OPERATIONS RESEARCH
Vol. 54, No. 5, September–October 2006, pp. 813–828
ISSN 0030-364X | EISSN 1526-5463 | 06 | 5405 | 0813

DOI 10.1287/opre.1060.0328
© 2006 INFORMS

OR CHRONICLE

Evaluating Academic Programs: With Applications to U.S. Graduate Decision Science Programs

Ralph L. Keeney
Fuqua School of Business, Duke University, Box 90120, Durham, North Carolina 27708-0120, keeney@duke.edu

Kelly E. See
Stern School of Business, New York University, 44 West 4th Street, New York, New York 10012-1106,
ksee@stern.nyu.edu

Detlof von Winterfeldt
School of Policy, Planning, and Development, University of Southern California, Ronald Tutor Hall, Room 310,
Los Angeles, California 90089-2902, winterfe@usc.edu

This paper describes a decision analysis methodology to evaluate academic programs. It avoids the shortcomings of the well-known evaluations of universities and academic programs produced by the public media. In addition to evaluating traditional departments and schools, the methodology is designed to evaluate interdisciplinary programs or fields that typically span many areas of a university, such as operations research, risk analysis, and decision science. We first discuss general principles of using this methodology for the evaluation of disciplinary or interdisciplinary academic programs. Next, we apply this methodology to interdisciplinary graduate decision science programs in United States universities, focusing on both prescriptive decision analysis and descriptive decision research. Finally, we suggest how the methodology might be implemented to evaluate operations research programs.

Subject classifications: academic program evaluation; decision analysis: applications; education systems: operations; utility/preference: multiattribute.
Area of review: OR Chronicle.
History: Received January 2005; revision received May 2005; accepted September 2005.

1. Background and Motivation

Over the past several years, the ranking of academic programs has become a very popular and anticipated activity. These academic programs include universities, schools within universities, and departments within schools. Two prominent examples are the *U.S. News and World Report's* rankings of colleges and graduate programs and the National Research Council's rankings of Ph.D. programs (National Research Council 1995). More specialized rankings of business school programs are provided in *Business Week* and in the *Wall Street Journal*. Other evaluations have focused on specific academic programs (e.g., Gourman 2002) or major research universities (e.g., Lombardi et al. 2003). Academic rankings take on great significance to universities, as they are widely distributed in the national media and perceived to be consequential. In particular, rankings are thought to influence donations to academic units, increase the number and quality of applicants, and indicate prestige. The emphasis on academic rankings and their associated effects is consistent with the notion that what gets measured is usually paid more attention.

Most methodologies used to evaluate universities or academic programs are fairly standard scoring and weighting techniques in which the universities or programs are first scored on a number of criteria, weights are assigned to these criteria in some fashion, and a weighted score is calculated to determine an overall ranking.

U.S. News and World Report uses several criteria to rank schools and programs, including judged quality by peer institution deans and faculty members, placement success, student selectivity, faculty resources, and faculty research activity. The criteria differ somewhat from program to program, but the *U.S. News and World Report* evaluators tend to place a high weight on assessments of program quality made by peers (25%–40%). Some of the data are subjective (e.g., the peer ratings) and some are objective (e.g., the mean GRE score of admitted students). The National Research Council (NRC) conducts an evaluation of several Ph.D. programs every five years or so, although the last evaluation by the NRC was conducted in 1995 (National Research Council 1995). Their evaluation used 20 criteria ranging from deans' judgments regarding the "Scholarly Quality of the Program Faculty" and the "Effectiveness of the Program Education," both rated on a scale of zero to five, with five being best for more quantitative measures, such as the percent of faculty publishing books and journal papers and citation records of these papers.

Most of the evaluation methodologies used in these studies suffer major flaws in both substance and process, and have been criticized by several authors (see, for example, McGuire 1995, Gater 2002, Ostriker and Kuh 2003, and Clarke 2004). Recently, at least two prestigious business schools, the Wharton School of the University of Pennsylvania and the Harvard Business School, have begun to discourage participation by students, faculty, and alumni in surveys used by public media like *Business Week* to determine rankings (Harker 2004). The main criticisms of these studies concern the selection of the criteria, the scoring of academic units on these criteria, and the assessment of the weights for the criteria, each of which will be discussed in turn below.

The selection of criteria often includes a mix of input measures (e.g., student selectivity) and output measures (e.g., placement success). As the former partially leads to the latter, this is a source of double-counting contributions. Other sources of overlap and double counting stem from the use of both overall subjective evaluations by deans (e.g., ratings of research productivity) and more objective criteria (e.g., faculty publication counts). Furthermore, in some cases the selected criteria seem to have implicit biases, such as the *Wall Street Journal* reporting of the number of companies recruiting at a business school, a measure that tends to favor larger schools in more accessible areas over smaller schools in remote places.

In addition to the problems discussed above regarding the selection of criteria, there are common limitations associated with the scoring of schools and programs on these criteria. The primary issue centers on the misuse of subjective judgments by administrators and faculty members. The use of self-reporting can easily lead to overestimation of some scores by the reporting schools or programs. Moreover, such self-reported subjective judgments are known to be fraught with biases, including the "halo effect" spurred by the university's history and status and the persistence of past perceptions of high or low performance of a school or program. To this point, Jacobs (1999) examined the effect of ascribed factors on the ranking of sociology departments in the NRC's evaluations of research doctorate programs. The author concludes that factors that should be irrelevant, such as whether a school has the word "State" in its name, do have an effect when sociologists assess departments' reputations.

Perhaps the most problematic aspect of evaluations of academic units is the weighting of the criteria. First, most of the aforementioned academic rankings provide no discussion of the assessment procedure to determine weights, yet it is well known in the decision analysis literature that weights are biased by the way in which criteria are defined and split into subcriteria, by the ranges of the alternatives on the criteria, and by the measures used for the criteria (for a summary, see von Winterfeldt 1998). Second, there is usually no documentation of who assigned the weights and how they were assigned. Regarding each of these shortcomings, there is very little, if any, justification given for the weights.

On the basis of these considerable shortcomings in current academic ranking systems, we sought to develop and apply a more rigorous methodology based on multiattribute utility analysis (Keeney and Raiffa 1976). An additional motivation for the methodology was to address the fact that current academic ranking systems were not designed to evaluate interdisciplinary programs that often span many areas of a university. Thus, one goal of our study was to extend our ranking system methodology to include such interdisciplinary programs as operations research, management science, and decision science. To illustrate this application, we evaluated 34 interdisciplinary decision science programs in the United States.

2. Methodology

A useful evaluation of academic programs should indicate both the relative strengths and weaknesses of various programs and provide an inventory of the information collected for the evaluation. As such, this evaluation provides information for many stakeholders: university administration, faculty, students and prospective students, users of university knowledge, and potential donors. For a university administration and its faculty, an academic evaluation identifies accomplishments and contributions and highlights areas for potential improvements. For students, an evaluation provides an inventory of resources at a university and indicates specific areas of expertise. For users of university knowledge, an evaluation indicates where to look for particular skills and for new employees. For donors interested in particular fields, evaluations suggest areas and academic units that could benefit from their support.

The tasks necessary to evaluate academic units or programs are the same as those necessary to conduct a decision analysis of other types of alternatives, such as consumer products or public policies (see, for example, Keeney 1992, Hammond et al. 1999, Clemen and Reilly 2001). Following guidelines from decision analysis practice, we can conduct such an evaluation in six steps:

Step 1. *Problem*. What is the unit or program type to be evaluated and what are the overall intended products of the study?

Step 2. *Alternatives*. What are the specific units or programs that will be inventoried and evaluated?

Step 3. *Criteria*. What criteria will be used to measure contributions of the alternatives?

Step 4. *Consequences*. What is the level of the contribution of each of the alternatives in terms of each of the criteria identified?

Step 5. *Trade-offs*. How should the different criteria be weighted to reflect their relative contributions to the overall evaluation?

Step 6. *Evaluation*. How should the information collected in Steps 1–5 be combined into an overall appraisal of the units or programs?

Guidelines for implementing each of these steps are discussed in detail in the following section.

3. Implementing the Methodology

General Guidelines. Any evaluation of academic programs should be logical and defensible. It needs to clearly frame the problem being addressed so that the users can appraise whether or not the study was comprehensive in including a reasonable set of alternatives and a full set of criteria. The evaluation should make substantial attempts to avoid motivational or cognitive biases and limitations due to data. Studies should try to utilize available objectively verifiable data wherever possible and clearly indicate the use of subjective judgments, when appropriate.

An implementation should be understandable to all those who wish to read or use the study. The results should be presented so that they can easily be interpreted and avoid any misinterpretations or misuses. For this reason, it is important to discuss what was done, why it was done, and how it was done. Although the importance of this information is seemingly obvious, current academic ranking systems typically do not include such explanations.

Finally, to enhance the quality of any evaluation of an academic program, it is important to involve people affiliated with the types of programs that are being evaluated. These people understand what is of value in those programs and can provide very useful information that is sometimes only privately available. Ideally, at least one individual from each of the academic units or programs being evaluated would be involved in the process of evaluating those units or programs. This involvement is discussed in more detail in the steps below.

Problem. When evaluating academic programs, the most important step is to clearly define the class of academic programs or units that are intended to be evaluated. In some cases, such as the evaluation of the universities or business schools, this is fairly straightforward because universities and business schools are clearly identified and contain organizational units. This is also true for many traditional disciplinary departments. However, the definition of the class of programs becomes much more complicated with less traditional departments, and particularly with interdisciplinary programs, which integrate concepts across different disciplines and departments. For example, biomedical engineering amalgamates elements of biology departments, medical schools, and engineering schools. Likewise, when evaluating departments or programs of "industrial engineering," how should one account for departments of "operations research and industrial engineering" or departments of "systems and industrial engineering?" Perhaps only part of a department of "operations research and industrial engineering" should count in an evaluation of industrial engineering departments. However this is done, it must be done with deliberation and care.

It is much more difficult to evaluate interdisciplinary academic programs that do not fall neatly into specific departments or units. Such programs often are defined less by a formal organizational structure and more by individual researchers and their interests, which may be pursued in various departments across a university. Furthermore, what is conceptually the same interdisciplinary program can be structured very differently in different universities. The same "program" may be within a single department at one university, across multiple departments at another university, or separate from any department in a third university. Some interdisciplinary programs, like decision science, may even span functionally different schools, such as engineering and business schools. Thus, to characterize an interdisciplinary program, it is necessary to first identify the activities and disciplines that define the interdisciplinary field in question, then identify the individuals engaged in those activities, and finally identify the contributions of those faculty to that activity.

After clearly defining the class of academic programs or units intended for evaluation, the other important aspect of defining the problem is clearly stating the purpose of the evaluation. For example, the purpose may be to compare the academic excellence of the programs. Another purpose may be to evaluate the impact of a program on society. Correctly and clearly stating the purpose of the evaluation is important for selecting and weighting criteria for the evaluation, as discussed below.

Alternatives. As discussed above, to assure that no academic unit is inadvertently left out, it is necessary to thoroughly identify all of the academic units in the class of units or academic programs to be evaluated. This is relatively straightforward when evaluating universities and schools. For departments, it is best to broadly interpret the candidates for evaluation. Drawing on the industrial engineering example above, an evaluation of "industrial engineering" departments would include, at this stage, departments of "industrial engineering and operations research." For interdisciplinary programs that are not defined by organizational structure and span many areas of a university, it is necessary to use the individuals pursuing the program activity as a basis for defining such programs. One might begin by searching through the professional societies concerned with the activities being evaluated and finding out how many faculty members at each of the universities are members. One would expect that at least one faculty member from a university with an interdisciplinary program would be affiliated with the corresponding professional society.

After a comprehensive list of all the academic units has been created, it is usually appropriate to narrow that list because many units will have few associated faculty members or do not meet other inclusion criteria. For instance, if the purpose is to appraise mainly those units that are making the largest contributions in a field, one might want to narrow the class to 25 or 50 units. In the process, it would be useful to have the assistance of individuals who are engaged in that activity at various universities, such as by having them vote for units that should be included in the evaluation.

Criteria. Selecting criteria for an evaluation follows the same procedures that have been routinely applied in decision analysis to identify objectives (Keeney 1992). When selecting the appropriate criteria for the evaluation, it is useful to involve different faculty members from various programs. To identify measures to indicate the contributions of the various alternatives in terms of those criteria, an important consideration is available data. For instance, if one criterion concerns faculty publication in academic journals, then a good measure would be the number of articles in peer-refereed journals in a specified time period. If another criterion concerns the number of students taking specific classes related to the program, then the measure might be the number of enrolled students in a given period. However, one should recognize that this measure possibly greatly benefits large universities relative to small universities. Another possible measure is the number of different course offerings that address the same type of material. A large university with four sections and 60 students each would have 240 students in a particular class, whereas a small university may have an equivalent course with only 25 students in one class. If the purpose of the study were to indicate the type of material to be learned, both universities should be evaluated equivalently on this dimension, so the number of different courses would be appropriate. If the purpose is to indicate how many people were trained, and hence facilitating the use of that material outside of the universities, perhaps the number of students would be more appropriate.

In identifying objectives and measures, it is important to avoid double counting. As noted earlier, this can often occur if one includes both inputs (e.g., student selectivity) and outputs (e.g., placement success) of an academic program. If the students coming into a university are very smart, they probably leave very smart; that alone does not mean that the university contributed much to the knowledge of those individuals. In actuality, a better indicator of the contribution of the university or academic problem would be the difference between the input and output.

Consequences. Both objective and subjective data may be necessary to describe the contributions of an academic program. Many objective data are available on the Internet or from individuals at the various universities with the academic programs of interest. For example, such sources can provide information on the number of books published, the number of students in various activities, and the number of funds donated to a university. For other, more subjective measures, individual judgments are necessary to describe the consequences. For example, current academic ranking systems commonly ask recruiters at business schools to rate the quality of the students that they interviewed on a 1–5 scale. Although information of this type might be valuable as part of the study, it comes directly from judgments of individuals and is thus potentially subject to more bias than objective data. For this reason, it is particularly important to take care in conducting surveys that gather such information and very carefully explain the process that was used, so that individuals can appropriately interpret the results. Specifically, one wants to avoid shortcomings such as sampling biases, response biases, motivational biases, and halo effects. For a general reference about sampling biases, see Saasford and Jupp (1996); for references about other biases, see Poulton (1979).

One important benefit of gathering the information to assess consequences is that the information provides a useful inventory of resources. This is particularly the case when the measures concern contributions of the faculty describing their academic output using objective measures, such as articles and books. An inventory of what recruiters said is just not as valuable as an inventory of objective data.

Trade-Offs. If the criteria are appropriately chosen, then an additive linear evaluation function is a reasonable approach to evaluate alternatives. An additive linear function provides a logical quantitative basis for an evaluation, while being accessible and understandable. It also allows for different weightings of the criteria based on the exact purposes of an evaluation.

Two aspects of this weighting are particularly important and are often done poorly in multiobjective evaluations (see Keeney 2002). Specifically, the weights that should be assigned to criteria are not simply the importance attached to the label of the criterion. Instead, weights should be explicitly related to the units and the ranges of the measures used for the criteria. For instance, it is a reasonable statement to say that one published book in an area is four times the contribution of one published article in that area. However, it does not make sense to say that books are four times more important than articles in the evaluation of an academic program. It should depend on how many books and how many articles are considered when comparing the importance of these criteria, and this needs to be made explicit.

The process by which weights are determined is also important. Weights are necessarily based on value judgments of individuals, and it is therefore appropriate to have a set of knowledgeable individuals provide them. The assessments should be done individually, with feedback to the participants to allow them to resolve any internal inconsistencies. Different individuals may feel that different weights are appropriate, and one could allow for such differences in an overall evaluation of academic programs. Subsequently, it is reasonable to aggregate the values of the various individuals by averaging, or by having a meeting to allow the participants to discuss their weights and come to an agreement about how to resolve differences.

Two specific points are worth noting. Obviously, different weights may be appropriate for different purposes. Students may place greater emphasis on teaching activities in a program, whereas users of university services and doctoral candidates may place greater emphasis on research activities. Moreover, there may be some interdependencies and synergies among criteria that are worthwhile to incorporate in the evaluation. An example might be when a particular program has both theoretical and practical aspects, and the combination of both is felt to be greater than the contributions of the two parts. In this type of situation, there are evaluation functions to accommodate this (see, for example, Keeney and Raiffa 1976).

Evaluation Functions. Once the criteria are scored and weighted, an evaluation function provides a numerical score for any given program aggregating across criteria. However, a critical decision is how to report the results. One can just report the overall numerical evaluations directly. In many of the prominent academic evaluation systems, the numerical scores are converted into a ranking, with higher overall scores having a higher rank. With this process, there is the tendency to exaggerate small differences. That is to say, the difference between a program ranked number 9 and a program ranked number 7 appears significant, when in actuality the underlying scores might not differ significantly. For this reason, most academics feel that small differences in rank are not meaningful and favor categories of contributions that aggregate academic units with similar scores. This approach is akin to the categorization of course grades with the A, B, C system, or the evaluation of restaurants by Michelin. Ultimately, the individuals conducting the study must decide on the most appropriate way to report the results to enhance their understandability and appropriate interpretation.

4. Application to the Evaluation of Decision Science Programs

In this section, we illustrate the methodology described above to evaluate “decision science” programs, which will be defined below. We selected the decision science field because it poses particular challenges due to its interdisciplinary nature, which mirrors many other types of important fields that are scattered across universities, such as operations research or risk analysis. Although there is a significant breadth and depth of decision-making resources in the universities of the United States, there has never been an inventory or appraisal of graduate decision science programs. A formal assessment of such resources can serve as the basis for identifying and recognizing the many quality graduate programs available and foster awareness of decision analysis and behavioral decision research as interesting and legitimate areas of study. Moreover, by disseminating these results directly to university administrators, and more broadly in public sources, the evaluation can increase current opportunities and future prospects for individuals in the decision sciences within universities, and externally to business and government.

Problem. The purpose of this evaluation is to appraise the societal impact of graduate decision science programs. This includes impact on the creation of new knowledge, generation of decision scientists, and improvement of decision making in society. This purpose is somewhat different and broader than that of traditional evaluations of academic programs, whose main purpose is to compare the academic quality of these programs.

We define decision science programs as (1) the collection of faculty at a university whose field of study includes a focus on how people do make decisions (“descriptive focus”) or how they should make decisions (“prescriptive focus”); and (2) the teaching, research, and service activities of those faculty members in the decision science field. This collection of decision science faculty, which typically crosses the boundaries of traditional departments and even schools, is found in business schools, engineering schools, schools of arts and sciences (e.g., economics, psychology, and statistics departments), and public policy and administration schools. Even within a single business school, faculty engaged in decision science research often are located in departments of management, decision science, marketing, operations management, and even accounting or finance.

There are two types of graduate decision science programs of interest that we refer to as “prescriptive” and “descriptive.” Prescriptive decision science programs address how people should gather information and make better decisions. Descriptive decision science programs address how people do process judgments and make decisions. Historically, both of these areas of decision science were based on the subjective expected utility (SEU) model of decision making (von Neumann and Morgenstern 1947, Savage 1954). This model breaks the evaluation of alternatives into separate parts that address the likely consequences of the various alternatives, typically formalized using probabilities, and the desirability of those possible consequences, typically formalized with utilities based on values and preferences. Prescriptive decision analysis continues to use the SEU model as a foundation. Descriptive decision research has focused on empirical deviations from this model and has developed alternatives, most notably Kahneman and Tversky’s (1979) prospect theory. In the current study, what are sometimes referred to as normative approaches to decision making are included in the prescriptive category.

We have deliberately defined the prescriptive and descriptive decision science programs somewhat narrowly to focus on the basic foundations of decision making. There are numerous categories of academic work dealing with other aspects of decision making that we do not include in our definitions or in the study. Examples of programs not included are systems analysis, decision support systems, mathematical programming, most mathematical optimization models, statistics, and multicriteria decision models, unless they are built on SEU concepts. Also, we do not include allied areas such as game theory and negotiation analysis unless they are directly tied to SEU ideas. Furthermore, individuals concerned with decision making in schools of medicine, law, and theology are not addressed in this study. Specifically, we are concerned with the fundamental ideas related to the prescriptive field of decision analysis and the descriptive field of decision research.

It is important to note that we have developed separate inventories and appraisals of the prescriptive and descriptive decision science programs. These programs are complementary, but they do address different substance matter, so we felt it was important to distinguish their contributions.

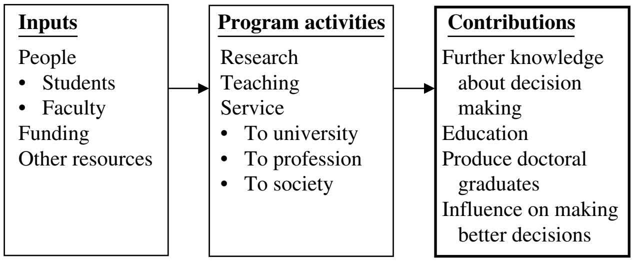
Alternatives. This study examines graduate decision science programs that are currently making important contributions to the prescriptive and descriptive decision fields. We used several steps to select these programs. First, we relied on the number of faculty from various universities in the professional societies concerned with prescriptive decision analysis and descriptive decision research to select 24 top universities. Next, a list of 10 faculty members at 10 different universities helped us select 10 additional universities in a polling process. The details are as follows.

The Decision Analysis Society is the main society concerned with prescriptive decision analysis, although many of the members also contribute to descriptive decision research. The Society for Judgment and Decision Making is the main professional society concerned with descriptive decision research, although many of its members are also interested in prescriptive decision analysis. Using membership lists from the two societies, we selected the top nine universities in terms of faculty membership in each of the societies. This provided a total of 14 universities, because four universities appeared on both lists. We then added those universities that were in the top 25 of faculty membership in both of the societies, which added six more universities. Next, we included two universities based on the number of graduate student members in each of the two societies, which brought the total to 24 universities.

At this stage, we relied on peer judgment to select additional programs for inclusion. Specifically, we sent a list of approximately 70 universities to 10 professors (five descriptively oriented and five prescriptively oriented) at 10 separate universities. These 70 universities were selected because they had at least two faculty members in the combined professional societies. Each faculty member was asked to identify between eight and 16 additional universities to include in the study. By selecting universities suggested by a majority of the 10 faculty, eight additional universities were added. As a result, 34 universities were included in the study (see the first column of Tables 2 and 3).

Criteria. To produce an inventory of resources or to evaluate the contribution of a decision science program, one needs to develop criteria and measures for those criteria. For example, one criteria of a graduate decision science program is to produce doctorates. A measure of such a contribution is the number of doctoral dissertations completed in a decision area in a given period of time.

Figure 1. Overview of the contributions of graduate decision science programs.



An overview of the contributions of graduate decision science programs is shown in Figure 1. This figure shows the input to a decision science program and the contributions of a decision science program. The main inputs are the people involved, specifically the faculty and students. Funding available is also important, as well as other resources such as facilities and support from university administration.

The program activities mainly consist of research, teaching, and service. The main beneficiaries of the research typically are other researchers and academics interested in decision areas. The main beneficiaries of the teaching are the students. The main beneficiaries of service include people in businesses, government, and individuals interested in making better decisions. There is also service provided to the university and to the profession via professional societies and organizations like the National Science Foundation.

The main contributions of a graduate decision science program are education, the production of doctoral graduates, the creation of knowledge about decision making, and the actual influence on quality of decisions. The specific criteria and measures selected for describing these contributions in the inventory and appraisal are listed in Table 1. We discuss each of these in turn.

Table 1. Criteria and measures of contributions.

Criteria	Measures
Further knowledge about decision making	Number of articles in refereed journals in 1999–2003 Number of academic books in print Number of edited books in print
Educate about decision making	Number of graduate courses offered Number of textbooks in print
Produce doctoral graduates in decision sciences	Number of decision-making dissertations completed in 1999–2003
Influence the making of better decisions	Number of popular-audience books in print

The first criterion is to "further knowledge about decision making." For each graduate program, the level of contributions on this criteria is indicated by three specific measures: the number of articles in refereed journals in the period 1999–2003, the number of scholarly academic books about decision making currently in print, and the number of scholarly edited books currently in print. There are some important implicit assumptions with each of these measures.

One assumption is that each article in a refereed journal counts the same as any other article. Also, articles developing theory or presenting an application count the same. In defense of these assumptions, it would be presumptuous as well as infeasible to determine the relative value of every article, especially when the long-term impact of articles could not be known at this time. Also, we would expect that for any graduate program, some of the articles would be exceptional, and some would just be good, so the assumption that all articles count the same may affect each of the programs equivalently. Analogous assumptions apply for scholarly books and edited volumes.

Another assumption with the three measures for research is that a particular university graduate program receives credit if they have one author affiliated with the article, academic book, or edited volume. Also, a single university program cannot get more than one credit for a particular item. Thus, two faculty members from the same university who coauthor a published article will contribute one article to that university's program using our measure. If two faculty members from different universities had coauthored that same article, each of their programs would get credit. The reasoning is that the people affiliated with the authors in each university would tend to learn about their findings more easily.

There are two measures for the criteria of "educate about decision making," namely, the number of graduate courses currently offered at the university and the number of textbooks written by faculty currently in print. The measure for courses assumes that each course is equivalent to any other course, whether it is mainly for first-year graduate students or a doctoral seminar, or whether it is a quarter or a semester long. Undergraduate courses are not counted because we are evaluating graduate programs only. Our measure for textbooks assumes that each text is equivalent to any other text. It is worth noting that a textbook, as we define it, is one that usually includes problems for students to solve or think about. Certainly, some of the books that we categorize as "scholarly books" in our evaluation might be used in graduate courses, but we maintain the description of a scholarly book rather than a textbook for the sake of clarity.

A third criterion is to "produce doctoral graduates in decision science programs." The measure selected for this is the number of dissertations with decision-making topics completed in the period 1999–2003. Doctoral graduates are important because they further the field by contributing new approaches and ideas.

The final criterion is to "influence the making of better decisions." The measure for this is the number of popular-audience books on decision making. This measure implicitly assumes that all popular-audience books are equivalent, which seems reasonable for reasons similar to those described above for academic publications.

The five-year period (1999–2003) for articles and dissertations was chosen to balance two main concerns. It is long enough to average over most of the annual fluctuations that occur in publications, and yet short enough to be a contemporary indicator of contributions. Moreover, especially with respect to professional articles, using a five-year window helps prevent the appraisal and evaluation from drastically changing from year to year. Such volatility does not accurately describe the relative contributions of academic programs, yet the public media that evaluates many academic programs exploits such volatility as it increases interest in the story on their evaluations.

Collectively, the seven measures in Table 1 cover the main contributions of decision science programs. There are other contributions that we recognize but decided not to include in this study. These include the contributions of faculty associated with a program to the professional societies in the field, to the journals in the field, to organizations such as the National Science Foundation or National Research Council that utilize expertise from individuals in the field, and from free consulting offered to public and nonprofit groups.

We also chose not to include any measures concerning either the number of students taking decision science courses or the number of master's degree graduates in decision science programs. Partly, these contributions are measured by the number of courses offered. The number of courses is a better indicator of the quality of the educational opportunity for an individual student than the number of students enrolled in courses. We also feel that there are few master's graduates whose main focus is a decision field, as compared to a more general focus (e.g., engineering or business) with a course or two in decision science.

Other potential measures not chosen had to do with the inputs to the decision science programs. We intentionally did not want to evaluate inputs such as the number of faculty, the amount of research funding, or the average scores of students on some standardized test entering a program. The main focus of this study is on outputs, not on inputs. We also chose not to include citations of literature as one of the measures. Citations do provide an indication of the quality of work and its impact, but the overall impact is not known until several years after a publication. As noted above, we chose the most recent five-year period because we wanted to use published articles to indicate the level of current contribution, as well as to minimize volatility in the annual appraisal.

Consequences. For each of the 34 graduate decision science programs, we determined how the programs contributed to the seven measures above. Using many individuals in the decision fields to help (see Keeney et al.

2004), we gathered information for all of the measures as described below.

The first step was to obtain a list of all faculty members at each of the 34 universities who were contributors to the decision science program. Professors were assigned to the universities where they had their permanent affiliation in 2002–2003, the most recent year included in our study. This seemed appropriate to interpret the evaluation of a program as its quality at the most recent time period of the appraisal. Thus, a university gets credit for recently added faculty that strengthen its decision science program. Counted articles published in earlier years, but no earlier than 1999, when the author might have been at a different university, are credited to the university where the faculty member resided at the end of 2003.

We began creating faculty lists for each university from the memberships of the two professional societies (the Decision Analysis Society and the Society for Judgment and Decision Making) and our own knowledge. We then sent this list to a professor at each university and asked them to add any faculty at their university involved in decision science programs. Note that we did not include visiting professors, emeritus professors, or postdoctoral researchers as members of the faculty. We did include adjunct professors regularly affiliated with the program.

The list of decision faculty at each university was necessary to gather professional articles and books that contribute to decision making. Hence, we intentionally tried to create faculty lists to include individuals with even a minor overlap of publications in the decision fields. As a result, many of the publications for some of the individual faculty members did not count as a decision science publication as defined in this study when they were rated later. Using the lists of publications, we could directly evaluate each publication to decide which to include and exclude (as described below) rather than indirectly evaluate all publications of an individual faculty member by including or excluding that faculty member from the original list of decision science faculty.

For each faculty member on our university lists, we conducted a search of Web of Science® and included all of the published articles in refereed journals of which they were an author or coauthor during the period 1999–2003. Web of Science® consists of five databases containing information gathered from thousands of scholarly journals in all areas of research. We restricted our search to the *Social Sciences Citation Index* (SSCI), which includes more than 1,725 journals across 50 social sciences disciplines and averages 2,900 new records per week. It is worth mentioning that the SSCI catalog does not include every publication, but it includes a large number of the peer-refereed publications in the decision science fields, e.g., *Operations Research, Management Science, Risk Analysis, Medical Decision Making, Journal of Risk and Uncertainty*, and IEEE Transactions journals. Articles in edited books and conference proceedings are not included in our lists of peer-reviewed articles.

The list of books was gathered in a similar manner. First, searches for the faculty of two universities were done at the Library of Congress website and at amazon.com. As the results were very similar, we chose to continue the search using amazon.com resources.

The initial lists of decision science courses offered by universities were gathered from university course catalogs on the Web. Additions to these lists were provided by the same individuals at each university who helped with obtaining the list of decision science faculty. Finally, these same individuals were asked to obtain a list of doctoral dissertations during the period 1999–2003. This request was difficult at some universities because there was often no central repository that categorizes dissertations by whether or not they have a decision science topic. Hence, typically the individual would send out separate e-mails to faculty in different departments (e.g., psychology, engineering, business) and asked if they could provide a list of decision science dissertations in the five-year period. In some cases, we directly e-mailed faculty in different departments and asked for dissertations completed. In addition, we augmented lists of dissertations by doing a search of the *Dissertation Abstracts* database. Moreover, for over half (24) of the universities reporting fewer than five dissertations from 1999–2003, we undertook a broad search of dissertation abstracts using the keyword "decision" in the abstract. This resulted in the addition of only 27 counted dissertations across all 34 universities.

This data collection effort provided a great deal of information. We identified over 3,700 articles, over 650 books, approximately 400 courses, and about 300 doctoral graduates. One of the more important tasks of our study was to apply a common basis to identify the items (articles, books, courses, and dissertations) that were part of and contributed to decision science programs as defined for this study. This was necessary because in collecting lists of faculty involved in decision science programs, some universities naturally treated the request more broadly than others. Rather than eliminating any of the suggested faculty as being not involved in a decision science program, we instead chose to compile articles and books of all listed faculty, then make judgment calls about individual publications.

A team of three faculty members—David E. Bell of Harvard University, James Shanteau of Kansas State University, and Detlof von Winterfeldt of the University of Southern California—went through all of the published articles and identified whether they were a prescriptive decision science contribution, a descriptive decision science contribution, or neither. There was unanimous agreement in a large majority of the cases. All articles that had either unanimous or majority (two out of three) judgments that the article was a decision science article were counted. In the two or three cases where one of the three individuals chose each of the three possibilities (prescriptive article, descriptive article, or not a decision article), the article was counted and the present authors decided whether it was prescriptive or descriptive. In cases where one evaluator said to count the article, the authors reviewed the article, including key words, and decided whether it was prescriptive, descriptive, or neither.

The examination of books, courses, and dissertations was done in a similar fashion except that the reviewers were the authors of this report. Each of the authors reviewed all the books and identified them as prescriptive, descriptive, or not a decision book. We also categorized each as an academic volume, text, edited book, or popular-audience book. The inclusion logic here was similar to those for articles with slightly more flexibility. If part of the book dealt with decision making, it was included. When we had differences of opinion about a particular book, which occurred in rare cases, we simply discussed the differences and came to a common conclusion.

The logic for courses was also essentially the same as those for articles and books. If the material covered was based on the subjected expected utility model, broadly defined, then the course was included. We also included courses that had overlap with related topics. For instance, a course that partly covered individual decision making and partly covered negotiation analysis would be included.

The criteria for including doctoral dissertations depended mainly on the topic of the dissertation. As the title of the dissertation is similar to the title of an article, the criteria that we used for articles were applied for dissertations.

Tables 2 (prescriptive contributions) and 3 (descriptive contributions) show the results of this analysis for the five-year period from 1999 to 2003. For the prescriptive decision analysis inventory, there are 152 articles, 38 books, 119 graduate decision courses, and 45 doctoral dissertations collectively for the 34 universities. For the descriptive decision research programs, the inventory is as follows: 275 articles, 43 total books, 62 graduate decision courses, and 86 doctoral dissertations. In about four cases of the over 400 articles published in 1999–2003 that were counted as decision science articles, we identified a coauthor not originally listed on our decision science faculty lists. This

Table 2. Contributions of prescriptive decision science programs.

	Measures						
Graduate programs	Articles	Academic books	Edited books	Graduate courses	Textbooks	Doctoral graduates	Popular books
Arizona State University	8	0	0	6	1	0	0
Carnegie Mellon University	14	3	2	2	0	2	1
Columbia University	4	0	1	1	1	0	1
Cornell University	0	0	0	1	0	4	2
Duke University	21	3	2	7	2	1	2
George Washington University	3	0	0	8	0	2	0
Georgia Institute of Technology	1	0	0	2	0	2	0
Harvard University	21	1	2	5	2	7	0
Johns Hopkins University	8	1	0	1	0	2	0
Massachusetts Institute of Technology	4	0	0	1	0	0	0
New York University	1	0	1	2	0	0	1
Northwestern University	2	0	0	8	0	0	0
Ohio State University	5	0	0	6	0	0	0
Princeton University	2	0	0	1	0	0	0
Stanford University	9	0	1	10	1	14	1
University of Arizona	0	0	0	4	0	0	0
University of California, Berkeley	7	0	1	4	0	0	0
University of California, Irvine	9	1	0	5	0	1	0
University of California, Los Angeles	2	0	0	7	0	1	0
University of Chicago	2	0	0	1	0	0	0
University of Colorado at Boulder	0	0	0	2	0	0	0
University of Illinois, Urbana–Champaign	5	0	0	4	0	0	0
University of Iowa	0	0	0	1	0	0	0
University of Maryland	0	0	0	1	0	1	0
University of Michigan	3	1	0	3	0	2	0
University of Minnesota	1	0	0	5	0	1	0
University of North Carolina at Chapel Hill	0	0	0	3	0	0	0
University of Oregon	1	1	0	1	0	0	0
University of Pennsylvania	12	2	3	3	0	0	0
University of Southern California	1	1	0	3	0	0	0
University of Texas at Austin	6	0	0	2	0	0	0
University of Virginia	6	0	0	4	0	2	0
University of Wisconsin–Madison	9	0	0	3	0	1	0
Yale University	1	0	0	2	0	2	0

Table 3. Contributions of descriptive decision science programs.

	Measures						
Graduate programs	Articles	Academic books	Edited books	Graduate courses	Textbooks	Doctoral dissertations	Popular books
Arizona State University	8	0	0	0	0	0	0
Carnegie Mellon University	28	2	2	3	1	6	1
Columbia University	19	1	1	6	0	3	0
Cornell University	10	0	0	4	0	3	0
Duke University	16	2	1	2	0	7	0
George Washington University	0	0	0	2	0	1	0
Georgia Institute of Technology	2	1	0	0	0	1	0
Harvard University	16	1	0	1	0	1	1
Johns Hopkins University	1	0	0	1	0	0	0
Massachusetts Institute of Technology	6	0	0	0	0	3	0
New York University	8	0	0	3	0	2	0
Northwestern University	4	0	0	2	0	1	0
Ohio State University	4	0	2	0	0	2	0
Princeton University	17	1	5	2	0	6	0
Stanford University	10	0	1	0	0	2	0
University of Arizona	8	1	1	2	0	6	0
University of California, Berkeley	26	0	1	2	0	2	0
University of California, Irvine	15	1	1	0	0	3	0
University of California, Los Angeles	13	0	0	0	0	5	0
University of Chicago	33	4	1	5	1	4	0
University of Colorado at Boulder	5	0	0	3	0	3	0
University of Illinois, Urbana–Champaign	10	0	1	2	0	2	0
University of Iowa	4	0	0	1	0	1	0
University of Maryland	13	1	0	1	0	1	0
University of Michigan	11	0	0	2	0	5	0
University of Minnesota	9	0	0	3	0	1	0
University of North Carolina at Chapel Hill	6	0	0	2	0	7	0
University of Oregon	8	0	2	1	0	2	0
University of Pennsylvania	37	2	6	2	0	5	1
University of Southern California	3	0	0	1	0	0	0
University of Texas at Austin	12	1	0	3	1	1	0
University of Virginia	0	0	0	1	0	0	0
University of Wisconsin–Madison	0	0	0	3	0	0	0
Yale University	5	0	0	2	0	0	0

finding indicates that the original faculty lists were broadly defined. We counted these coauthor publications for their respective universities.

There are two general comments about the inventories. First, there is surely some undercounting of items in the inventories. Even though the Web of Science® search identified hundreds of journals where decision science articles are published, it did not include every journal that could contain such an article. Likewise, our source for books was exhaustive enough to identify most books in print that are selling even a few copies annually. It was more difficult to get complete listings of decision courses and decision dissertations as there are no repositories of these. Importantly, any undercounting or limitations arising from the databases we used to compile information is likely to be randomly distributed across universities.

Second, we do not perceive any other systematic biases. One possible bias we considered is that some universities included many more faculty than others in their original lists. In general, we felt that the main core of decision science faculty were included in all cases. Faculty members on the periphery of the decision science field, as defined in this report, are not likely to have written books in these fields. However, they may have been an author of a decision science article. This circumstance could lead to a smaller undercounting of articles for the universities that provided the more expansive faculty lists. To reduce any such potential systematic bias, we utilized the same procedures to gather the initial inventories for all of the university programs and used a common criteria to decide which items in those initial inventories to count as decision items. These procedures led to the result that some listed faculty at each university were outside the productive group of decision science faculty, suggesting that this potential bias is not large.

Trade-Offs. We chose an additive linear function for this study:

$$\mathrm{E}(x_1, x_2, \ldots, x_7) = k_1 x_1 + k_2 x_2 + \cdots + k_7 x_7, \tag{1}$$

where $x_1 =$ number of articles, $x_2 =$ number of academic books, $x_3 =$ number of edited volumes, $x_4 =$ number of text books, $x_5 =$ number of graduate courses, $x_6 =$ number of doctoral dissertations, and $x_7 =$ number of popular-audience books. At some point, additional courses do not add much value, as a typical graduate student would take a number of decision science courses and then tend to study further on his or her own. To assess this effect, we asked six of the eight professionals listed in the following paragraph to determine the number of courses that would provide the solid core of a decision science program. The mean response was 4.67 courses, and as a result, we capped the value function for courses at five.

The weights (k_i, $i = 1, \ldots, 7$) in Equation (1) are based on value judgments of eight individuals with experience in the decision science field. The individuals were David E. Bell (Harvard University), John Butler (Ohio State University), Robin Dillon (Georgetown University), Ward Edwards (University of Southern California, emeritus), Peter C. Fishburn (ATT Bell Laboratories, retired), R. Duncan Luce (University of California, Irvine), Howard Raiffa (Harvard University, emeritus), and Detlof von Winterfeldt (University of Southern California). They include six people who have won the Frank P. Ramsey Medal for significant contributions to decision analysis, which is the highest professional award given by the Society for Decision Analysis, and two less senior professionals in the field. These people were chosen because of their familiarity with both prescriptive decision analysis and descriptive decision research.

Each individual was sent forms for use in assigning relative weights to unit contributions for each measure, either by spreading 100 points or by providing ratio weights (see von Winterfeldt and Edwards 1986). One of us (Ralph Keeney) subsequently conducted the assessments of the weights in a half-hour telephone discussion with each individual. Two individuals selected slightly different evaluations for the prescriptive and descriptive decision science programs and the other six felt that the same trade-offs were appropriate.

To better compare and aggregate the judgments, all of the assessments were converted to the "spread of 100 points" and averaged. These averages were then converted to the relative (ratio) weights on a single unit of each of the measures to be used in the linear additive evaluation function (1). In Table 4, we show the average weights expressed as multiples of the weight on "edited volumes," which received the lowest relative weight, which we assigned as one. In general, there was reasonable agreement among most of the individuals whose values were assessed (see Keeney et al. 2004).

Evaluation. The evaluation function (1), when applied to the description of the seven measures for each decision science program, provides a numerical score that indicates the overall contribution. The results of the overall evaluation are shown in Tables 5 and 6, expressed in the form of a five-star system similar to the system used in the Michelin restaurant guide. The verbal meanings of the different evaluations are as follows:

Table 4. Relative weighting factors used to evaluate decision science programs.

	Weighting factor	
Measures	Prescriptive	Descriptive
Articles	1.1	1.2
Academic books	4.1	3.8
Popular-audience books	2.2	2.2
Textbooks	4.0	4.2
Edited volumes	1.0	1.0
Graduate courses (value limited to five courses)	2.2	2.4
Doctoral graduates	1.8	1.9

Exceptional (5 stars): Recognized as a national leader for the outstanding quality of its annual contributions to education, research, and service;

Excellent (4 stars): Recognized nationally for the superior quality of its routine contributions to education, research, and service;

High Quality (3 stars): Recognized for the high quality of its regular contributions in both educational and research areas;

Quality (2 stars): Regularly makes significant contributions to either education or research;

Contributing (1 star): Makes some important educational or research contributions.

The reason for a five-star system is that it reflects the relative quality of academic programs without providing the appearance of differentiation at a level that cannot truly be supported by the available information and analysis. Numerical ranking systems of academic units stress small differences between similarly ranked programs or institutions. These differences are based on information that is not significant enough to warrant different overall ratings. Hence, using categories in our evaluation system seems much more reasonable. It also results in a less volatile evaluation than the ranking system, which can jump around from year to year based on such factors as whether a few publications at one university were in December of one year rather than January of the next year, and vice versa.

To interpret the five-star evaluation system, it is necessary to recognize that only decision science programs that were making important contributions at the graduate level were included in this study. Indeed, students can take courses in decision making at numerous other universities, and there are faculty members in the Decision Analysis Society or the Society for Judgment and Decision Making from over 100 universities. Hence, the 34 universities selected for this study all have good programs. Every university was awarded at least one star for either their prescriptive or descriptive decision science program, and a majority have at least one three-star program and some stars in each. The appropriate interpretation of a one-star program is the same as a one-star restaurant: It has something

Table 5. Evaluation of prescriptive graduate decision science programs.

☆☆☆☆☆☆	☆☆☆☆☆	
Duke University Harvard University Stanford University	Arizona State University Carnegie Mellon University University of California, Irvine University of Pennsylvania	
☆☆☆	☆☆	
George Washington University Johns Hopkins University Ohio State University University of California, Berkeley University of Michigan University of Virginia University of Wisconsin	Columbia University Cornell University Northwestern University University of California, Los Angeles University of Illinois, Urbana–Champaign University of Minnesota University of Southern California University of Texas at Austin	
☆	☆☆☆☆☆☆	Exceptional
Georgia Institute of Technology Massachusetts Institute of Technology New York University University of Arizona University of North Carolina at Chapel Hill University of Oregon Yale University	☆☆☆☆☆ ☆☆☆☆ ☆☆☆ ☆☆ ☆	Excellent High Quality Quality Contributing

Note. Programs are listed alphabetically within categories.

special. We all know that most restaurants are not awarded any stars and many of them are also very good.

We conducted some sensitivity analyses to determine how the changes in weights and consequences affect the assignment of stars to programs. Figure 2 shows the cutoff scores used to assign stars for the prescriptive evaluation. Doubling the weight on the measure "refereed journal publications" and using correspondingly increased cutoffs left essentially the same programs in each star category; only one program would have a changed number of stars. This program moves down from one star to no star, as it had no refereed journal publications. The correlation between overall ratings calculated with the two weighting schemes was 0.99, which is consistent with the literature showing a limited effect of weights on overall evaluations (see von Winterfeldt and Edwards 1986). In general, we believe that the star system is quite robust against changes in the weights and consequences. Similar results were obtained for the descriptive programs.

The change in weights had no effect on the ranks of the first seven schools and only minor effects on the ranks of the last six. However, many ranks changed in the middle range. In total, 23 of the 34 schools changed ranks, with two changes of five ranks, three changes of three ranks, nine changes of two ranks, and nine changes of one rank. Thus, ranks seem to be relatively sensitive to this weight change.

There are misinterpretations that one should avoid in examining the results of the evaluations. This evaluation took the perspective of appraising the overall contribution of graduate decision science programs to teaching, research, and service. Hence, this evaluation would likely be different from one that an individual student might use to evaluate graduate decision science programs for attending graduate school. Some students would likely be more interested in the specific courses offered, the interests of the faculty in decision making (which are partly indicated by the article titles and dissertation topics), and other concerns such as research support for graduate students and employment opportunities when graduating. Other students might be interested in both prescriptive and descriptive decision making and wish to somehow combine the two evaluations or the specific information on both programs to judge universities. Faculty examining what was best for their career might have very different weights on the measures used in this study. Specifically, faculty who are untenured would likely put a much greater weight on published articles in refereed journals than on other considerations, as it is these articles that are a major consideration in tenure decisions.

One of the goals of this appraisal of decision science programs was to provide a benchmark for current contributions. As such, it can be used to evaluate and understand changes in the future. It can also indicate ways to improve each program and provide one way to compare

Table 6. Evaluation of descriptive graduate decision science programs.

☆☆☆☆☆☆

Carnegie Mellon University
University of Chicago
University of Pennsylvania

☆☆☆☆☆

Columbia University
Duke University
Princeton University
University of California, Berkeley

☆☆☆

Cornell University
Harvard University
University of Arizona
University of California, Irvine
University of California, Los Angeles
University of Maryland
University of Michigan
University of North Carolina at Chapel Hill
University of Texas at Austin

☆☆

New York University
Stanford University
University of Colorado at Boulder
University of Illinois, Urbana–Champaign
University of Minnesota
University of Oregon

☆

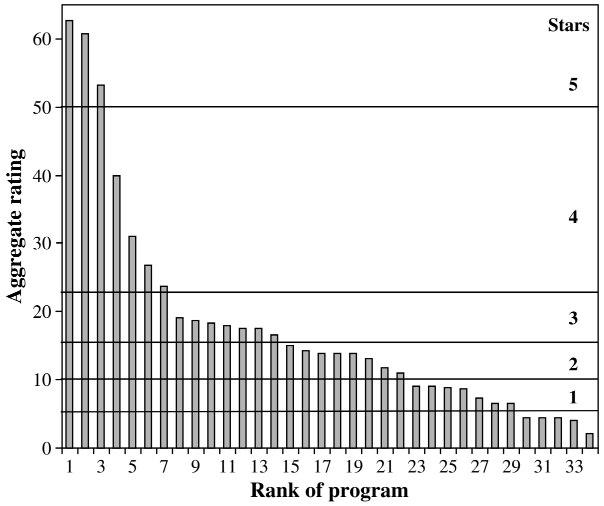
Arizona State University
Georgia Institute of Technology
Massachusetts Institute of Technology
Northwestern University
Ohio State University
University of Iowa
Yale University

☆☆☆☆☆☆	Exceptional
☆☆☆☆☆	Excellent
☆☆☆☆	High Quality
☆☆	Quality
☆	Contributing

Note. Programs are listed alphabetically within categories.

the significance of different products to that contribution. The evaluation also suggests how the infusion of additional resources of faculty, courses, and funds might make an improvement in the contribution of each program. An important ancillary benefit of conducting this evaluation is that the data gathered provide a useful repository of information. The inventories of articles, books, dissertations, and courses produced by major decision science programs in the United States are a resource for anyone interested in decision making. These inventories, as well as a detailed report on this study, can be downloaded free of charge at the website of the Decision Analysis Society at http://decision-analysis.society.informs.org.

Figure 2. Cutoffs to determine stars assigned to prescriptive programs.



5. Outline of Application for Operations Research Programs

If an evaluation of operations research programs in the United States were undertaken, the organization to support that evaluation is particularly important because of the breadth of the field. It seems appropriate that such an evaluation be sponsored by INFORMS, as it is well recognized and is the only professional society exclusively for operations researchers. Also, it is important to establish a management team that provides guidance throughout the study, as discussed in the steps below. This team should be composed of 8–12 senior operations researchers with broad experience, coverage of the different aspects of the field, and with different affiliations.

Problem. It is critical to clearly define what is meant by operations research in this step. We envision that the management team plays a key role in this, although it could certainly elicit the help of others. Essentially, this group needs to define the boundaries of what they consider to be operations research for the evaluation, including a careful delineation of related areas such as systems analysis, industrial engineering, and decision sciences.

One way to characterize the material of interest might be that which is appropriate for publication in our flagship journals, *Operations Research* and *Management Science*. That does not mean that the material would have to appear there, but the areas in those journals might characterize our definition of operations research for the evaluation. This definition of operations research would have significant impact on what would be an appropriate list of criteria for evaluating such programs, which we consider later.

Alternatives. In this step, the operations research programs that will be evaluated are identified. One key issue in defining programs at a university involves characterizing exactly what a program is. For operations research, it seems reasonable that the same approach taken in the decision science evaluation is probably appropriate. One first characterizes the set of people at a university doing operations research (as defined above). Then, the operations research program in the university would be the research and teaching in operations research done by those faculty anywhere in the university. Other approaches, such as requiring a department of operations research, would likely be inadequate because many schools with good programs in operations research do not have departments. In some cases they are centered within a department; in other cases they are a part of a department or an interdisciplinary center not included in the department.

To create a very large initial set, it may be useful to define all programs as those that have at least one faculty member who is a member of INFORMS. To select a final list of operations research programs from this large set, the management team would again have a key role. As a first cut, it seems reasonable to select those programs that have a minimum specified number of faculty members that are members of INFORMS. Then, using the individual judgments of members of the management team, one could select programs at additional universities. However it is done, it is important that the selection method be logical, clear, and transparent.

An important element for the inclusion of the university is that at least one operations researcher of that university be designated to help provide information to describe the contributions of the operations research program at his or her university. This becomes particularly important when one is trying to get a list of classes of operations research taught, the numbers of students taking them, or dissertations in operations research.

Criteria and Measures. The identification of the criteria and measure to evaluate programs is obviously critical. For this role, the management team should provide input about the appropriate criteria that relate to the definition of an operations research program provided earlier. It would likely be the case that the operations research program should be evaluated for both master's-level students and doctoral students, so there should be separate criteria that addresses each of these. Furthermore, because of the breadth of operations research, one may wish to have separate criteria and measures for aspects of the field, such as mathematical programming, decision analysis, applications, transportation, information systems, and so forth. The team should also consider criteria that concern separately the areas of teaching, research, service to professional societies and government panels, and impact of the program on the real world.

Measures should be selected for the criteria that provide a clear and logical indication of the degree to which the corresponding criteria is achieved. Details on the selection of appropriate measures are discussed in Keeney and Gregory (2005).

Consequences. This task involves describing the contributions of all of the operations research programs with respect to each of the objectives. This can be very time consuming. The management team would have responsibility for identifying how the data and information should be collected. Some of the information can be collected from the Internet. This is especially the case for lists of publications in operations research from faculty at the various programs. Also, information may be needed on operations research courses at the university, on dissertations, and on the contributions of service of societies by faculty at those universities. This information is likely much easier to obtain from an individual at the university.

Collecting the data should be done in two steps. The first is to obtain all of the information that may be appropriate to measure criteria. It will naturally be the case that individuals providing information to different schools will use a different basis for inclusion. Hence, they should be given a characterization of operations research to use in collecting their information and be told to broadly interpret that characterization to include any contributions that may be appropriate. Subsequently, the management team, or some group appointed by them, needs to establish common "cutoffs" for the information that is acceptable for each of the criteria. The purpose here is to have fair and comparable evaluations of the different programs.

Some of the criteria require the use of a constructed measure. For instance, this might be the case if one of the criteria chosen concerns the reputation of the program with a certain group, such as university provosts. Then, a survey instrument would need to be created to gather information from those provosts. Again, such surveys are often used by operations researchers and should follow typical guidelines. It is particularly important in these cases that the methods of collecting the information are transparent and reliable.

Trade-Offs. With an appropriate selection of criteria, an additive linear evaluation function is likely to be appropriate. However, some tests of the additivity assumptions should be conducted (see Keeney and Raiffa 1976). Determining weights for the various criteria is critical. We would suggest that this task be performed by the individuals on the management team, or by members of some other group appointed by the management team. It might be possible that this evaluation be done by a voting procedure if the group for evaluation was chosen to be large, such as the group of current Fellows of INFORMS.

However the evaluation is conducted, it would be useful to have input from several individuals and then average the weights provided by them. If the evaluation group is small, say under 12, it would be desirable to have a meeting to try to select a set of weights the group deemed appropriate for the evaluation. Alternatively, the team might deem two or more evaluations of different facets of programs as important to distinguish. For instance, it may be appropriate to select one set of weights for evaluating programs for master's-level students and another set of weights for doctoral programs. Whether or not they are appropriate relates back to the defined purpose of the evaluation at the beginning of the study.

Evaluation. Given the consequences and trade-offs, one can calculate an overall contribution of the program given the evaluation function. At this stage, one needs to carefully think about what should be reported about the evaluation. Simply reporting the numbers from the evaluation or the rankings implied by those evaluations is perhaps more interesting to some, but it does not fully characterize the reality of operations research programs. A program with a slightly lower evaluation than another is probably more appropriately considered to be in the same category, rather than to have a different rank. The management team should think about the pros and cons of categorization and decide which to report.

An advantage to the categorization approach is that some university programs will be much better on some and other programs would be better on others. Different programs that contribute roughly the same overall should be categorized the same.

As part of an overall evaluation, one should also indicate evaluations on the separate subsets of criteria. For example, one may wish to report evaluations in terms of total student output, total research output, total teaching output, and total service. Once one has all the information describing the consequences, there are many different ways that one can decide to report it. It might be helpful to the different operations research programs and useful to promote these programs outside of the operations research community.

6. Summary and Conclusions

There are several different evaluations of academic programs that are periodically reported. Many of these are conducted by the media (*Business Week, US News and World Report*, and the *Wall Street Journal*). The individuals conducting these studies do not appear to have knowledge of research on evaluation systems involving multiple objectives. The logic and procedures used to select the criteria, gather the data, and assign weights to the different criteria are often not done in a scientifically valid manner, nor are the procedures clearly reported. The most prominent academic evaluations report the results in the form of rankings, which change quite frequently and thus are more "newsworthy," even if the changes are not really significant.

In spite of their shortcomings, current academic rankings have a great deal of impact on issues that are critical to the functioning of universities, such as financial contributions and applications for admissions. While it is understandable that academic evaluations are highly valued and beneficial, such evaluations also create significant incentives for the academic programs being evaluated to take actions to enhance their evaluation. For example, media studies evaluating universities often include average SAT scores of the freshman class entering in the fall. Universities who would like certain students without high scores sometimes admit them for the spring quarter or as transfer students, and those scores are not included in the fall average SAT evaluation. Another criterion of university evaluations is the percentage of admitted students accepting the offer to attend. Universities have pushed or tried hard to ascertain whether students would accept the invitation to attend the university before they send the official admission letter.

The methodology outlined in this paper for evaluating academic programs is based on sound multiobjective decision analysis principles. The procedures followed are thorough in identifying criteria, gathering information, and assigning weights. Furthermore, the procedure clearly documents the implementation of the methodology for others to appraise. The results categorize the contributions of academic programs into groups with similar contributions. This tends to unify a field and lessens the divisive nature of rankings.

The method outlined has other strengths as well. First, it requires and benefits from the involvement of individuals in the academic programs being evaluated at each of the universities being evaluated. Second, the methodology requires a precise definition of the programs being evaluated. Third, unlike all of the media studies, this method can be used to evaluate interdisciplinary programs, which have not been evaluated prior to now, in a comprehensive manner. Finally, a benefit of the evaluation is a rather comprehensive inventory of contributions in the field, which can be of service to many different individuals and groups.

By identifying the contributions of the various programs and allowing individuals in those programs to compare their contributions to others, one would hope that ideas to improve various programs will be generated. News about the evaluation of the programs and their contributions will increase the understanding outside of those in the program about the field, and promote opportunities for those in the field.

Acknowledgments

This study was sponsored by the Decision Analysis Society, a professional society that is part of the Institute for Operations Research and the Management Sciences (INFORMS). Numerous people in the prescriptive decision analysis and descriptive decision research fields helped with this study. Ten faculty members (five from the descriptive area and five from the prescriptive area) helped select the graduate programs to be included in this study. Each of the 34 universities was represented by a faculty contact. Eight decision analysis professionals provided value judgments for the evaluation function. In addition, David Bell of Harvard University and James Shanteau of Kansas State University helped categorize all of the over 3,700 articles that we found on Web of Science® as prescriptive or descriptive or neither. Kate McCormick, an undergraduate research assistant at Duke University working with Kelly See, compiled the original lists of articles for each of the universities. Renay Cleary of the University of Washington, working with Ralph Keeney, compiled all of the original lists of books for the universities. The authors and the Decision Analysis Society owe a debt of gratitude to all of these individuals. For a complete listing of all individuals who helped with this study, see Keeney et al. (2004).

Detlof von Winterfeldt was partially supported by the United States Department of Homeland Security through the Center for Risk and Economic Analysis of Terrorism Events (CREATE) grant number EMW-2004-GR-0112. Any opinions, findings, and conclusions or recommendations in this paper are those of the authors and do not necessarily reflect views of the U.S. Department of Homeland Security.

References

Clarke, M. 2004. Weighting things up: A closer look at the *U.S. News and World Report*'s ranking formulas. *College Univ. J.* **79** 3–9.

Clemen, R. T., T. Reilly. 2001. *Making Hard Decisions with Decision Tools*. Duxbury Press, Pacific Grove, CA.

Gater, D. S. 2002. A review of measures in *U.S. News and World Report*'s "America's Best Colleges." Lombardi Program Occasional Paper, Center for Studies in the Humanities and Social Sciences, University of Florida, Gainesville, FL.

Gourman, J. 2002. *Gourman Report: Graduate Programs*, 10th ed. Princeton Review Publishing, New York.

Hammond, J. S., R. L. Keeney, H. Raiffa. 1999. *Smart Choices: A Practical Guide to Making Better Decisions*. Harvard Business School Press, Boston, MA.

Harker, P. 2004. Wharton on the rankings: We can't have it both ways. *eNewsline*, Wharton School, University of Pennsylvania, Philadelphia, PA. http://www.wharton.upenn.edu/mbaexecutive/rankings.

Jacobs, D. 1999. Ascription or productivity? The determinants of departmental success in the NRC quality rankings. *Soc. Sci. Res.* **28** 228–239.

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

Keeney, R. L. 1992. *Value-Focused Thinking*. Harvard University Press, Cambridge, MA.

Keeney, R. L. 2002. Common mistakes in making value trade-offs. *Oper. Res.* **50** 935–945.

Keeney, R. L., R. S. Gregory. 2005. Selecting attributes to measure the achievement of objectives. *Oper. Res.* **53** 1–11.

Keeney, R. L., H. Raiffa. 1976. *Decisions with Multiple Objectives*. Wiley, New York. Reprinted by Cambridge University Press, Cambridge, UK, 1993.

Keeney, R. L., K. E. See, D. von Winterfeldt. 2004. Inventory and appraisal of U.S. graduate decision programs. Technical report, Fuqua School of Business, Duke University, Durham, NC. http://faculty.fuqua.duke.edu/daweb.

Lombardi, J. V., E. D. Capaldi, K. R. Reeves, D. D. Craig, D. S. Gater, D. Rivers. 2003. The top American research universities. Lombardi Program on Measuring University Performance, University of Florida, Gainesville, FL. http://thecenter.ufl.edu/research2003.pdf.

McGuire, M. D. 1995. Validity issues for reputational studies. *New Directions for Institutional Res.* **88** 45–60.

National Research Council. 1995. *Research Doctoral Programs in the United States: Continuity and Change*. National Academy Press, Washington, D.C.

Ostriker, J. P., C. V. Kuh. 2003. *Assessing Research-Doctorate Programs: A Methodology Study*. National Academy Press, Washington, D.C.

Poulton, E. C. 1979. Model of biases in judging sensory magnitudes. *Psych. Bull.* **86** 777–803.

Saasford, R., V. Jupp, eds. 1996. *Data Collection and Analysis*. Sage, Thousand Oaks, CA.

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

von Neumann, J., O. Morgenstern. 1947. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ.

von Winterfeldt, D. 1998. On the relevance of behavioral decision research for the practice of decision analysis. J. Shanteau, B. Mellers, D. Schum, eds. *Decision Research from Bayesian Approaches to Normative Systems: Reflections on the Contributions of Ward Edwards*. Kluwer, Norwell, MA, 63–92.

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