or situations. The second approach is to create a hypothetical case from a combination of attributes.

Relying on hypothetical rather than real cases is generally preferable for two reasons. First, the analyst does not often have time or resources to pull together a minimum of 30 real cases. Second, attributes in real cases are positively correlated, and any model in these circumstances, even models with incorrect attribute weights, will produce ratings similar to the experts. In generating hypothetical cases, a combination of attributes, called *orthogonal design*, is used to generate cases more likely to detect differences between the model and the expert. In an orthogonal design, the best and worst of each attribute are combined in such a manner that there is no correlation between the attributes.

The test of the accuracy of a model depends in part on what cases are used. If the cases are constructed in a way that all of the attributes point to the same judgment, the test will not be very sensitive, and any model, even models with improper attribute weights, will end up predicting the cases accurately. For example, if a hypothetical applicant is described to have all of the desired features, then both the model and the decision maker will not have a difficult time accurately rating the overall value associated with the applicant. A stricter test of the model occurs only when there are conflicting attributes, one suggesting one direction and the other the opposite. When cases are constructed to resemble real situations, attributes are often correlated and point to the same conclusions. In contrast, when orthogonal design is used, attributes have zero correlation, and it is more likely to find differences between the model score and expert's judgments.

The steps for constructing orthogonal cases, also called *scenario generation*, are as follows:

1. Select two extreme levels for each attribute (best and worst).
2. Start with two to the power of number of attribute cases. For example, if there are four attributes, you would need 16 cases.
3. Divide the cases in half and assign to each half the level of the first attribute.

4. Divide the cases into quartiles and assign to each quartile the level of the second attribute.
5. Continue this process until every alternate case is assigned the best and worst levels of the last attribute.
6. Review the cases to drop those that are not possible (e.g., pregnant males).
7. If there are too many cases, ask the expert or decision maker to review a randomly chosen sample of cases.
8. Summarize each case on a separate piece of paper so that the decision maker or expert can rate the case without being overwhelmed with information from other cases.

Table 2.3 shows an orthogonal design of cases for a three attribute model.

Rate Cases

The second step in comparing model scores to expert's judgments is to ask the expert or decision maker to review each case and rate it on a scale from zero to 100, where 100 is the best (defined in terms of the task at hand) and zero is the worst (again defined in terms of task at hand). If multiple experts are available, experts can discuss the cases in which they differ and rate again. This process is known as *estimate-talk-estimate* and is an efficient method of getting experts to come to agreement on their numerical ratings. In this fashion, a behavioral consensus and not just a mathematical average can emerge.

When asking an expert to rate a case, present each case on a separate page so that information from other cases will not interfere. Table 2.4 shows an orthogonal design for cases needed to judge severity of HIV/AIDS based on three attributes: skin disease, lung disease, and GI disease.

TABLE 2.3 Orthogonal Design for Three Attributes

Scenario/Case	*Attribute 1*	*Attribute 2*	*Attribute 3*
1	Best	Best	Best
2	Best	Best	Worst
3	Best	Worst	Best
4	Best	Worst	Worst
5	Worst	Best	Best
6	Worst	Best	Worst
7	Worst	Worst	Best
8	Worst	Worst	Worst

These cases are presented one at a time. Figure 2.1 shows an example case and the question asked of the expert.

Compare Experts

In step three, if there are multiple experts, their judgments are compared to each other by looking at pairwise correlations between the experts. Two experts are in excellent agreement if the correlation between their ratings are relatively high, at least more than 0.75. For correlations from 0.50 to 0.65, experts are in moderate agreement. For correlations lower than 0.5, the experts are in low agreement. If experts are in low agreement, it is important to explore the reason why. If there is one decision maker or one expert, this step is skipped.

TABLE 2.4 Orthogonal Design for Three Attributes in Judging Severity of AIDS

Scenario/ Case	*Skin Disease*	*Lung Disease*	*GI Disease*
1	No skin disorder	No lung disorder	No GI disease
2	No skin disorder	No lung disorder	GI cancer
3	No skin disorder	Kaposi's sarcoma	No GI disease
4	No skin disorder	Kaposi's sarcoma	GI cancer
5	Thrush	No lung disorder	No GI disease
6	Thrush	No lung disorder	GI cancer
7	Thrush	Kaposi's sarcoma	No GI disease
8	Thrush	Kaposi's sarcoma	GI cancer

FIGURE 2.1 An Example of a Scenario

Case number 4:
Rated by expert: XXXX
Patient has the following conditions:

Skin disorder:	None
Lung disorder:	Kaposi's sarcoma
GI disorder:	GI cancer

On a scale from 0 to 100, where 100 is the worst prognosis (i.e., a person with less than six months to live) and 0 is the best (i.e., a person with no disorders), where would you rate this case?

First rating before consultations: _____
Second rating after consultations: _____

Compare Model to Average of Experts

In step four, the average scores of experts (in cases where there are multiple experts) or the experts' ratings (in cases where there is a single expert) are compared to the model scores. For each case, an MAV model is used to score the case. The correlation between the model score and the expert's scores is used to establish the validity of the model. This correlation should be at least as high as agreement between the experts on the same cases.

Preferential Independence

Independence has many meanings. Following are various definitions for what it means to be independent:[1]

- Not subject to control by others
- Not affiliated with a larger controlling unit
- Not requiring or relying on something else
- Not looking to others for one's opinions or for guidance in conduct
- Not bound by or committed to a political party
- Not requiring or relying on others (for care or livelihood)
- Free from the necessity of working for a living
- Showing a desire for freedom
- Not determined by or capable of being deduced or derived from or expressed in terms of members (as axioms or equations) of the set under consideration
- Having the property that the joint probability (as of events or samples) or the joint probability density function (as of random variables) equals the product of the probabilities or probability density functions of separate occurrence
- Neither deducible from nor incompatible with another statement

To these definitions should be added yet another meaning known as *preferential independence.* Preferential independence can be defined as follows:

- One attribute is preferentially independent from another if changes in shared aspects of the attribute do not affect preferences in the other attribute.
- Two attributes are mutually preferentially independent from each other if each is preferentially independent of the other.

For example, the prognosis of patients with high cholesterol levels is always worse than the prognosis of patients with low cholesterol levels

independent of shared levels of age. To test this, the expert should be asked which one of two patients has the worst prognosis:

Analyst: Let's look at two patients. Both of these patients are young. One has high cholesterol levels, and the other has low levels. Which one has the worst prognosis?

Expert: This is obvious—the person with high cholesterol levels.

Analyst: Yes, I agree it is relatively obvious, but I need to check for it. Let me now repeat the question, but this time both patients are frail elderly. Who has the worst prognosis, the one with high cholesterol or the one with low cholesterol?

Expert: If both are elderly, then my answer is the same: the one with high cholesterol.

Analyst: Great, this tells me in my terminology that cholesterol levels are preferentially independent of age.

Please note that in testing the preferential independence, the shared feature is changed but not the actual items that the client is comparing: the age for both patients is changed, but not the cholesterol levels.

Experts may say that two attributes are dependent (because they have other meanings in mind), but the attributes remain preferentially independent when the analyst checks. In many circumstances, preferential independence holds despite appearances to the contrary. However, there are occasional situations where preferential independence does not hold. Now take the previous example and add more facts in one of the attributes so that preferential independence does not hold:

Analyst: Let's look at two patients. Both of these patients are young. One has high cholesterol levels and low alcohol use. The other has high alcohol use and low cholesterol levels. Which one has worst prognosis?

Expert: Well, for a young person, alcohol abuse is a worse indicator than cholesterol levels.

Analyst: OK, now let's repeat the question, but this time both patients are frail elderly. The first patient has high cholesterol and low alcohol use. The second patient has low cholesterol and high alcohol use.

Expert: If both are elderly, I think the one with high cholesterol is at more risk. You see, for young people, I am more concerned with alcohol use; but for older people, I am more concerned with cholesterol levels.

Analyst: Great, this tells me that the combination of alcohol and cholesterol levels is not preferentially independent of age.

To assess preferential independence, a large number of comparisons need to be made, as any pair of attributes must be compared to any other attribute. Keeney and Raiffa (1976) show that if any two consecutive pairs are mutually preferentially independent from each other, then all possible pairs are mutually preferentially independent. This reduces the number of assessments necessary to only a comparison of consecutive pairs, as arranged by the analyst or the decision maker.

When preferential independence does not hold, the analyst should take this as a signal that the underlying attributes have not been fully explored. Perhaps a single attribute can be broken down into multiple attributes.

An additive or multiplicative MAV model assumes that any pair of attributes is mutually preferentially independent of a third attribute. When this assumption is not met, as in the above dialog, there is no mathematical formula that can combine single-attribute functions into an overall score that reflects the decision maker's preferences. In these circumstances, one has to build different models for each level of the attribute. For example, the analyst would need to build one model for young people, another for older people, and still another for frail elderly.

When the analyst identifies preferential independence, several different courses of actions could be followed. If the preferential dependence is not systematic or large, it could be ignored as a method of simplifying the model. On the other hand, if preferential independence is violated systematically for a few attributes, then a different model can be built for each value of the attributes. For example, in assessing risk of hospitalization, one model can be built for young people and a different model can be built for older people. Finally, one can search for a different formulation of attributes so that they are preferentially independent.

Multi-Attribute Utility Models

Utility models are value models that reflect the decision maker's risk preferences. Instead of assessing the decision maker's values directly, utility models reflect the decision maker's preferences among uncertain outcomes. Single-attribute utility functions are constructed by asking the decision maker to choose among a "sure return" and a "gamble." For example, to estimate return on investment, the decision maker should be asked to find a return that will make him indifferent to a gamble with a 50 percent chance

of maximum return and a 50 percent chance of worst-possible return. The decision maker's sure return is assigned a utility of 50. This process is continued by posing gambles involving the midpoint and the best and worst points. For example, suppose you want to estimate the utility associated with returns ranging from zero to $1000. The decision maker is asked how much of a return she is willing to take for sure to give up a 50 percent chance of making $1,000 and a 50 percent chance of making $0.

If the decision maker gives a response that is less than midway (i.e., less than $500), then the decision maker is a risk seeker. The decision maker is assigning a utility to the midway point that is higher than the expected value of returns. If the decision maker gives a response above the midway point, then the decision maker undervalues a gamble and prefers the sure return. He is risk averse. The utility he assigns to gambles is less than the expected value of the gamble; risk itself is something this decision maker is trying to avoid. If the decision maker responds with the midpoint, then she is considered to be risk neutral. A risk-neutral person is indifferent between a gamble for various returns and the expected monetary value of the gamble.

Suppose the decision maker has responded with a value of $400. Then, 50 utilities should be assigned to the return of $400. The midpoint of the scale is $500. The decision maker is a risk seeker because he assigns to the gamble a utility more than its expected value. Of course, one point does not establish risk preferences, and several points need to be estimated before one has a reasonable picture of the utility function. The analyst continues the interview to assess the utility of additional gambles. The analyst can ask for a gamble involving the midpoint and the best return. The question would be stated as follows: "How much do you need to get for sure to give up a 50 percent chance of making $400 and a 50 percent chance of making $0." Suppose the response is $175; the return is assigned a utility of 25. Similarly, the analyst can ask, "How much do you need to get for sure to give up a 50 percent chance of making $400 and a 50 percent chance of making $1,000." Suppose the response is $675; the response is assigned a utility of 75. After the utility of a few points has been estimated, it is possible to fit the points to a polynomial curve so that a utility score for all returns can be estimated. Figure 2.2 shows the resulting utility curve.

Sometimes you have to estimate a utility function over an attribute that is not continuous or that does not have a natural physical scale. In this approach, the worst and the best levels are fixed at zero and 100 utilities. The decision maker is asked to come up with a probability that would make her indifferent between a new level in the attribute and a gamble involving the worst and best possible levels in the attribute. For

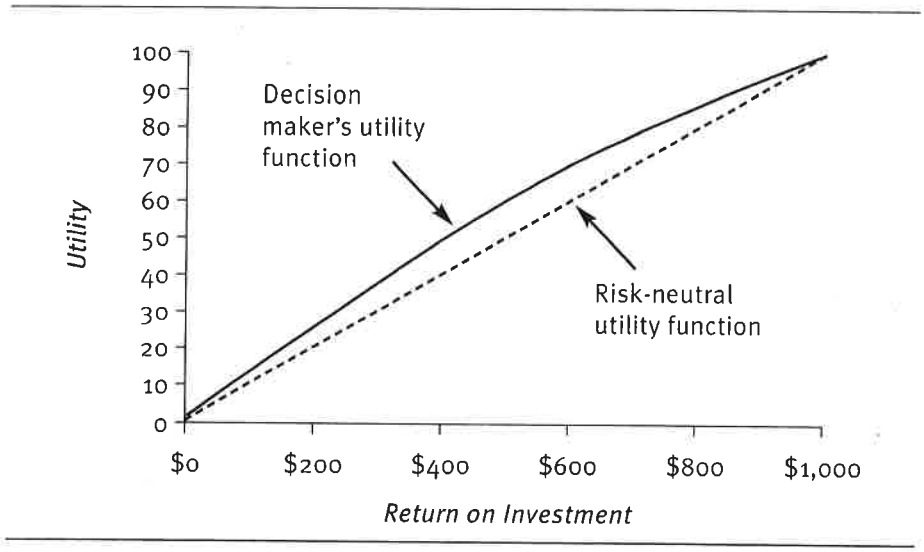

FIGURE 2.2 A Risk-Seeking Utility Function

example, suppose you want to estimate the utility (or dis-utility) associated with the following six skin conditions (listed in increasing order of severity): (1) no skin disorder, (2) Kaposi's sarcoma, (3) shingles, (4) herpes complex, (5) *candidiasis,* and (6) thrush. The analyst then assigns the best possible level a utility of zero. The worst possible level, thrush, is assigned a utility of 100. The decision maker is asked to think if she prefers to have Kaposi's sarcoma or a 90 percent chance of thrush and a 10 percent chance of having no skin disorders. Regardless of the response, the decision maker is asked the same question again but with probabilities reversed: "Do you prefer to have Kaposi's sarcoma or a 10 percent chance of thrush and a 90 percent chance of having no skin disorders?" The analyst points out to the decision maker that the choice between the sure disease and the risky situation was reversed when the probabilities were changed.

Because the choice is reversed, there must exist a probability at which point the decision maker is indifferent between the sure thing and the gamble. The probabilities are changed until the point is found where the decision maker is indifferent between having Kaposi's sarcoma and the probability P of having thrush and probability $(1 - P)$ of having no skin disorders. The utility associated with Kaposi's sarcoma is 100 times the estimated probability, P. A utility function assessed in this fashion will reflect not only the values associated with different diseases but also the decision maker's risk-taking attitude. Some decision makers may consider a sure disease radically worse than a gamble involving a chance, even though remote, of having

no diseases at all. These estimates thus reflect not only the decision makers' values but also their willingness to take risks. Value functions do not reflect risk attitudes; therefore, one would expect single-attribute value and utility functions to be different.

Hierarchical Modeling of Attributes[2]

It is sometimes helpful to introduce a hierarchical structure among the attributes, where broad categories are considered first and then, within these broad categories, weights are assigned to attributes. By convention, the weights for the broad categories add up to one, and the weight for the attributes within each category also add up to one. The final weight for an attribute is the product of the weight for its category and the weight of the attribute within the category. The following example shows the use of hierarchy in setting weights for attributes.

Chatburn and Primiano (2001) employed an additive, compensatory, multi-attribute utility model to assist the University Hospitals of Cleveland in their purchase of new ventilators for use in the hospitals' intensive care units. A decision-making model was useful in this instance because ventilators are expensive, complicated machines, and the administration and staff needed an efficient way to analyze the costs and benefits of the various purchase options.

The decision process began with an analysis of the hospitals' current ventilator situation. Many factors suggested that the purchase of new ventilators would be advantageous. First, all of the ventilators owned by the hospitals were between 12 and 16 years old, while the depreciable life span of a ventilator is only ten years. Thus, the age of the equipment put the hospitals at a greater risk to experience equipment failures. Because ventilators are used primarily for life support, the hospitals would be highly liable should this equipment fail. Second, the costs to maintain the older equipment were beginning to outweigh the initial capital investment. Third, the current fleet of ventilators varied in age and model. Some ventilators could be used only for adults, while others could only be used for infants or children, and different generations of machines ran under different operating systems. The result was that not all members of the staff were facile with every model of ventilator, yet it seemed impractical to invest in the type of extensive staff training that would be required to correct this problem. Therefore, the goals for the ventilator purchase were to advance patient care capabilities and increase staff competence, to reduce maintenance costs and staff training costs.

To begin the selection process, the consultants wanted to limit the analysis to only the most relevant choices: those machines that were designed for use in intensive care units with an ability to ventilate multiple types of patients. In addition, it was important to select a company with good customer support and the availability of software upgrades. Also, the analysis involved both a clinical and technical evaluation of each ventilator model, as well as cost analysis. Each possible ventilator was used in the hospital's units on a trial basis for 18 months so that staff could familiarize themselves with each model. The technical evaluation utilized previously published guidelines for ventilators as well as vendor-assisted simulations of various ventilator situations so that administrators and staff could compare the functionality of the different models. A checklist was used in this instance to evaluate each ventilator in three major areas: control scheme, operator interface, and alarms. Figure 2.3 depicts the attributes, their levels, and relative weights used in the final decision model.

Note that weights were first assessed across broad categories (cost, technical features, and customer service). Two of these broad categories were broken into additional attributes. Weights for broad categories were assessed; note that these weights add up to one. In addition, the weights for each attribute within the categories were also assessed; note that these weights also add up to one within the category. In the end, the model had eight attributes in total, and the weight for attributes was calculated as the product of the weight for the broad category and the weight of the attribute within that category.

Summary

In this chapter, a method is presented for modeling preferences. Often, decisions must be made through explicitly considering the priorities of decision makers. This chapter teaches the reader how to model decisions where qualitative priorities and preferences must be quantified so that an informed decision can be made. The chapter provides a rationale for modeling the values of decision makers and offers words of caution in interpreting quantitative estimates of qualitative values. The chapter concludes with examples of the use of value models, and it explains in detail the steps in modeling preferences.

The first step is determining if a model would be useful in making a particular decision. This includes identifying decision makers, objectives of the decision makers, what role subjective judgments play in the decision-making process, and if and how a value model should be employed.

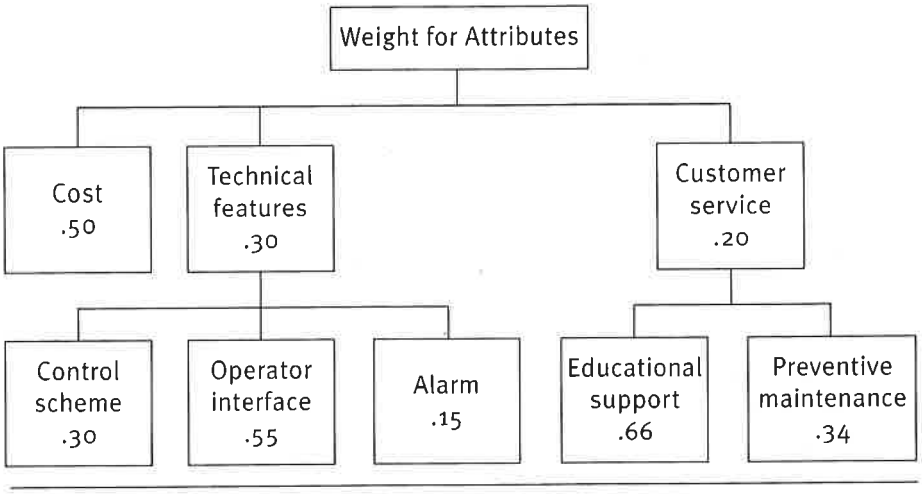

FIGURE 2.3
A Hierarchy for Assessing Attribute Weights

Next, the decision analyst must identify the attributes needed for making the judgment, and several suggestions are offered for completing this step. The third step entails narrowing the list of identified attributes to those that are the most useful. Once the attributes to be used in the decision have been finalized, the decision maker assigns values to the levels of each attribute. The next step entails the analyst determining how to aggregate single-attribute functions into an overall score evaluated across all attributes. These scores are then weighted based upon importance to the decision makers. The analyst finishes with an examination of the accuracy of the resulting decision model. The chapter concludes by providing several alternative methods for completing various steps in the process of modeling preferences.

Review What You Know

1. What are two methods for assessing a decision maker's preferences over a single attribute?
2. What are two methods for aggregating values assigned to different attributes into one overall score?
3. Make a numbered list of what to do and what not to do in selecting attributes.
4. Describe how attribute levels are solicited. In your answer, describe the process of soliciting attribute levels and not any specific list of attributes or attribute levels.

Rapid-Analysis Exercises

Construct a value function for a decision at work. Be sure to select a decision that does not involve predicting uncertain outcomes (see examples listed below). Select an expert who will help you construct the model and make an appointment to do so. Afterwards, prepare a report that answers the following questions:

1. What is the problem to be addressed? What judgment must be made, and how can the model of the judgment be useful? (Conduct research to report if similar studies have been done using MAV or utility models.)
2. Who is the decision maker?
3. What are the assumptions about the problem and its causes?
4. What objectives are being pursued by each constituency?
5. Do various constituencies have different perceptions and values?
6. What options are available?
7. What factors or attributes influence the desirability of various outcomes?
8. What values did the expert assign to each attribute and its levels?
9. How were single-attribute values aggregated to produce one overall score?
10. What is the evidence that the model is valid?
11. Is the model based on available data?
12. Did the expert consider the model simple to use?
13. Did the expert consider the model to be face valid?
14. Does the model correspond with other measures of the same concept (i.e., construct validity)?
15. Does the model simulate the experts' judgment on at least 15 cases?
16. Does the model predict any objective gold standard?

Audio/Visual Chapter Aids

To help you understand the concepts of modeling preferences, visit this book's companion web site at ache.org/DecisionAnalysis, go to Chapter 2, and view the audio/visual chapter aids.

Notes

1. *Merriam-Webster's Collegiate Dictionary*, 11th ed., s.v. "Independent."

2. This section is a summary prepared by Jennifer A. Sinkule based on Chatburn, R. L., F. P. Primiano, Jr. 2001. "Decision Analysis for Large Capital Purchases: How to Buy a Ventilator." *Respiratory Care* 46 (10): 1038–53.

References

Alemi, F., B. Turner, L. Markson, and T. Maccaron. 1990. "Severity of the Course of AIDS." *Interfaces* 21 (3): 105–6.

Alemi, F., L. Walker, J. Carey, and J. Leggett. 1999. "Validity of Three Measures of Severity of AIDS for Use in Health Services Research Studies." *Health Services Management Research* 12 (1): 45–50.

Anthony, M.K., P. F. Brennan, R. O'Brien, and N. Suwannaroop. 2004. "Measurement of Nursing Practice Models Using Multiattribute Utility Theory: Relationship to Patient and Organizational Outcomes." *Quality Management in Health Care* 13 (1): 40–52.

Bernoulli, D. 1738. "Spearman theoria novai de mensura sortus." *Comettariii Academiae Saentiarum Imperialses Petropolitica* 5:175–92. Translated by L. Somner. 1954. *Econometrica* 22:23–36.

Chapman, G. B, and E. J. Johnson. 1999. "Anchoring, Activation, and the Construction of Values." *Organizational Behavior and Human Decision Processes* 79 (2): 115–53.

Chatburn, R. L., and F. P. Primiano. 2001. "Decision Analysis for Large Capital Purchases: How to Buy a Ventilator." *Respiratory Care* 46 (10): 1038–53.

Chiou, C. F., M. R. Weaver, M. A. Bell, T. A. Lee, and J. W. Krieger. 2005. "Development of the Multi-Attribute Pediatric Asthma Health Outcome Measure (PAHOM)." *International Journal for Quality in Healthcare* 17 (1): 23–30.

Chow, C. W., K. M. Haddad, and A. Wong-Boren. 1991. "Improving Subjective Decision Making in Health Care Administration." *Hospital and Health Services Administration* 36 (2):191–210.

Cline, B., F. Alemi, and K. Bosworth 1982. "Intensive Skilled Nursing Care: A Multi-Attribute Utility Model for Level of Care Decision Making." *Journal of American Health Care Association* 8 (6): 82–87.

Cook, R. L., and T. R. Stewart. 1975. "A Comparison of Seven Methods for Obtaining Subjective Description of Judgmental Policy." *Organizational Behavior and Human Performance* 12:31–45.

Edwards, W., and F. H. Barron. 1994. "SMARTS and SMARTER: Improved Simple Methods for Multiattribute Utility Measurement." *Organizational Behavior and Human Decision Processes* 60: 306–25.

Eriksen, S., and L. R. Keller. 1993. "A Multiattribute-Utility-Function Approach to Weighing the Risks and Benefits of Pharmaceutical Agents." *Medical Decision Making* 13 (2): 118–25.

Fields, W. 1995. "Brainstorming: How to Generate Creative Ideas." *Nursing Quality Connection* 5 (3): 35.

Fos, P. J., and M. A. Zuniga. 1999. "Assessment of Primary Health Care Access Status: An Analytic Technique for Decision Making." *Health Care Management Science* 2 (4): 229–38.

Freedberg, K. A., J. A. Scharfstein, G. R. Seage, E. Losina, M. C. Weinstein, D. E. Craven, and A. D. Paltiel. 1998. "The Cost-Effectiveness of Preventing AIDS-Related Opportunistic Infections." *JAMA* 279 (2): 130–36.

Goodwin, P., and G. Wright. 2004. *Decision Analysis for Management Judgment.* 3rd ed. Hoboken, NJ: John Wiley and Sons.

Health Services Research Group, Center for Health Systems Research and Analysis, University of Wisconsin. 1975. "Development of the Index for Medical Under-service." *Health Services Research* 10 (2): 168–80.

Jia, J., G. W. Fischer, and J. S. Dyer. 1998. "Attribute Weighting Methods and Decision Quality in the Presence of Response Error: A Simulation Study." *Journal of Behavioral Decision Making* 11 (2): 85–105.

Kahneman, D. 2003. "A Perspective on Judgment and Choice: Mapping Bounded Rationality." *American Psychologist* 58 (9): 697–720.

Keeney, R. 1996. *Value-Focused Thinking: A Path to Creative Decisionmaking.* Cambridge, MA: Harvard University Press.

Keeney, R. L., and H. Raiffa. 1976. *Decisions and Multiple Objectives: Preferences and Value Tradeoffs.* New York: John Wiley and Sons.

Kim, S., D. Goldstein, L. Hasher, and R. T. Zacks. 2005. "Framing Effects in Younger and Older Adults." *Journals of Gerontology: Series B, Psychological Sciences and Social Sciences* 60 (4): P215–8.

Kinzbrunner, B., and M. M. Pratt. 1994. "Severity Index Scores Correlate with Survival of AIDS Patients." *American Journal of Hospice and Palliative Care* 11 (3): 4–9.

Krahn, M., P. Ritvo, J. Irvine, G. Tomlinson, A. Bezjak, J. Trachtenberg, and G. Naglie. 2000. "Construction of the Patient-Oriented Prostate Utility Scale (PORPUS): A Multiattribute Health State Classification System for Prostate Cancer." *Journal of Clinical Epidemiology* 53 (9): 920–30.

McNeil, B. J., S. H. Pedersen, and C. Gatsonis. 1992. "Current Issues in Profiles: Potentials and Limitations." In *Physician Payment Review Commission Conference on Profiling*, 46–70. Washington, DC: Physician Payment Review Commission.

Naglie, G., M. D. Krahn, D. Naimark, D. A. Redelmeier, and A. S. Detsky. 1997. "Primer on Medical Decision Analysis: Part 3—Estimating Probabilities and Utilities." *Medical Decision Making* 17 (2): 136–41.

Nisbett, R., and L. Ross. 1980. *Human Inferences.* Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R.E., and Y. Miyamoto. 2005. "The Influence of Culture: Holistic versus Analytic Perception." *Trends in Cognitive Sciences* 9 (10): 467–73.

Salo, A. A., and R. P. Hämäläinen. 2001. "Preference Ratios in Multi-Attribute Evaluation (PRIME)—Elicitation and Decision Procedures." *IEEE Transactions on Systems, Man, and Cybernetics* 31 (6): 533–45.

Spoth, R. 1991. "Multi-Attribute Analysis of Benefit Managers' Preferences for Smoking Cessation Programs." *Health Values* 14 (5): 3–15.

Sutton, R. I. 2001. "The Weird Rules of Creativity." *Harvard Business Review* 79 (8): 94–103, 161.

Torrance, G. W, W. Furlong, D. Feeny, and M. Boyle. 1995. "Multi-Attribute Preference Functions: Health Utilities Index." *Pharmacoeconomics* 7 (6): 503–20.

Vibbert, S. 1992. "Illinois Blues Target Doctors." *Medical Utilization Review*, April 2.

Von Winterfeldt, D., and W. Edwards. 1986. *Decision Analysis and Behavioral Research.* New York: Cambridge University Press.

Appendix 2.1: Severity of the Course of AIDS Index

Step 1: Choose the lowest score that applies to the patient's characteristics. If no exact match can be found, approximate the score by using the two markers most similar to the patient's characteristics.

Age

Less than 18 years, do not use this index

18 to 40 years, 1.0000
40 to 60 years, 0.9774
Over 60 years, 0.9436

Race

White, 1.0000
Black, 0.9525
Hispanic, 0.9525
Other, 1.0000

Defining AIDS diagnosis

Kaposi's sarcoma, 1.0000
Candida esophagitis, 0.8093
Pneumocystis carinii pneumonia, 0.8014
Toxoplasmosis, 0.7537
Cryptococcosis, 0.7338
Cytomegalovirus retinitis, 0.7259
Cryptosporidiosis, 0.7179
Dementia, 0.7140
Cytomegalovirus colitis, 0.6981
Lymphoma, 0.6981
Progressive multi-focal leukoencephalopathy, 0.6941

Mode of transmission

Blood transfusion for non-trauma, 0.9316
Drug abuse, 0.8792
Other, 1.0000

Skin disorders

No skin disorder, 1.0000
Herpes simplex, 0.8735
Kaposi's sarcoma, 1.0000
Cutaneous candidiasis, 0.8555
Shingles, 0.9036
Thrush, 0.8555

Heart disorders

No heart disorders, 1.0000
HIV cardiomyopathy, 0.7337

GI diseases

No GI disease, 1.0000
Herpes esophagitis, 0.7536
Isosporidiasis, 0.8091
Mycobacterium avium-intracellulare, 0.7494
Candida esophagitis, 0.8058
Salmonella infectum, 0.7905
Cryptosporidiosis, 0.7369
Tuberculosis, 0.7897
Kaposi's sarcoma, 0.7324
Nonspecific diarrhea, 0.7803
Cytomegalovirus colitis, 0.7086
GI cancer, 0.7060

Time since AIDS diagnosis

Less than 3 months, 1.0000
More than 3 months, 0.9841
More than 6 months, 0.9682
More than 9 months, 0.9563
More than 12 months, 0.9404
More than 15 months, 0.9245
More than 18 months, 0.9086
More than 21 months, 0.8927
More than 24 months, 0.8768
More than 36 months, 0.8172
More than 48 months, 0.7537
More than 60 months, 0.6941

Lung disorders

No lung disorders, 1.0000
Pneumonia, unspecified, 0.9208
Bacterial pneumonia, 0.8960
Tuberculosis, 0.8911
Mild *Pneumocystis carinii* pneumonia, 0.8664
Cryptococcosis, 0.8161
Herpes simplex, 0.8115
Histoplasmosis, 0.8135
Pneumocystis carinii pneumonia with respiratory failure, 0.8100
Mycobacterium avium-intracellulare, 0.8020
Kaposi's sarcoma, 0.7772

Nervous system diseases

No nervous system involvement, 1.0000
Neurosyphilis, 0.9975
Tubercular meningitis, 0.7776
Cryptoccoccal meningitis, 0.7616
Seizure, 0.7611
Myelopathy, 0.7511
Cytomegalovirus retinitis, 0.7454
Norcardiosis, 0.7454
Meningitis encephalitis unspecified, 0.7368
Histoplasmosis, 0.7264
Progressive multifocal leukoencephalopathy, 0.7213
Encephalopathy/HIV dementia, 0.7213
Coccidioidomycosis, 0.7189
Lymphoma, 0.7139

Disseminated disease

No disseminated illness, 1.0000
Idiopathic thrombocytopenic pupura, 0.9237
Kaposi's sarcoma, 0.9067
Non-*Salmonella* sepsis, 0.8163
Salmonella sepsis, 0.8043
Other drug-induced anemia, 0.7918
Varicella zoster virus, 0.7912
Tuberculosis, 0.7910
Norcardiosis, 0.7842
Non-tubercular mycobacterial disease, 0.7705
Transfusion, 0.7611
Toxoplasmosis, 0.7591
AZT drug-induced anemia, 0.7576
Cryptococcosis, 0.7555
Histoplasmosis, 0.7405
Hodgkin's disease, 0.7340
Coccidio-idomycosis, 0.7310
Cytomegalovirus, 0.7239
Non-Hodgkin's lymphoma, 0.7164
Thrombotic thrombocytopenia, 0.7139

Recurring acute illness

No, 1.0000 Yes, 0.8357

Functional impairment

No marker, 1.0000
Boarding home care, 0.7933
Hospice care, 0.7416
Home health care, 0.7655
Nursing home care, 0.7535

Psychiatric comorbidity

None, 1.0000
Psychiatric problem in psychiatric hospital, 0.8872
Psychiatric problem in medical setting, 0.8268
Severe depression, 0.8268

Drug markers

None, 1.0000
Lack of prophylaxis, 0.8756
Starting AZT on 1 gram, 0.7954
Starting and stopping of AZT, 0.7963
Dropping AZT by 1 gram, 0.7673
Incomplete treatment in herpes simplex virus, varicella-zoster virus, Mycobacterium avium-intracellulare, or Cytomegalovirus retinitis, 0.7593
Prescribed oral narcotics, 0.7512
Prescribed parenteral narcotics, 0.7192
Incomplete treatment of *Pneumocystis carinii* pneumonia, 0.7111
Incomplete treatment in toxoplasmosis, 0.7031
Incomplete treatment in cryptococcal infection, 0.6951

Organ involvement

None, 1.0000

Organ	*Failure*	*Insufficiency*	*Dysfunction*
Cerebral	0.7000	0.7240	0.7480
Liver	0.7040	0.7600	0.8720
Heart	0.7080	0.7320	0.7560
Lung	0.7120	0.7520	0.8000
Renal	0.7280	0.7920	0.8840
Adrenal	0.7640	0.8240	0.7960

Comorbidity

None, 1.0000
Hypertension, 1.0000
Alcoholism, 0.8406
Influenza, 0.9203
Legionella, 0.9402

Nutritional status

No markers, 1.0000
Antiemetic, 0.9282
Nutritional supplement, 0.7687
Payment for nutritionist, 0.7607
Lomotil®/Imodium®, 0.7447
Total parenteral nutrition, 0.7248

Step two: Multiply all selected scores and enter here:
Step three: Subtract 1 from above entry and enter here:
Step four: Divide by –0.99 and enter here: