The Evaluation of Research Papers (Or How to Get an Academic Committee to Agree on Something)

MATTHEW J. LIBERATORE
College of Commerce and Finance
Villanova University
Villanova, Pennsylvania 19085

ROBERT L. NYDICK
College of Commerce and Finance
Villanova University

PETER M. SANCHEZ
College of Commerce and Finance
Villanova University

The College of Commerce and Finance at Villanova University sponsors an annual research awards program. Its objectives are to encourage and reward faculty research efforts. The process of selecting papers to win the award turned out to be far more complex than we initially expected. After many unsuccessful attempts at developing a manageable procedure, we used the analytic hierarchy process to structure the selection process. We have used this approach for the last three years without any serious problems. The faculty and external advisors using the system believe that it can better capture the subjective judgments of the evaluators than any of the previous approaches.

In the fall of 1983, the College of Commerce and Finance at Villanova University instituted a research awards program with the purpose of recognizing outstanding research efforts and contributing to an overall research atmosphere within the college. The awards program was developed and continues to be administered by the college's research committee with the assistance of an external advisory board drawn from the local business community. Each member of the research committee and the advisory board provides input in the selection process. The research committee is responsible for recommending award-winning research papers to the dean, who in turn makes the final selections. The college awards two or three cash prizes each year to the authors of the outstanding research papers.

0091-2102/92/2202/0092$01.25
This paper was refereed.

DECISION ANALYSIS—SYSTEMS

INTERFACES **22:** 2 March–April 1992 (pp. 92–100)

Since the research committee has many other duties, it wishes to carry out the process as efficiently as possible. The members believe that, although the awards program is important, it should not consume an inordinate amount of the committee's time. In spite of previous efforts to implement this objective, however, unforeseen circumstances (and academics' innate desire for protracted discussion) made the process very complex.

Shortly after the program began, several problems became apparent. First, evaluating a broad variety of research papers ranging from the theoretical to the empirical and encompassing the full spectrum of business disciplines proved to be quite formidable. Various attempts at producing evaluation instruments were unsuccessful. Over a four-year period, the number of evaluative criteria contained in the instrument ranged between five and 15. Each year, the committee modified the instrument to some extent with mixed results. The final rankings of the papers were often so nebulous that the committee decided on awards through a long process of qualitative reasoning and discussion. The members usually felt they had taken far more time than necessary. Although the committee was confident that it chose the appropriate papers to win awards each year, it also believed that it could develop more efficient ways of accomplishing the task.

To develop an efficient method, the committee needed to determine a set of generic criteria applicable to many types of research papers. Whatever evaluative criteria it used, it also needed an accompanying weighting mechanism to reflect importance of the criteria. While the members of the committee recognized that evaluative criteria did vary in importance, they struggled to decide on a weighting method.

The committee tried a few different scoring models; however, none were acceptable to all of the members. To maintain a sense of collegiality, they desperately needed a new system. The chairperson of the research committee (and third author) and a research committee member (and first author) decided to embark on this adventure. Later, an additional individual (and second author) was shanghaied into participating in later modifications and implementation. The new system was to be a cure for a common refrain echoed in the hallowed halls by the various committee members, "We have to stop meeting like this!"

The New System

Our objective was to develop a new and improved evaluation system that the committee would find similar in concept and application to the scoring model it currently used. An important benefit of this system should be its formal approach for achieving consensus on the various evaluation factors and their influence on the final ranking decisions. This, in turn, should allow the committee members to reach agreement with much less effort. Because of the multi-criteria nature of the problem and the need to easily elicit the judgments of the committee members, we decided to use an analytic-hierarchy-process-based approach.

The AHP, developed by Saaty [1980], is a decision-making method for prioritizing alternatives when multiple criteria must be considered. The approach allows the decision maker to structure complex problems

in the form of a hierarchy or a set of integrated levels. Generally, the hierarchy has at least three levels. Two standard approaches for representing the hierarchy follow:

Approach 1

(1) The goal,
(2) The criteria, and
(3) The alternatives.

Approach 2

(1) The goal,
(2) The criteria,
(3) The ratings scale, and
(4) The alternatives.

The hierarchy lends itself to an analysis based on the impact of a given level on the next higher level. The process begins by determining the relative importance of the criteria in meeting the goal. In Approach 2, one must also determine the relative importance of the ratings categories for each of the criteria. Next, the focus shifts to measuring the extent to which the alternatives achieve each of the criteria. Finally, the results of the analyses are synthesized to compute the relative importance of the alternatives in meeting the goal.

The AHP has been extensively described and debated elsewhere, for example, the recent set of articles in *Management Science* (Dyer [1990a, 1990b], Harker and Vargas [1990], Saaty [1990], Winkler [1990]). Although comparisons with utility theory and methodological issues, such as rank reversal, are important, one should recognize that the AHP has been widely and successfully applied in practice [Zahedi 1986].

Our implementation links the AHP software package *Expert Choice* [Forman et al. 1990] with a Lotus 1-2-3 based scoring spreadsheet to rate the research papers. The basic approach was developed and applied by one of the authors to evaluate research and development projects [Liberatore 1987].

The committee developed and implemented the new system in four phases:

(1) We established definitions of criteria.
(2) We established consensus weights for the criteria themselves and the rating scale associated with each criterion.
(3) We read, critically reviewed, and rated each paper according to the established criteria.
(4) We reviewed the results and discussed the consistency of the preference rankings and the final scores for the papers.

For the first phase, we had to find a way to consider the many different types of papers that can be submitted. Papers could be discipline specific, could vary according to methodology (statistical, mathematical, case, survey, and so forth), and could vary according to target audience (academic or practitioner). It was important to identify and define criteria that can be used to evaluate many different types of papers. After considerable discussion, we decided that we needed a small, clearly defined, and independent set of criteria. The committee agreed on five evaluation criteria:

— Objectives: Are the objectives of the research clear? Is the intended purpose clear?

We realized that academic committees can rarely achieve consensus on anything expeditiously.

— Justification: Is a clear rationale presented for the research? Is the research positioned in light of existing knowledge in the area? Is it clear how the research will extend the body of knowledge in this area?
— Design: Is the research design (methodology) appropriate for the topic? Is the research design adequate to reach the intended objectives? Would other approaches have been better?
— Execution-implementation: Has the research design been adequately implemented? Were research procedures executed in scientific fashion? Are there any shortcomings that might have compromised the research?
— Recommendations and implications: Do recommendations flow logically from the research results? Are future directions for research adequately specified? Are implications developed so as to adequately place results in perspective?

These criteria can be used to distinguish the different phases or components of most research papers. (For the most part, business law remained a statistical outlier, although we did try to accommodate its needs.)

During the second phase we asked each committee member to judge the relative importance of the five criteria in ranking any research paper. We provided the committee members with a brief explanation of the task at hand and the standard one-to-nine AHP scale, where 1 means equally preferred, 3 means moderately preferred, 5 means strongly preferred, 7 means very strongly preferred, and 9 means extremely preferred [Saaty 1980]. Since there are five criteria, only 10 judgments ($n(n-1)/2$) are required from each committee member (25 = ($n*n$), if the committee is composed entirely of economists!). Figure 1 shows the form we used to elicit judgments from the committee members about each paper.

If the number of papers to be evaluated is small (seven or less), the papers can be judged with respect to each criteria (Approach 1). However, when the number of papers is large, such methods are generally computationally infeasible. For example, for 20 papers, $n(n-1)/2 = 190$ judgments are required for each of the five criteria. The explosion in the number of required comparisons is a disadvantage of the basic AHP approach. Fortunately, other methods can be used to reduce the number of judgments required.

Liberatore [1987] used a simple method based on Approach 2. Associated with each of the five criteria is a five-point ratings scale (that is, outstanding (O), good (G), average (A), fair (F), and poor (P)) which was used to rate each paper according to each criterion. In order to maintain the ratio scale of measurement throughout the hierarchy, one must obtain judgments to determine the relative importance of each of the five ratings categories. One can then establish ratio-scale weights for a given rating category under each evaluation criteria. A potential complication arises if, for example, the relative value of an "outstanding" versus a "good" rating

To maintain a sense of collegiality, they desperately needed a new system.

For each comparison, evaluate the relative importance of the options by placing a number next to the preferred option.

Example 1: If OBJECTIVE is **Strongly Preferred** or **Strongly More Important** than JUSTIFICATION, then:

__5__ **Objective** as compared to **Justification** _____

Example 2: If JUSTIFICATION is **Strongly Preferred** or **Strongly More Important** than OBJECTIVE, then:

_____ **Objective** as compared to **Justification** __5__

_____ **OBJECTIVE** as compared to **JUSTIFICATION** _____

_____ **OBJECTIVE** as compared to **DESIGN** _____

_____ **OBJECTIVE** as compared to **EXECUTION** _____

_____ **OBJECTIVE** as compared to **RECOMMENDATIONS** _____

_____ **JUSTIFICATION** as compared to **DESIGN** _____

_____ **JUSTIFICATION** as compared to **EXECUTION** _____

_____ **JUSTIFICATION** as compared to **RECOMMENDATIONS** _____

_____ **DESIGN** as compared to **EXECUTION** _____

_____ **DESIGN** as compared to **RECOMMENDATIONS** _____

_____ **EXECUTION** as compared to **RECOMMENDATIONS**

Figure 1: Each evaluator provided judgments concerning the relative importance of the various criteria in judging each research paper. Each judge used the standard AHP scale, where 1 means equally preferred, 3 means moderately preferred, 5 means strongly preferred, 7 means very strongly preferred, and 9 means extremely preferred.

differs for different criteria. Because we thought making such fine discriminations in judgment would be very difficult and because we wanted to keep the process as simple as possible, we asked for only one set of ratings-scale judgments (KISS applies in academia, too!). We used a form similar to that shown in Figure 1 to elicit ratings-scale judgments from the committee members.

We needed a method for combining the judgments of all committee members concerning both the criteria weights and the ratings categories. One approach would be to review all of the judgments generated and then seek a consensus. However, we realized that academic committees can rarely achieve consensus on anything expeditiously. We note that in the few cases when this does occur, the dean or other administrator usually "modifies" (a euphemism for rejects) the decision or recommendation. Therefore, we elected to achieve consensus mathematically by computing the geometric mean of each of the judgments (Aczel and Saaty [1983] give a justification of this method). This can be done automatically using *Expert Choice*. The resulting inconsistencies in the consensus judgments were well within the standard 0.10 acceptance limit.

The AHP hierarchy for this problem (Figure 2) has three levels:

— The goal, the evaluation of research papers;

— The criteria; and

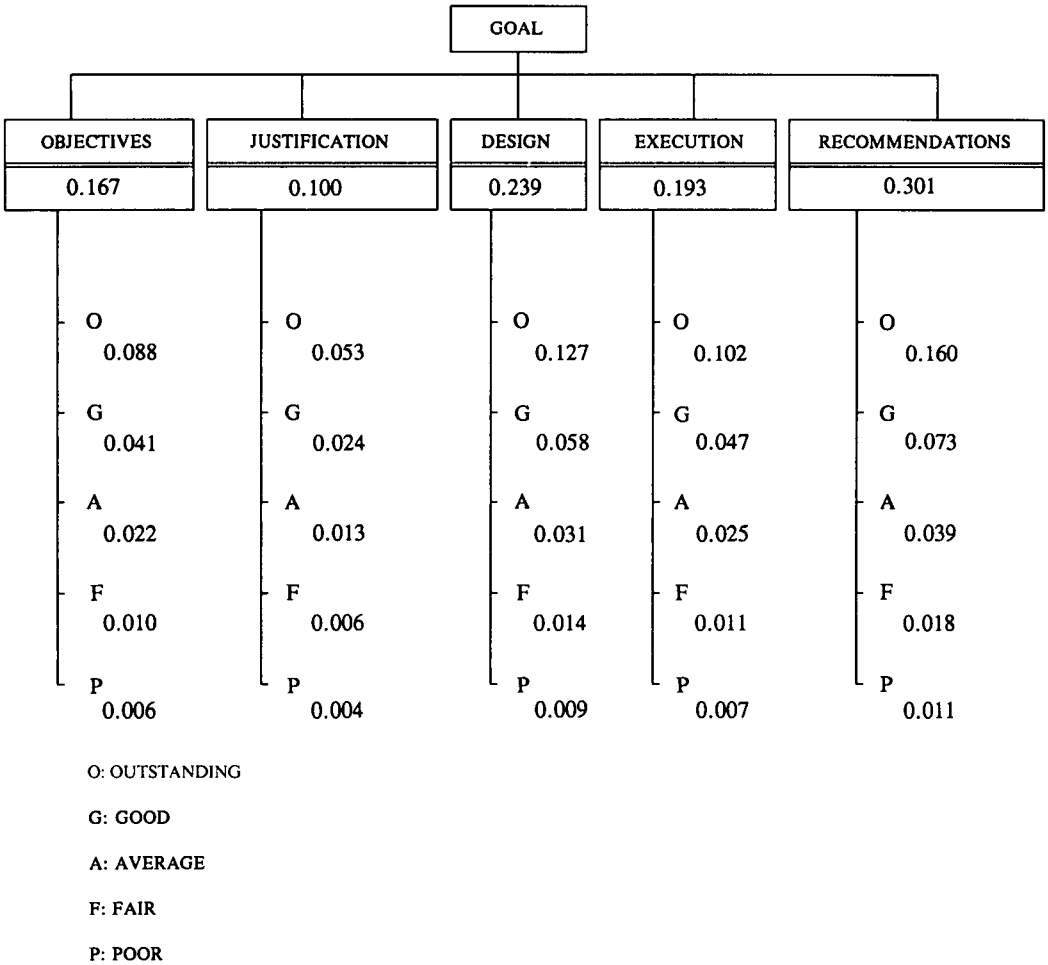

Figure 2: The goal of the AHP hierarchy was to develop the best ranking of all of the research papers considered. The hierarchy shows the five criteria selected by the committee and their consensus weights. Individual papers are rated using the five-point scale (outstanding through poor) shown under each criterion. The weights for the ratings scale shown were also generated by consensus.

— The five-point ratings scale, that is, O, G, A, F, and P.

The fourth level, namely the papers to be evaluated, does not appear in the AHP hierarchy. This occurs because the papers are not being pairwise-compared with respect to each criteria, but rated instead on the five-point scale.

We used only one set of "local" weights for the five-point ratings scale for all five criteria: 0.53, 0.24, 0.13, 0.06, and 0.04 for the five ratings in this example. Figure 2 shows the "global" weights for these categories, which are the local weights scaled by the criteria weights shown. The criteria, the ratings scale, and the associated weights are transferred to a Lotus 1-2-3 spreadsheet program so that each judge can select a rating for each evaluation criteria for each paper.

Typically, 10 to 15 papers have been entered in the competition annually, and

they are judged by approximately 12 judges (both external and internal). To illustrate the process and protect the innocent, we will present some hypothetical data for a simplified example. We will assume that three papers are to be evaluated by five judges.

We developed a standard scoring sheet to capture the judgments of each committee member for each paper. We made the

Criteria Rating	Papers						Weights
	1		2		3		
OBJECTIVES							0.167
OUTSTANDING		0.000	1	0.088		0.000	0.088
GOOD		0.000		0.000		0.000	0.041
AVERAGE		0.000		0.000		0.000	0.022
FAIR	1	0.010		0.000	1	0.010	0.010
POOR		0.000		0.000		0.000	0.006
JUSTIFICATION							0.100
OUTSTANDING		0.000		0.000		0.000	0.053
GOOD	1	0.024		0.000	1	0.024	0.024
AVERAGE		0.000	1	0.013		0.000	0.013
FAIR		0.000		0.000		0.000	0.006
POOR		0.000		0.000		0.000	0.004
DESIGN							0.239
OUTSTANDING		0.000	1	0.127		0.000	0.127
GOOD	1	0.058		0.000		0.000	0.058
AVERAGE		0.000		0.000	1	0.031	0.031
FAIR		0.000		0.000		0.000	0.014
POOR		0.000		0.000		0.000	0.009
EXECUTION-IMPLEMENTATION							0.193
OUTSTANDING		0.000		0.000		0.000	0.102
GOOD		0.000	1	0.047		0.000	0.047
AVERAGE		0.000		0.000		0.000	0.025
FAIR	1	0.011		0.000	1	0.011	0.011
POOR		0.000		0.000		0.000	0.007
RECOMMENDATIONS AND IMPLICATIONS							0.301
OUTSTANDING		0.000		0.000	1	0.160	0.160
GOOD	1	0.073		0.000		0.000	0.073
AVERAGE		0.000		0.000		0.000	0.039
FAIR		0.000	1	0.018		0.000	0.018
POOR		0.000		0.000		0.000	0.011
Total score		0.177		0.293		0.236	
Renormalized score		0.251		0.414		0.335	

Table 1: A spreadsheet model was used to process each judge's ratings of each research paper according to the five accepted evaluation criteria. The weights for the ratings scale shown under each criterion were developed using the AHP (as shown in Figure 2). Hypothetical results of the evaluation of three research papers by one committee member are given in the table.

committee members aware of the local ratings-scale weights but not the criteria weights in order to minimize potential biases. We transferred the ratings from each committee member's scoring sheet to a corresponding Lotus 1-2-3 spreadsheet for processing (Table 1).

After collecting all of the rating sheets, we prepared a summary spreadsheet (Table 2). It contains the renormalized scores (from Table 1) for each judge. It also shows the individual judges' preference rankings for the papers based on these renormalized scores. We also computed the mean and median scores for each paper to help make the final selection.

Benefits of the New System

To achieve a consensus on the award-winning papers under the original system, the judges met as many as five times for three-hour sessions. It took them that much time to evaluate and discuss the papers. The first year they used the new system a miraculous event occurred. After minimal discussion at the second meeting, the committee reached complete consensus on the rankings of the top three choices within one hour. The committee chairperson looked up to heaven and proclaimed: "Eureka! The impossible has occurred!" The committee's recommendations were accepted by the dean.

We have used this system successfully for three years, without any serious glitches. During this time, both the committee and the dean have accepted the rankings generated by the new system. Many faculty members of the college regard the AHP-based system as a more structured method that is better able to capture the subjective judgments of the evaluators. For this reason they prefer it over any previous scoring system. Periodically we review and modify all of the components of the system to reflect changes in committee composition and in the research

	Papers					
	1		2		3	
Judges	Renormalized Score	Rank	Renormalized Score	Rank	Renormalized Score	Rank
Judge 1	0.251	3	0.414	1	0.335	2
Judge 2	0.221	3	0.378	2	0.401	1
Judge 3	0.168	3	0.451	1	0.381	2
Judge 4	0.261	2	0.492	1	0.247	3
Judge 5	0.255	3	0.382	1	0.363	2
Mean Score	0.231	3	0.423	1	0.345	2
Median Score	0.251	3	0.414	1	0.363	2

Based on the median scores, the following are the final rankings of all papers: Paper 2, Paper 3, Paper 1.

Table 2: The scores in this table were obtained after five hypothetical judges applied the spreadsheet method given as Table 1. The rankings and other summary results shown were used in obtaining a consensus on the final paper rankings. Hypothetical summary scores of the evaluation of three papers are provided in the table.

objectives of the university.

Conclusions and Directions for Future Work

Using the AHP-based spreadsheet scoring model, our committee reduced the number of its meetings from five to two, saving approximately 11 hours for each of the 12 committee members. Probably even more important than the time savings, which are not trivial given the members' nine-hour teaching loads, is their reduced consumption of Mylanta (or possibly Tagamet).

Applying this ratings system successfully requires only two ingredients; first, some initial commitment of time for operationalizing the process, and second, the support of the committee members. Its ongoing usage shows that management science can be applied, even in an academic setting.

Similar applications are possible in other areas within colleges and universities. For example, the same basic approach could be used to evaluate proposals for summer research grants or sabbaticals.

References

Aczel, J. and Saaty, T. L. 1983, "Procedures for synthesizing ratio judgments," *Journal of Mathematical Psychology*, Vol. 27, No. 1, pp. 93–102.

Dyer, J. S. 1990a, "Remarks on the analytic hierarchy process," *Management Science*, Vol. 36, No. 3, pp. 249–258.

Dyer, J. S. 1990b, "A clarification of 'remarks on the analytic hierarchy process,' " *Management Science*, Vol. 36, No. 3, pp. 274–275.

Forman, E.; Saaty, T. L.; Selley, M.; and Whittaker, R. 1990, *Expert Choice*, a propriety software package developed by Decision Support, Inc., Pittsburgh, Pennsylvania.

Harker, P. T. and Vargas, L. G. 1990, "Reply to 'Remarks on the analytic hierarchy process' by J. S. Dyer," *Management Science*, Vol. 36, No. 3, pp. 269–273.

Liberatore, M. J. 1987, "An extension of the analytic hierarchy process for industrial R&D projects selection and resource allocation," *IEEE Transactions on Engineering Management*, EM-34, No. 1, pp. 12–18.

Saaty, T. L. 1980, *The Analytic Hierarchy Process*, McGraw-Hill, New York, New York.

Saaty, T. L. 1990, "An exposition of the AHP in reply to the paper 'Remarks on the analytic hierarchy process,' " *Management Science*, Vol. 36, No. 3, pp. 259–268.

Winkler, R. L. 1990, "Decision modeling and rational choice, AHP and utility theory," *Management Science*, Vol. 36, No. 3, pp. 247–248.

Zahedi, F. 1986, "The analytic hierarchy process—A survey of the methods and its applications," *Interfaces*, Vol. 16, No. 4, pp. 96–108.

Alvin A. Clay, Dean, College of Commerce and Finance, Villanova University, Villanova, Pennsylvania 19085-1678, writes "I am delighted to write this letter to confirm the successful work that was conducted by the Commerce and Finance Research Committee. The College of Commerce and Finance Research Awards Program serves an important function in the college in both rewarding and encouraging faculty research efforts. The committee attacked the recurrent problem of evaluating research papers and recommending those that were worthy of recognition and award. Their newly developed rating and evaluation process, described in the paper submitted to you, has been successfully used during the last three years, and I am pleased with the results.

"The College of Commerce and Finance will definitely continue to use this evaluation system. We expect that only minor modifications will be required periodically."