IDA PAPER P-2387

SELECTED JUDGMENTAL METHODS IN DEFENSE ANALYSES

Volume I: Main Text

Jeffrey H. Grotte
Lowell Bruce Anderson
Mitchell S. Robinson

July 1990

Prepared for
Joint Chiefs of Staff

INSTITUTE FOR DEFENSE ANALYSES
1801 N. Beauregard Street, Alexandria, Virginia 22311-1772

Series B
IDA Log No. HQ 90-35423

DEFINITIONS

IDA publishes the following documents to report the results of its work.

Reports

Reports are the most authoritative and most carefully considered products IDA publishes. They normally embody results of major projects which (a) have a direct bearing on decisions affecting major programs, (b) address issues of significant concern to the Executive Branch, the Congress and/or the public, or (c) address issues that have significant economic implications. IDA Reports are reviewed by outside panels of experts to ensure their high quality and relevance to the problems studied, and they are released by the President of IDA.

Group Reports

Group Reports record the findings and results of IDA established working groups and panels composed of senior individuals addressing major issues which otherwise would be the subject of an IDA Report. *IDA Group Reports are reviewed by the senior individuals* responsible for the project and others as selected by IDA to ensure their high quality and relevance to the problems studied, and are released by the President of IDA.

Papers

Papers, also authoritative and carefully considered products of IDA, address studies that are narrower in scope than those covered in Reports. IDA Papers are reviewed to ensure that they meet the high standards expected of refereed papers in professional journals or formal Agency reports.

Documents

IDA Documents are used for the convenience of the sponsors or the analysts (a) to record substantive work done in quick reaction studies, (b) to record the proceedings of conferences and meetings, (c) to make available preliminary and tentative results of analyses, (d) to record data developed in the course of an investigation, or (e) to forward information that is essentially unanalyzed and unevaluated. The review of IDA Documents is suited to their content and intended use.

The work reported in this document was conducted under contract MDA 903 89 C 0003 for the Department of Defense. The publication of this IDA document does not indicate endorsement by the Department of Defense, nor should the contents be construed as reflecting the official position of that Agency.

This Paper has been reviewed by IDA to assure that it meets high standards of thoroughness, objectivity, and appropriate analytical methodology and that the results, conclusions and recommendations are properly supported by the material presented.

This Paper does not necessarily represent the views of the Joint Chiefs of Staff for whom it was prepared and to whom it is forwarded as independent advice and opinion.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 1990	Final

4. TITLE AND SUBTITLE
Selected Judgmental Methods in Defense Analyses, Volume I: Main Text

5. FUNDING NUMBERS
C-MDA 903 89 C 0003

TA-T-I6-593

6. AUTHOR(S)
Jeffrey H. Grotte, Lowell Bruce Anderson, Mitchell S. Robinson

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
1801 N. Beauregard St.
Alexandria, VA 22311-1772

8. PERFORMING ORGANIZATION REPORT NUMBER
IDA Paper P-2387

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
JCS, J-8
Room 1E965, The Pentagon
Washington, DC 20301

Director, FFRDC Programs
1801 N. Beauregard St.
Alexandria, VA 22311-1772

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This paper examines selected methodologies for collecting and using judgment data. Such methodologies may have applications in the consideration of "qualitative "aspects of military effectiveness, such as morale and leadership, as well as in the estimation of values for which good empirical bases do not exist. This paper examines several established approaches to using judgment to provide numerical values and ordinal rankings. Underlying principles, ease of implementation, and criticisms of the methods are discussed. Substantial bibliographic references are provided.

14. SUBJECT TERMS
Judgment, qualitative methods, Delphi method, utility theory, Analytical Heirarchy Process, AHP, subjective transfer function, STF, voting, paired comparison theory

15. NUMBER OF PAGES
228

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

IDA PAPER P-2387

SELECTED JUDGMENTAL METHODS IN DEFENSE ANALYSES

Volume I: Main Text

Jeffrey H. Grotte
Lowell Bruce Anderson
Mitchell S. Robinson

July 1990

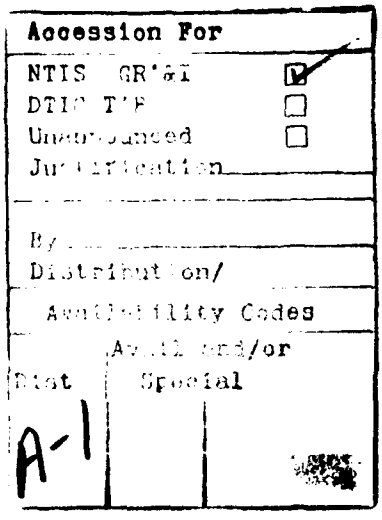
INSTITUTE FOR DEFENSE ANALYSES
Contract MDA 903 89 C 0003
Task T-I6-593

PREFACE

This paper was prepared by the Institute for Defense Analyses (IDA) for the Joint Chiefs of Staff under contract No. MDA903-89-C-0003, Task T-I6-593, Survey of Qualitative Methods in Military Operations Research.

The objective of this analysis is to summarize and evaluate methodologies for collecting and using judgmental data.

This paper was reviewed by Dr. Jesse Orlansky, Dr. Robert Kuenne, Mr. John Cook, and Ms. Laurna Hansen.



ACKNOWLEDGEMENTS

Dr. Clairice T. Veit, Dr. Thomas Saaty, and Dr. Luis Vargas helped us to clarify the chapters on the Subjective Transfer Function and the Analytic Hierarchy Process. Mrs. Renee Harper and Mrs. Eva Wiggins prepared the manuscript, and Ms. Eileen Doherty provided invaluable editorial assistance in bringing the document together. This paper was reviewed by Dr. Jesse Orlansky, Dr. Robert Kuenne, Mr. John Cook, and Ms. Laurna Hansen. However, all errors and omissions are entirely the responsibility of the authors.

ABSTRACT

The Institute for Defense Analyses was requested by the Organization of the Joint Chiefs of Staff to summarize and evaluate methodologies for collecting and using judgment data. Such methodologies have applications in the consideration of qualitative aspects of military effectiveness, such as morale and leadership, as well as in the estimation of values for which good empirical bases do not exist. Based on a survey of the literature, six methodologies were selected for evaluation. Two of these are well-defined methodologies that have had extensive application in defense analysis (the Delphi Method and the Analytic Hierarchy Process), two have theoretical importance that merit closer attention (Utility Theory and the Subjective Transfer Function Method), and two are not generally associated with defense analysis, but were explored to determine what, if any, defense applications might exist (Voting Theory and Paired Comparison Theory).

In general, this evaluation found that requirements for judgmental quantification are *not uncommon in military analysis.* However, ad hoc methods rather than methods that have received critical examination or that have a proven track record are frequently used. Well-scrutinized methods are sometimes not used because of the costs of implementing them relative to the amount and significance of the data required. However, these methods are also not used simply because of a lack of awareness of them. There is also not a great deal of ongoing communication among defense analysts regarding these methods. In addition, popularly used methods often suffer from inadequate understanding. Thus, better dissemination of the strengths and weakness of these methods, and of variants and extensions designed to address these weaknesses may be warranted. Finally, inordinate weight can be put on the judgment of expert sources. The concept of "expertise" and related questions about identifying experts and the accuracy of their judgments have received attention in the psychological literature, but may not be given due weight by users of expert judgment data. Undue weight given to expert judgments may lend undeserved authority to analyses based on these judgments.

With regard to specific methodologies, this evaluation found:

- The Delphi method is more a philosophy about designing group judgment collection processes than it is a tightly bounded methodology.

- The Analytic Hierarchy Process is a method for assigning numerical weights to the elements of a set. It does this by estimating the psychological scale values underlying pairwise comparative judgments. However, a number of serious criticisms suggest that the methodology should not be used in the form originally described in the literature, popularly used today, and implemented in some commercially and privately available software packages. The implications of these criticisms range from disregarding all but the method's ordinal results to discontinuing its use entirely.
- The Subjective Transfer Function Method attempts to bring important ideas about psychological measurement and psychological judgment processes to the analytical community. However this method has not yet received the attention necessary to determine its theoretical quality or practical value.
- The implementation of Utility Theory requires proper determination of utility functions, which has proven to require care in how information is elicited and requires respondents willing to participate in the type of exercises needed to determine such matters as risk averseness. If such conditions are not met, other methods may prove more informative.
- Where Voting Theory applies, it is obviously applicable, and forcing it to fit where it doesn't obviously apply does not appear to be useful.
- Paired Comparison Theory is not well known outside of a relatively small group of theorists. It has been applied, but not extensively, and apparently not in major defense analyses. Several paired comparison methods exist, and one (which has been regularly rediscovered) seems to have generally more desirable properties than the others. More widespread dissemination of this theory, together with sufficient ingenuity in adapting it to particular situations, could lead to many additional applications.

CONTENTS

VOLUME II: APPENDICES

A. Ratio and Subtractive Processes in Psychophysical Judgment

B. DTIC Bibliography on the Delphi Method

TABLES

FIGURES

EXECUTIVE SUMMARY

A. OBJECTIVE

The Institute for Defense Analyses was requested by the Organization of the Joint Chiefs of Staff to summarize and evaluate methodologies for collecting and using judgment data. Such methodologies have applications in the consideration of qualitative aspects of military effectiveness, such as morale and leadership, as well as in the estimation of values for which good empirical bases do not exist.

B. APPROACH

We surveyed the literature and held informal conversations with analysts. As a result of this investigation, we selected six methods for evaluation. Four of these methods have had extensive application in defense analysis (the Delphi Method and the Analytic Hierarchy Process), two have theoretical importance merited closer attention (Utility Theory and the Subjective Transfer Function Method), and two are not generally associated with defense analysis, but were explored to determine what, if any, applications might exist (Voting Theory and Paired Comparison Theory). We also identified a number of topics that apply to several judgment methods and that are important enough to analytical practice to merit separate attention.

C. SUMMARY OF FINDINGS

Both the literature review and the conversations with practitioners suggested that requirements for judgmental quantification are not uncommon in military analysis. However, many practitioners use "ad hoc" methods rather than methods that have received critical examination or that have a proven track record. Well-scrutinized methods are sometimes not used because of the costs of implementing them relative to the amount and significance of the data required. However, these methods are also not used simply because of lack of awareness of them. In addition, there is not a great deal of ongoing communication among defense analysts regarding these methods, although islands of expertise can be found. Even popularly used methods often suffer from inadequate understanding. Thus, better dissemination of the strengths and weakness of these

methods, and of variants and extensions designed to address their weaknesses, may be warranted.

Our literature review and conversations with practitioners also found some insensitivity to issues revolving around collecting judgment data and conducting social science research in general. For example, question design is often not thought of as an activity that requires careful attention to ensure interpretable and applicable results. Finally, and of particular importance, inordinate weight may be put on the judgment of expert sources. The concept of "expertise" and related questions about identifying experts and the accuracy of their judgments has received attention in the psychological literature on judgment and decisionmaking, but may not be given its due weight by users of expert judgment data. Undue weight given to expert judgments may lend undeserved authority to the analysis based on it.

With regard to specific methods, we found:

- The Delphi method is more a philosophy about designing group judgment collection processes than it is a tightly bounded methodology. Under the Delphi "philosophy," steps should be taken to mitigate the effects of undesired group social processes on the formation and reporting of group judgment. However, there is little good evidence regarding efficacy of particular steps under particular conditions.
- The Analytical Hierarchy Process is a method for assigning numerical weights to the elements of a set. It does this by estimating the psychological scale values underlying pairwise comparative judgments. However, a number of serious criticisms suggest that the methodology should not be used in the form originally described in the literature, popularly used today, and implemented in some commercially and privately available software packages. Some other criticisms suggest that some assumptions underlying the method are questionable and should be evaluated more thoroughly. The implications of these criticisms range from disregarding all but the method's ordinal results to discontinuing its use entirely.
- The Subjective Transfer Function Method is an approach to collecting and representing knowledge about complex systems. The method also brings important ideas about psychological measurement and about psychological judgment processes to the analytical community. The method itself, however, has not yet received the kind of thorough attention necessary to determine its theoretical or practical value to the analytical community.
- The implementation of Utility Theory requires proper determination of utility functions, which has proven to require care in how information is elicited and

requires respondents willing to participate in the type of exercises needed to determine such matters as risk averseness. If such conditions are not met, other methods may prove more informative.

- Where Voting Theory applies, it is obviously applicable, and forcing it to fit where it doesn't obviously apply does not appear to be useful. There are many different voting methods, and the choice of which to use can be so important that, given a fixed set of voters' preferences, choosing different methods can result in different winners. Various voting methods have various properties. In a sense, none is perfect. However, based on their properties and on the voting situations involved, some may be deemed better than others. A commonly used method, plurality voting, may be one of the worst for all situations. Plurality voting may be used so frequently because of its extreme simplicity, but this extreme simplicity can result in serious flaws.
- Paired Comparison Theory is not well known outside of a relatively small group of theorists. It has been applied, but not extensively, and apparently not in major defense analyses. Several paired comparison methods exist, and one (which has been regularly rediscovered) seems to have generally more desirable properties than the others. Specific applications, however, may have specific characteristics that are more suitably addressed by one of the other methods. More widespread dissemination of this theory, together with sufficient ingenuity in adapting it to particular situations, could lead to many additional applications.

I. SURVEYING JUDGMENTAL METHODS

A. INTRODUCTION

Information for military analysis often takes the form of concrete characteristics -- number, mass, length, distance, time, penetration capability and relationships among them. These attributes (as well as some societally defined attributes such as monetary value) are measured using models that have been applied for so long that we usually accept them intuitively and take them for granted. However, a number of important attributes, e.g., readiness, are not covered well by models of measurement for concrete quantities; other "resistant" attributes concern the products of psychological processes, e.g., morale. Moreover, conventional measurement methods may not be sufficient for yet other attributes that depend on events either rare in natural occurrence or yet to materialize (e.g., the advent of new technologies). The relative difficulty of measuring these resistant attributes using the techniques of the physical sciences has led us to think of them as qualitative. Yet we do have an intuitive sense about the magnitude of many of these attributes, and this allows us to conceptualize them in terms of categories (poor to excellent, one-to-ten scales, etc.) that in turn convey information about ordering and differences (e.g., adjacent categories on a scale may be more similar than distant categories).

In these and other instances where we lack hard data, analysts frequently take one of three paths: to estimate some values by drawing analogies from related data; to adopt a more tractable proxy for the ill-behaved quantities; or to subjectively estimate the desired quantities. The focus of this examination is on the third approach, the collection and use of judgment data. This chapter presents an overview of the six methods examined in this paper, briefly discusses a taxonomy for judgment methods, and discusses issues that should be considered in conducting judgment-based research. Chapters II through VI present the examinations of these methods. Observations and summation are offered as Chapter VII.

B. SELECTED JUDGMENT METHODS

Any enumeration of methods for collecting and using judgments will be large. In addition to the large variety of different methods, for any single one of these there

frequently are numerous extensions and modified versions developed to cover special cases and to correct problems. In addition, there are methods that use or depend exclusively on other judgment methods for data. For these reasons, we organize these methods into families related by purpose, theory and historical development.

The *Delphi Method* is a set of loosely defined rules for eliciting judgments in a group setting. Broadly construed, the Delphi method provides a framework for minimizing undesirable group social influences on the judgments of group members.

The *Subjective Transfer Function Method* is a method for modeling knowledge about how inputs relate to outputs in complex systems. The method provides guidelines on how to decompose a complex system into constituents, how to collect information on the relationships among the constituents, and how to build up a subjective model of the system from the analysis of the data. The subjective model can then be used to estimate system outputs for new inputs.

The *Analytic Hierarchy Process* is a theory and methodology about how to measure preferences among alternatives. The method provides guidelines on how to decompose decision problem and on how to measure strength of preference. It also specifies how to combine preference evaluations for multiple criteria into an aggregate preference index.

Utility Theory, and more generally, *Multi-Attribute Utility Theory* concerns modeling preferences when the outcomes are uncertain and making choices on the basis of these preference models. The theory provides guidelines on how to decompose the decision problem and on how to measure strength of preference. It also specifies how to combine preference evaluations for multiple criteria into an aggregate preference index. A variant of the theory applies to outcomes for which there is no uncertainty.

Voting Theory concerns selecting a winner or a set of winners from (or developing a full ranking of) a set of alternatives based on the preferences (judgments or beliefs) of several voters. If there are only two alternatives, the situation is trivial. Voting theory addresses situations involving multiple voters and three or more alternatives.

Paired Comparison Theory concerns developing a ranking (say, from best to worst) of a set of alternatives based on a set of paired comparisons between these alternatives. Each such comparison involves only two alternatives, with one winning and one losing that comparison, or (optionally) the comparison resulting in a tie. Magnitudes of victory in these comparisons are not usually considered. Some alternatives may be compared with each other more than once, while other alternatives may not be compared with each other at all. Further, when two alternatives are compared more than once, one of them does not necessarily win each such comparison -- it might win some, but lose others.

C. CHARACTERISTICS OF METHODS

In this section, we examine some of the characteristics of the various subjective assessment methods in order to assist the analyst in matching a method to a particular application. The most important concern, of course, is whether or not the theoretical and methodological underpinnings of a particular method make it appropriate for the issue at hand (these are discussed in the later sections of this paper). There is no quick and easy way of characterizing these differences; there are, however, a number of less fundamental differences among these methods that provide additional insight into applicability. These can be described briefly and, while not a formal taxonomy, can help to categorize the various methods. Table I-1 summarizes the discussion that follows.

1. Number of Judges

There are two cases with regard to the number of individuals to be involved in the process. Some methodologies are suited to extracting the judgments of a single individual (or a group of individuals who have reached consensus through some mechanism), while others address situations that involve a number of individuals whose differing views have to be consolidated into a single evaluation. In the former case, achieving *a priori* group consensus may not be a trivial requirement and may bias results, if forced.

2. Types of Judgments

Different methodologies require different types of judgments, some of which may be easier to produce than others. Among the methods we have investigated, many different types of evaluations are required -- preference/indifference, scale (e.g., from 1 to 10), and ratio (e.g., option A is twice as good as option B). For a given situation, however, some judgments may be harder to make than others. Voting methods involve numerous variations on these themes, including, for instance, the choice of a most preferred option out of several, a set of preferred but otherwise undistinguished options out of a larger set, or an ordinal ranking (say from most preferred through least preferred) of a set of options.

3. Ease of Implementation

Ease of implementation, itself a qualitative judgment, subsumes such considerations as how many steps might be required to implement a methodology, the amount of training needed for both the analysts and those providing judgments, and the manpower and time needed to apply the methodology.

Table I-1. Characteristics of Qualitative Measurement Methods

CRITERIA	DELPHI	ANALYTIC HIERARCHY PROCESS	SUBJECTIVE TRANSFER FUNCTION	UTILITY	VOTING	PAIRED COMPARISONS
NUMBER OF JUDGES	Many	Single	Single	Single	Many	Many
TYPES OF JUDGMENTS	Many types	Scale, Ratio	Scale	Preference/ Indifference	Many types	Preference/ Indifference
EASE OF IMPLEMENTATION	Moderate	Moderate	Difficult	Difficult	Easy	Easy
SOFTWARE AVAILABILITY	Not commercially available	Commercially available	Not available	Commercially available	Very limited availability	Not commercially available
MEASURES OF CONSISTENCY	Deviations from mean/median	Formal index of inconsistency available	Not available	Sets of preference relations can be examined for consistency	Not available	Not available
DEGREE OF ACCEPTANCE	Widely accepted	Widely accepted	Not widely known	Relatively accepted	Little known in defense context	Little known in defense context

4. Software Availability

The availability of software to facilitate the application of a methodology is an ease of implementation issue, but one that deserves special note. Good software can help structure an analysis, minimize errors, perform complicated or extensive calculations and provide documentation. These advantages can obtain, however, only if the software is well documented, validated and operated by knowledgeable individuals. If used without an adequate level of understanding, software simply makes it easier to misuse a methodology.

5. Measures of Consistency

Some methods make available measures of the consistency of the judge's responses. These can be used not only to evaluate results, but also to flag problems as they arise during an analysis.

6. Degree of Acceptance

Other things being equal, it is desirable to select methodologies that generally have been accepted in the defense community either because there is an established track record of analyses that have used those methodologies, or because practitioners are simply more comfortable with underlying concepts.

D. RELATED SURVEY

In an earlier study, Kneppreth et al. (1974) discussed a number of techniques for the assessment of "worth". The techniques ranged from procedures for eliciting rank order preferences to utility assessment methods. However, our understanding of "worth assessment" has changed considerably since that report. Some methods have been superseded by more sophisticated extensions (e.g., single-anchored estimation by the Analytic Hierarchy Process). Others, such as direct scaling methods, have been questioned by a number of researchers studying psychological measurement (Shepard, 1976, Birnbaum, 1981). Others still, such as Voting Theory, and Paired Comparison Theory, were not addressed at all. Finally, the Subjective Transfer Function Method was developed after the publication of that work.

E. ISSUES FOR JUDGMENT METHODS

In this section we review some ideas that are important to consider when addressing judgment studies. We refer the interested reader to the papers cited at the end of this chapter for additional discussion (e.g., Einhorn, 1974, Shanteau, 1988, Meyer and Booker, 1989).

1. The Concept of Expertise

We frequently are interested in the opinions and judgments of "experts," respondents who are assumed to be knowledgeable in an area of inquiry. We use expert respondents because we expect their special knowledge of facts, relationships, and reasoning strategies to render them a better source of information, judgment, and opinion than would be the nonexpert. However, this assumption may be unwarranted. Without reliable criteria for identifying experts, it is difficult to evaluate how much weight we should give to expert responses?

Among the first questions that emerge from a consideration of expertise are -- What is an expert? and How do we identify expert respondents? There exists a considerable

literature on differences between experts and novices in various domains (Larkin, McDermott, Simon, and Simon, 1980; Chase and Ericsson, 1981, 1982); this research, however, does not provide useful criteria for recruiting expertise to collect judgment data. We frequently "know experts when we see them" by their command of the subject area -- their "knowledgeability." Yet we do not have reliable criteria for assessing "expertise". Should expertise be defined by titles or credentials?, by years of education or relevant experience?, by number of relevant publications?, by identification by peers?, by reputation?, by standardized rating schemes?

While indexing knowledgeability directly or indirectly, none of these criteria guarantees that an "expert" respondent knows what is necessary to satisfactorily answer a given question.[1] Significant mismatches between what an expert knows and the information we seek may make a knowledgeable person appear decidedly nonexpert. This view is consistent with Armstrong's suggestion (1978, 1980) that beyond a minimal level, expertise is not a reliable index of accuracy in forecasting and prediction. In the studies he reviews, expert respondents are identified using knowledgeability criteria similar to those we suggest above, yet their expertise does not appear to correspond to an enhanced ability to accurately forecast and predict outcomes. Rather than belittling their value, Armstrong's observations suggest that there is information that even a knowledgeable individual cannot be expected to provide.[2]

A more sophisticated view of expert respondents recognizes that there is an unknown degree of mismatch between what the expert knows and what information the analyst seeks. The analyst's task, then, is to distinguish between those responses to place confidence in, and those to ignore. How can we make use of this operational view of the role of the expert as a source of judgment? First, we need to assess the relationship between what we ask the respondent and what the respondent knows. ..."Exactly what information are we requesting?"..."What does the respondent know with respect to answering the question?"..."What assumptions must the respondent make?"..."What can the respondent do with his knowledge to form a response?" For nontrivial questions, wherein we ask the respondent to estimate or predict some unknown quantity, the answers to these questions are crucial; if the respondent cannot be presumed to have sufficient

---

1 Shanteau, 1988, 1989, personal communication, has made very promising progress in identifying "expertise" in a less narrow sense than we have taken the term here. Prillaman [1989], has made progress in developing a procedure to flag potentially non-expert judgments.

2 Martino (1980) questions whether the "experts" of Armstrong's review were exceptionally knowledgeable in the subject under investigation.

knowledge upon which to base a response, then he should not be expected to form an "expert" response. Meehl (1951), for instance, in *Clinical Versus Statistical Prediction* has pointed out that "expert prediction" of long-term social outcomes (e.g., success in college) requires a model of mediating events of such detail that no one should be expected to make accurate predictions on this basis.

Respondents who do not have the knowledge required to make an "expert" response may nonetheless make an "informed" response. We then must question how the experts actually *formulated* the response in order to determine our degree of confidence in it. In some cases, the respondent may be able to report information on how he formed a judgment (see Ericsson and Simon, 1980, 1984, for discussions on when such reports may and may not be accurate, and on how to collect them). This information then might be used to evaluate the "goodness" of the judgment. In other cases, analysis of the respondent's judgments may reveal biases or inaccuracies. For instance, informed responses may well be based on "simplified judgment strategies" similar to those reported in the literature on human judgment and decisionmaking (Kahneman, Slovik, and Tverksy, 1982). These strategies are valid under some circumstances, but certainly not under many others. We discuss a number of them below in Section 2, Heuristic Judgment Strategies, but illustrate this idea with research by Neff and Solick (1983; also see Ryan-Jones, 1978). Neff and Solick asked "expert" respondents to predict human performance during continuous military operations. In both cases, the authors found that the "experts" were not able to predict the actual performance well. However, the authors also never made a convincing argument that their respondents *should have been able to make accurate predictions* by detailing the "steps" they would need to go through to form valid and acceptable responses.

What kind of knowledge would Neff and Solick's respondents have needed to accurately predict performance during continuous operations? This is not a trivial question and will not be addressed in detail here. However, it is sufficient to suggest that if performance under continuous operations is predictable at all, a reasonable model of performance on which to base predictions may well be nonlinear, involving interactions among several variables. With the benefit of hindsight, we are not surprised that "expert" respondents could not well predict performance on various tasks during continuous operations. On the other hand, Neff and Solick did observe that their respondents made "informed" responses consistent with "simplified" models of how performance changes over time. Eighty five of 99 predictions conformed to the rule that performance would remain the same or deteriorate with time. One person accounted for eight of the 14

deviations from this model by consistently predicting that performance would recover in the last time interval of the operations. Sixty seven of 99 predictions conformed to the rule that performance would strictly deteriorate over time. One additional person deviated from this concept by predicting that performance would not deteriorate over the first time interval.

Thus it appears that the "expert respondents" brought to bear performance models with some applicability, but not good enough for the purposes of the study. Ironically, as a result of their research Neff and Solick themselves may have become more accurate predictors of performance during continuous operations than were their expert respondents. The authors did not test this conjecture.

A more sophisticated view of expert respondents also recognizes that when the analyst does not have confidence in the "experts'" responses, the "experts" may nonetheless have useful knowledge. For instance, rather than asking a respondent to make a prediction for which useful predictive models already exist, the analyst might solicit information on parametric inputs to the models or on data necessary to estimate the parametric inputs. The expert respondent may be more able to provide factual data more accurately than deriving a judgment or a forecast.

Our operational view of expertise points to the possibility that useful experts do not exist for some areas of interest. In a controversial critique of the Delphi method, Sackman also questioned the availability of expert judges:

> "A tacit, largely unchallenged assumption of the Delphi is that authentic experts do in fact exist for predicting the extremely complex socio-economic-technological events so common in Delphi questionnaires. Closer scrutiny reveals this to be wishful thinking. Many of these events are initial forays into unknown areas requiring unknown skills, hence, unknown 'experts.' Even if such events are understood to some extent, they typically presuppose a fantastic array of real, not shallow, skills....When we match predictions of complex sets of social events against 'experts,' we get something like the fabled blind men examining the Indian elephant." (pp. 34-35)

2. Heuristic Judgment Strategies

The literature on human judgment and decisionmaking is replete with studies demonstrating systematically inaccurate or "biased" judgments by experts and nonexperts alike (Rubin, 1989). There are many sources for these inaccuracies. One such, described above in connection with Neff and Solick's (1983) research on expert military judges, is the use of simplified models as the basis for predictions. More generally, when asked to make difficult estimates or forecasts, nonexpert judges may well use so-called heuristic

strategies to simplify the problem, to make the best use of available information, or to simply respond to a difficult task request. Heuristic judgment strategies are not accurate under all conditions, and may result in predictably "biased" judgments when they are not.

Spetzler and Von Holstein (1975) characterize these judgment strategies with regard to subjective probability assessment:

> "People seem to assess uncertainty in a manner similar to the way they assess distance. They use intuitive assessment procedures that are often based on cues of limited reliability and validity.
>
> "To pursue the example with estimation of distance, it is known that people consistently overestimate the distance of a remote object when visibility is poor and underestimate the distance when the sky is clear. In other words, they exhibit a regular systematic bias. This is because they rely on the haziness of an object as a cue to its distance. This cue has some validity, because more distant objects are usually seen through more haze. At the same time, this mode of judgment may lead to predictable errors.
>
> "These same characteristics apply to the assessment of uncertain quantities. Here too, one relies on certain modes of judgment that may introduce systematic biases." (pg. 344)

We hypothesize that there is a similarity between the nonexperts observed in psychological studies and expert respondents who are "knowledge poor" with regard to particular probes for information. When an expert cannot make an "expert response" based on sufficient knowledge he may nonetheless make an informed response based on a simplified heuristic concept of the judgment task and the system in question.

The similarity between nonexpert respondents in the lab and expert respondents in more realistic contexts makes it worthwhile to briefly review the literature on heuristic judgment strategies.

Among the first of the well-known studies on predictable bias in human decision-making are those published by Kahneman and Tversky (Kahneman and Tversky 1971, 1973; Tversky and Kahneman, 1973; Tversky, 1974). Tversky and Kahneman (1973) demonstrated that in a variety of experimental tasks, people frequently evaluate likelihood using *availability*, the ease with which relevant instances of an event or concept come to mind as a cue. In one study, Tversky and Kahneman read respondents lists of public personalities. Respondents heard a list of 19 more famous people and 20 less famous people. After listening to the list, respondents recalled more names of the more famous people than names of less famous people. In addition, eighty of ninety-nine respondents judged the class of more famous names to be more numerous in the list. Tversky and

Kahneman suggest that respondents judged the relative frequency of the two types of names by quickly recalling names and extrapolating recall success to a frequency judgment.

In extending their work out of the laboratory, Kahneman and Tversky suggest how the availability heuristic may be applicable to more realistic problems:

> "We often construct *scenarios*, i.e., stories that lead from the present situation to the target event. The plausibility of the scenarios that come to mind, or the difficulty of producing them, then serve as a clue to the likelihood of the event. If no reasonable scenario comes to mind, the event is deemed impossible or highly unlikely. If many scenarios come to mind, or if the one scenario that is constructed is particularly compelling, the event in question appears probable." (pg. 229)

Spetzler and Holstein (1975) continue:

> "The credibility [likelihood] of a scenario to a subject seems to depend more on the coherence with which its author has spun the tale than on its intrinsically 'logical' probability of occurrence." (pg. 347)

In other papers, Kahneman and Tversky (1972) demonstrated respondents judging the probability of an event according to the degree to which it "represents" the essential characteristics of its parent population or generating process. For instance, respondents judge that a coin-tossing sequence of three heads followed by three tails (HHHTTT) should occur less frequently than a more random appearing sequence, such as (THHTHT). Both sequences are equally likely. Kahneman and Tversky term this judgment strategy the *representativeness heuristic*.

Feller (1968) recounts a more naturalistic example of reasoning by representativeness. He writes that the spatial distribution of flying-bomb hits on London during the Second World War was random and well-approximated by a Poisson distribution. However, he reports that most people believed that the hits were not random, but deliberately aimed, because many areas were not hit at all while several areas were hit several times. The clustering they observed appeared to be unrepresentative of a random process and the operation of an intuitive law of large numbers in which events "even out," given enough repetitions.

Armstrong (1978) and Spetzler and Von Holstein (1975) discuss another judgment heuristic, *adjustment and anchoring*. A particular value, the "anchor," is used as an initial basis for formulating responses. The subsequent responses are formed by making adjustments on this value.

Fallon (1976) and others (Tversky and Kahneman, 1974, and Slovik, Fischoff, and Lichtenstein, 1982, summarize some research results) have demonstrated the effect of anchoring on subjective estimates of magnitude. Fallon asked respondents questions like "What do you assess as the probability of the Pentagon having an area greater than (500,000 square feet/3 million square feet)?" and "What do you assess as the probability of Sophia Loren being older than (55/35)?" where different respondents received the alternative numerical referents. Then asked to estimate the actual quantities, respondents receiving a larger referent also made larger estimates on the average: 3 million versus 1.2 million square feet for the area of the Pentagon and 44.38 versus 42.96 for Sophia Loren's age at the time.

Spetzler and Von Holstein comment that the adjustments from the anchor are often insufficient because the anchor exerts such a "dominating influence" on the estimation process. Further, the choice of anchor frequently is not chosen to be reasonable with respect to the quantity being estimated (i.e., a high-valued anchor might be reasonable for estimating a high-valued quantity).

Of course it is not inevitable that such judgmental biases take place. However, documenting bias in simple laboratory demonstrations points to the need to guard against them in more important situations. In this vein, Spetzler and Von Holstein recommend specific measures to preclude and ameliorate the effects of judgmental biases. Other authors have discussed methods for training good forecasters and estimators (Kahneman and Tversky, 1977; Lichtenstein, Fischoff, and Phillips, 1977; Fong, Krantz, and Nisbett, 1986; Agnoli and Krantz, 1989, Meyer and Booker, 1990) and have discussed why some estimators are better than others (Rubin, 1989; Winkler and Murphy, 1968). However, the effectiveness of several of these procedures has yet to be assessed.

3. The Costs of Using Expert Respondents

We noted earlier that knowledgeable respondents may not be able to provide accurate judgments when the requirements of doing so exceed the respondents' resources. Recognizing this, a number of authors have argued that using respondents labelled "expert" may create undesirable perceptions regarding the accuracy of the studies of which they are a part. Specifically, an *"expert halo effect"* may lend unjustified authority to findings and interfere with the responsibility of the analyst for the findings of the study and its accuracy.

Sackman (1974) warns in his critique of the Delphi method --

> "Delphi is enmeshed in a pervasive halo effect. The director, the panelists, and the users of Delphi results tend to place excessive credence on the opinions of 'experts' .... [Expert] panelists bask under the warm glow of a kind of mutual admiration society. The director [of the Delphi study] has the prestige of pooled authority behind his study, and the uncritical user [of the study's results] is more likely to feel snug and secure under the protective wing of an impressive phalanx of experts.
>
> The result of the expert halo effect for Delphi is to make no one accountable. The director merely reports expert opinion objectively, according to prescribed procedure; he is not responsible or liable for outcomes .... The user can always claim that he was simply following the best advice available, and that he is not responsible for what the experts say. Everyone has an out, no one needs to take any serious risks, and no one is ultimately accountable. With so much to gain, so little to invest at such low risk, no wonder this method is so popular. The Delphi belief structure is psychologically held together by the cementing influence of the expert halo effect." (pg. 34).

Armstrong (1980) concurs, remarking that a client who calls in the best expert available avoids blame if the forecasts are inaccurate.

4. Validity, Generalizability, and Reliability: Three Dimensions of Judgment Research

There are a number of ways that we evaluate and make sense out of research results. However, what we shall refer to as *validity*, *generalizability*, and *reliability* are especially important.

a. Validity

One of the more important criteria in evaluating research results is whether the research "measures what it purports to measure" (Armstrong, 1978, on "construct validity"). We refer to this concept as *validity* .[3]

The validity of some studies may be questionable because they actually *do not* measure what they purport to measure (e.g., the validity of an opinion about the quality of a product may be questionable if the respondent has a financial interest in the product.). However, also important is that the validity of a research result is threatened because it does not allow for an unambiguous interpretation.

---

[3] Different authors (Campbell and Stanley, 1963, Armstrong, 1978, Brinberg and McGrath, 1985) define validity more broadly or as being of different types (e.g., *internal validity, external validity*). We restrict our definition of validity solely for the sake of clarity.

One important source of threats to unambiguous interpretability of judgment study results is the questions used to elicit information. Inadequately specified questions require the respondent to make assumptions to fill in gaps in order to answer the them. Inadequately specified questions also allow the respondent to elaborate upon the questions in ways that may not be consistent with the intent of the study. As a result, the questions that respondents are actually answering may not be the questions we intended to ask them.

Consider, for example, an analyst who wants to estimate the cost of software for a system. The cost-estimating tool requires an estimate of software size in terms of words of software, but the database measures software size in terms of lines of code. The analyst consults with several local experts, asking *How many words of software correspond to a software line of code?* The relationship between lines-of-code and words-of-software varies from project to project, so there is no single number that is correct. Not wanting to underestimate software size, the analyst picks a number somewhat larger than the mean response, a "conservative" choice, as the number he will use. However, the analyst has not specified what kind of number the respondents should provide. Thus, they may respond with a mean, a median, a conservative number, a liberal number, etc. Further, by not specifying the kind of software application, the respondents may be referring to applications for which the factor is large relative to other applications. By applying his own conservative screen to the respondent's estimates, the analyst may be using a number which, in fact, is more conservative than he actually desires.

A recent study by a Federal agency sought to evaluate the public's knowledge of risk factors for contracting a dangerous virus. One means of transmitting the virus is receiving blood from an infected person. The agency developed a question, *If you recently injected illegal drugs, should you give blood at a blood drive?* The question strictly requires an answer of "don't know," because there is no mention of the cleanliness of the hypodermic needle, which is the critical attribute with regard to transmitting the virus. The study leaders reported that "don't know" would be scored as a lack of knowledge about risk factors for transmitting the virus. Their "correct" answer is that blood should not be given. However, to knowledgeably give this "correct" answer requires an assumption that the hypodermic needles in question were infected. This assumption may be warranted if we assume that intelligent respondent will "psych out" the question, understanding its purpose and underlying assumptions. However, properly viewed, this answer indicates lack of knowledge of the risk factors for the virus. Indeed, poorly informed respondents may respond "no, don't give blood" because they believe that it is the illegal drug rather than the hypodermic needles that may be the transmission modality of the virus. This

question and several others like it in the study survey are threatened by invalidity because we don't know exactly what they mean.

In his critique of the Delphi method, Sackman (1974) (see also Watson and Freeling, 1982, 1983; Schoner and Wedley, 1989; and Dyer, 1990 for corresponding comments on "typical Analytic Hierarchy Process questions") argues that Delphi questionnaire items frequently are ambiguous.

> "We find vague, generalized descriptions of future events, permitting the respondent to project any one of a large number of possible scenarios as his particular interpretation of that event.
>
> For example, the Delphi inquiry might be concerned, as in Baran's study (1971, [Institute for the Future, R-26)], with the 'Potential Market Demand for Two-Way Information Service to the Home.' Baran had to leave vast areas unspecified in asking panelists when such services were likely to be available and how much they would cost the consumer. These unspecified areas included the configuration of hardware, software, and communications; the nature of federal, state, and local regulation of such mass computer services; the mix of public and private support of the information services considered; very brief general descriptions of the 30 information services (typically one paragraph); no indication of how the public will be taught to use such services; and many other socio-economic-technological areas impacting directly on these services....As presently practiced, Delphi is -- in many respects -- a psychological projective technique for future inkblots." (pp. 50-51)

One step to ensure validity in subjective judgment studies is to make requests for information as explicit as possible in terms of definitions, assumptions, and concepts referenced. Spetler and Von Holstein offer the following guidelines:

> "Clearly define the quantity. A good test is to ask whether the clairvoyant could reveal the value of the quantity by specifying a single number without requesting clarification. For example, it is not meaningful to ask for 'the price of wheat in 1975,' because the clairvoyant would need to know the quantity, kind of wheat, the date, the exchange, and whether you wanted to know the buying or selling price. However, 'the closing price of 10,000 bushels of durum wheat on June 30, 1975, at the Chicago Commodity Exchange' is a well-defined quantity." (pg. 344)

Payne (1951) in *The Art of Asking Questions* makes many useful suggestions for crafting requests for information. Salancik, Wenger and Helfer (1971) considered principles for constructing Delphi method questions that may have some generality.

b. Generalizability

Generalizability is a second important perspective from which to view research results. Generalizability refers to the ability to apply results beyond the particulars of the

study that generated them. In principle, there is no way of assessing with certainty whether the results of one study can be generalized. Campbell and Stanley (1963) comment about generalizability ("external validity"),

> "Generalization always turns out to involve extrapolation into a realm not represented in one's sample.
>
> Thus, if one has an internally valid [experimental] design, one has demonstrated the effect only for those specific conditions which the experimental and control group have in common." (pg. 17)

However, we usually *do* generalize the results of a study by guessing at those aspects of the study which can be disregarded and over which we can generalize. We naturally assume some attributes to be irrelevant to applying the results of the study. For instance, we typically do not worry that judgments collected on a Monday will only be applicable on that Monday or on Mondays in general. We also frequently assume that the closer two things are in their "significant" dimensions, the more confidently we can extend the results concerning one to the other. Of course, we rarely know "how close" the concepts must be for the approximation to be "good enough". Thus, assumptions underlying a generalization of results always should be carefully examined.

There are several kinds of threats to generalizability; a number of them have been discussed as threats to validity in the literature on research design (e.g., see Campbell and Stanley, 1963). However, whether they should be thought of as threats to validity or to generalizability is really immaterial; neither is desirable, and serious threats to generalizability may so severely limit the applicability of results that the survey is rendered as useless as if it were invalid.

The order in which questions are asked is one threat to generalizability. A respondent's estimate for quantity may depend on whether he estimated other quantities before or after it; this may happen for several reasons. The first question may create a particular perspective for the second question (e.g., an anchor); or the first question may elicit information pertinent to the second question that would not have been present otherwise. Answering the first question may simply introduce a learning or practice effect on answering questions. Alternatively, the respondent may lose interest in participating as a respondent after answering the first question. The results of a study may thus be specific to the particular ordering in which questions are asked.

Another threat to generalizability comes from respondent selection procedures. Respondents may vary in their demographic characteristics, their level of experience and

comfort in the role of "expert respondent," their self-interest in the outcome of the study, etc. That careful attention should be paid to these factors is self-evident.

In any study requiring recruitment of respondents, those who volunteer to participate may differ from those who decide not to participate. Participants may have a strong interest in the area of study or in the study's outcome. They may be those who disagree with what they perceive the study's purpose to be or the use of anticipated results. Respondents may be coerced into participation or they may be acquaintances of the study director. Any of these factors may influence the results of the study in a way that limits its applicability. In addition, when a study requires extended participation over time, respondents who persevere may be different from those who drop out in ways that compromise the generalizability of the results. Campbell and Stanley (1963), Armstrong (1978), Sackman (1974) and others discuss these as well as other threats to generalizability. They also discuss measures to take when these threats occur (also see Winkler and Murphy, 1974 for a discussion of the generalizability of judgment research to nonlaboratory settings.).

c. Reliability

Reliability refers to the repeatability of a measurement or result, the extent to which repetition of the measurement process will yield the same results. We are concerned here with variability in measurement which is not systematic, and which is attributable to the measuring instrument and its use rather than to change sin the quantity of interest.

In the presence of variability, we can only estimate the quantity of interest through the measurement sample we have collected. Therefore, subject to cost-benefit considerations, more reliability (less variability) should be preferred to less (more variability).

Basic techniques for assessing the reliability of a measurement instrument are well established (Carmines and Zeller, 1988) and are of three types. Retest methods involve making measurements on multiple occasions using the same measuring instrument on each occasion. Alternate form methods are like retest methods, except that the measurements are made using related but not identical measurement instruments. For example, different lotteries could be used to assess a utility function on multiple occasions (see Chapter V describing utility methods). Finally, the split halves method is like to the alternative-form methods except that the multiple measurements using related instruments are made on a single occasion rather than on multiple occasions. For example, two sets of lotteries for assessing a utility function could be interspersed within the same assessment session.

For all three methods, high reliability corresponds to obtaining similar results over occasions, over related measurement instruments, or over both. Each method for evaluating reliability has associated with it different methods for measuring reliability, as well as different advantages and disadvantages. For example, the attribute being measured may be more likely to change over dispersed measurements than over closely spaced measurements. Thus, observed variability may not be due to the measurement instrument. Similarly, the variability observed when related measurement instruments are used may be attributable to variation in measurement instruments rather than to reliability of the instruments per se.

We have not observed reliability studies of any of the procedures reviewed in this paper.

5. "Identical" Studies with Different Results: Limits to Validity and Generalizability

We have argued above that research studies involving judgment data should be carefully evaluated for validity and generalizability. Familiar reports on the sensitivity of survey research results to how information requests are phrased illustrates this view. A simple experiment run by Schoemaker (Withers, 1990) illustrates this concretely. Schoemaker asked a class to evaluate a business proposition in one of two forms. The idea was either stated as having an 80 percent chance of success or as having a 20 percent chance of failure. Students given the "success" version of the idea were overwhelmingly likely to favor the idea, whereas most of the students given the failure "version" of the idea rejected it. The prognosis that there is an 80 percent chance of surviving an operation is similarly viewed more optimistically than a 20 percent chance of not surviving. (Simon, 1990, pp. 9)

These and related results suggest that we should carefully examine our assumptions regarding the generalizability of research results. The results of this literature also suggest that we should carefully assess whether the questions we ask will elicit responses that actually measure the quantity of interest. (see Payne, 1951)

Underlying the sensitivity of answers to question-phrasing is the human psychology that occurs between the analyst's probe for information and the respondent's reply. Often overlooked but of great significance is that the respondent must gain an understanding of the question or task posed. We term this understanding *representation*.

Representation involves more than the "literal" understanding of the linguistic elements of a verbal request. It also involves augmenting the request with assumptions, related information, inferences, perspective-taking (Is a pint glass filled with a cup of water thought of as half full or half empty? Is a risk thought of as an opportunity to win or as a chance to lose?), and the "adoption" of alternative representational formats (Is the linguistic request best thought of as an algebraic relation, as a mental image, as a statement in propositional logic?).

A good illustration of the systematic effects of language, task and representation judgment has been reported by Hershey, Kunreuther, and Schoemaker (1982) for utility function assessment procedures (also see Johnson and Schkade, 1989, and Johnson, Payne and Bettman, 1988. Spetzler and Von Holstein (1975) have written a related paper on subjective probability assessments).

Hershey et al first observed that there are several types of utility assessment procedures. The basic utility assessment procedure involves presenting the respondent with a choice between obtaining a sure payoff (S) or a two-outcome gamble. The chance payoffs are either a "large" payoff, denoted G and received with probability p, or a "small" payoff, denoted L and received with probability $(1-p)$. Certainty equivalence (CE) methods require respondents to state a level of S for fixed values of G, L, and p which makes them indifferent between the sure payoff and the chance outcome of the gamble. For instance, what value of a sure payoff would make you indifferent between receiving the sure payoff or accepting a gamble in which you receive \$100 with probability 0.5 or receive nothing with probability 0.5. In the probability equivalence (PE) method we fix S,G, and L, and respondents state the value of p which makes them indifferent between the sure payoff and the gamble.

In making CE- and PE-type judgments, a response will be of one of three types: risk-seeking, risk-averse, or risk-neutral. A risk-seeking response sets the value of the sure payoff to be more than the expected value of the gamble (e.g., I am indifferent between a sure payoff of \$200 and a gamble which has equally likely payoffs of \$300 and \$0. The expected value of the payoff is (.5)(\$300) + (.5)(\$0) = \$150.) A risk-averse response sets the value of the sure payoff to be less than the expected value of the gamble. A risk-neutral response sets the value of the sure payoff equal to the expected value of the gamble.

Hershey et al ran two utility assessment experiments. In one, each respondent answered only CE- or PE-type questions for functionally identical lotteries. In a second

experiment, each respondent answered both types of questions. In both studies, risk-seeking responses were more likely to occur in PE tasks than in CE tasks. Risk-averse responses were correspondingly more likely to occur in CE tasks than in PE tasks. The authors also found this result when they modified the experiment to remove a confounding factor.

In a third experiment, Hershey et al presented respondents with pure-loss decisions and mixed-outcome decisions. In pure-loss decisions, both the sure payoff S and the payoffs G and L associated with the gamble are negative. In mixed-outcome decisions, $G>0$ and $L<0$, but the expected value of the gamble is zero. Hershey et al created the mixed-outcome decisions by adding a single constant to S,G, and L in the pure-loss decisions so that $G>0$, $L<0$,and the expected value of the lottery equalled zero. Consistent with observations of other studies, risk-seeking responses were more likely to occur for pure-loss decisions and risk-averse responses were more likely to occur for mixed-outcome decisions.

From these studies and other studies we observe that functionally identical tasks stated in different language are represented differently by respondents. The result is a large difference in the judgments made.

These effects of language and task on subjective judgments has been termed a "biasing effect". We argue, however, that this will usually be a misleading statement because it suggests that there is some form of the task that is neutral and is therefore "correct" or "best." Rather, language and task always establish a context for an information request, and there may not be a neutral standard against which all other contexts are biased. Thus, there may not be a universally best way of requesting information, although there may be a way that is "better" with regard to the requirements of a particular study. Choosing a good method for eliciting responses requires a correspondingly clear knowledge of the goals of the study and how they relate to ways of requesting information.

6. The Value of Methodological Rigor: You Get What You Pay For

In the absence of hard data to support an analysis, our discussions with analysts and our reading of the literature suggest that expert judgment is seen as an acceptable means for quantifying intangibles and other resistant quantities. One reason for this is that expert judgment sometimes appears to be the only means to the end. However, another reason appears to be that analysts frequently view judgment methods as easy to implement,

requiring no more than a "cookbook level" understanding of the method and how to implement it. Moreover, when judgment data are required quickly or in large quantity, the "cookbook" attitude leads analysts to use ad hoc methods that appear to be reasonable, but whose assumptions have not been examined. Ad hoc methods require few assumptions, little effort, and less analysis.

Our discussion above and our reviews to follow suggest otherwise. We find that careful attention needs to be paid to research design in order to preserve validity and generalizability as best as possible. We also argue that expert respondents may not always be able to provide expert responses. In fact, as a fallback, experts may adopt simplified judgment strategies that are systematically inaccurate under specified conditions. To combat this, expert respondents may be trained to minimize the use of heuristic judgment strategies. However, it is not unusual for experts to not be so trained. Unfortunately, since generally we cannot measure the value or interpretability of a study, it is frequently difficult to characterize the degree of damage done by inattention to these threats or by failing to test the assumptions of the research. We question the value of casually applying subjective judgment methods when hard data are lacking. If the results of the study are important enough, then time and care should be taken to do it well -- careful attention to research design, pretesting, ameliorating the effects of judgmental bias, screening expert respondents for their ability to address the questions posed to them, and evaluating the assumptions underlying the methodology adopted. Not doing so compromises the quality of the study results to a degree that usually cannot be estimated. Not being able to answer the question "Are such results still "good enough" to be taken seriously?" may well imply an answer of "no" more frequently than we realize.

BIBLIOGRAPHY

L. Adelman, J. Mumpower, "The Analysis of Expert Judgment," *Technological Forecasting and Social Change*, Vol. 15, 1979, pp.191-204.

F. Agnoli, D.H. Krantz. "Suppressing Natural Heuristics by Formal Instruction: The Case of the Conjunction Fallacy," *Cognitive Psychology*, Vol. 21, 1989, pp. 515-550.

J.S. Armstrong. "Long Range Forecasting, From Crystal Ball to Computer," New York: John Wiley and Sons, 1978.

J.S. Armstrong. "The Seer-Sucker Theory: The Value of Experts in Forecasting," *Technology Review*, June/July 1980, pp. 18-24.

M.H. Birnbaum. "Controversies in Psychological Measurement," in B. Wegener (Ed.) "Social Attitudes and Psychophysical Measurement," Hillsdale, NJ: Erlbaum Publishers, 1982.

D. Brinberg, J.E. McGrath. "Validity and the Research Process," Newbury Park, CA: Sage University Press, 1985.

D. Campbell, J.C. Stanley. "Experimental and Quasi-experimental Designs for Research," Chicago: Rand McNally College Publishing Company, 1963.

W.G. Chase, K.A. Ericsson. "Skilled Memory," in J.R. Anderson (Ed.), "Cognitive Skills and Their Acquisition," Hillsdale, N.J.: Lawrence Earlbaum Associates, 1981.

W.G. Chase, K.A. Ericsson. "Skill and Working Memory," in G.H. Bower (Ed.), "The Psychology of Learning and Motivation," (Vol. 16), New York: Academic Press, 1982.

E.G. Carmines, R.A. Zeller. "Reliability and Validity Assessment," Beverly Hills, CA: Sage University Press, 1979.

Dyer, J. "Remarks on the Analytic Hierarchy Process," *Management Science*, Vol. 36 No. 3, 1990, pp. 248-258.

H. J. Einhorn. "Expert Measurement and Mechanical Combination," *Organizational Behavior and Human Performance*, Vol. 7, 1972, pp. 86-106.

H. J. Einhorn, "Expert Judgment: Some Necessary Conditions and an Example," *Journal of Applied Psychology*, Vol. 59, No. 5, pp. 562-571.

K.A. Ericsson, H.A. Simon. "Verbal Reports as Data," *Psychological Review*, Vol. 87, 1980, pp. 215-251.

K.A. Ericsson, H.A. Simon. "Protocol Analysis," Cambridge, MA: The MIT Press, 1984.

R. Ettenson, J. Shanteau, J. Krogstad. "Expert Judgment: Is More Information Better," *Psychological Reports*, Vol. 60, 1987, pp. 227-238.

W. Feller. "An Introduction to Probability Theory, Volume I," New York: John Wiley and Sons, 1968.

G.T. Fong, D.H. Krantz, and R.E. Nisbett. "The Effects of Statistical Training on Thinking About Everyday Problems," *Cognitive Psychology*, Vol. 18, 1986, pp. 253-292.

J.C. Hershey, H.C. Kunreuther, and P.J.H. Schoemaker. "Sources of Bias in Assessment Procedures for Utility Functions," *Management Science*, Vol. 28. No. 8, 1982, pp. 936-953.

E.J. Johnson, D.A. Schkade. "Bias in Utility Assessments: Further Evidence and Explanations," *Management Science*, Vol. 35 No. 4, 1989, pp 406-424.

E.J. Johnson, J.W. Payne and J.R. Bettman. "Information Displays and Preference Reversals," *Organizational Behavior and Human Decision Processes*, Vol. 42, 1988, pp. 1-21.

D. Kahneman, P. Slovik, and A. Tversky. "Judgment Under Uncertainty: Heuristics and Biases," Cambridge: Cambridge University Press, 1982.

D. Kahneman, A. Tversky. "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychology*, Vol. 3, 1972, pp. 430-454.

D. Kahneman, A. Tversky. "On the Psychology of Prediction," *Psychological Review*, Vol. 80 (4), July 1973, pp. 237-251.

D. Kahneman, A. Tversky. *Intuitive Prediction: Biases and Corrective Procedures*, Decision Research Technical Report PTR-1042-77-6, June 1977.

N.P. Kneppretn, D.H. Gustafson, R.P. Leifer, E.M. Johnson, "Techniques for the Assessment of Worth," U.S. Army Institute for the Behavioral and Social Sciences Technical Paper 254, DTIC AD-784 629, August, 1974.

J.H. Larkin, J.McDermott, D.P. Simon, H.A. Simon. "Models of Competence in Solving Physics Problems," *Cognitive Science*, Vol. 4, 1980, pp. 317-345.

S. Lichtenstein, B. Fischoff, L.D. Phillips. "Calibration of Probabilities: The State of the Art," in H. Jungermann and G. De Zeeuw (Eds.) "Decision Making and Change in Human Affairs," Dordrecht, Holland: D. Reidel Publishing Company, 1977.

J.P. Martino. "Review of *Long Range Forecasting: From Crystal Bell to Computer*," *Technological Forecasting and Social Change*, Vol. 16, 1980, pp. 269-273.

P.F. Meehl. "Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence," Minneapolis: University of Minnesota Press, 1954.

M.A. Meyer, J.M. Booker. Eliciting and Analyzing Expert Judgment, A Practical Guide. Los Alamos National Laboratory LA-11667-MS, NUREG/CR-5424, 1989. (Also London: Academic Press, in press.)

J.L. Mumpower, L.D. Phillips, O. Renn, V.R.R. Uppuluri (Eds.) "Expert Judgment and Expert Opinion," Proceedings of the 1986 NATO Advanced Research Workshop on Expert Judgment and Expert Systems, New York: Springer Verlag, 1987.

K.L. Neff, R.E. Solick. *Military Experts' Estimates of Continuous Operations Performance (or Close but No Cigar)*, U.S. Army Research Institute for the Behavioral and Social Sciences Technical Report 600, November 1983.

V. Neufeldt, D.B. Guralnik. "Webster's New World Dictionary of American English," Third College Edition, New York: Simon and Schuster, Inc., 1988.

S. Payne. "The Art of Asking Questions," Princeton: Princeton University Press, 1951.

J. Prillaman. "Identification of Expert Inadequacies," presented at The 26th Annual Symposium, Washington Operations Research/Management Science Council, November 1989.

D.L. Ryan-Jones. *A Comparison of Expert Ratings of Task Difficulty with an Independent Criterion*, U.S. Army Research Institute for the Behavioral and Social Sciences Technical Report 418, November 1979.

H. Sackman. *Delphi Assessment: Expert Opinion, Forecasting, and Group Process*, The Rand Corporation, R-1283-PR, April 1974.

J.R. Salancik, W. Wenger and E. Helfer. "The Construction of Delphi Event Statements," *Technological Forecasting and Social Change*, Vol. 3, 1971, pp. 65-73.

B. Schoner, W.C. Wedley. "Ambiguous Criteria Weights in AHP: Consequences and Solutions," *Decision Sciences*, Vol. 20, 1989, pp. 462-475.

J. Shanteau. "Psychological Characteristics and Strategies of Expert Decision Makers." *Acta Psychologica,* Vol. 68, 1988, pp.203-215.

J. Shanteau. "The 3 C's of Expert Audit Judgment: Creativity, Confidence, and Communication," presented at the USC Audit Judgment Symposium, February 1989.

R.N. Shepard. "On the Status of 'Direct' Psychological Measurement," in C.W. Savage (Ed.), "Minnesota Studies in the Philosophy of Science," Vol. IX, Minneapolis, MN: University of Minnesota Press, 1976.

H.A. Simon, J.R. Hayes. "The Understanding Process: Problem Isomorphs," *Cognitive Psychology,* Vol. 8, 1976, pp. 165-190.

H.A. Simon. "Prediction and Prescription in Systems Modelling," *Operations Research*, Vol. 38, No. 1, 1990, pp. 7-14.

P. Solvik, B. Fischoff and S. Lichtenstein. "Facts Versus Fears: Understanding Perceived Risk," in D. Kahneman, P. Slovik, A. Tversky (Eds.), "Judgment Under Uncertainty: Hueristics and Biases," Cambridge: Cambridge University Press, 1982.

C.S. Spetzler, C.A. VonHolstein. "Probability Encoding in Decision Analysis," *Management Science.* Vol. 22, 1975, pp. 340-358.

A. Tversky, and D. Kahneman. "Belief in the Law of Small Numbers," *Psychological Bulletin*, Vol. 76, 1971, pp. 105-110.

A. Tversky, and D. Kahneman. "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology*, Vol. 5, 1973, pp. 207-232.

A. Tversky, and D. Kahneman. "Judgment Under Uncertainty: Hueristics and Biases," Science, Vol. 185, 1974, pp. 1124-1131; also in D. Kahneman, P. Slovik, A. Tversky (Eds.), "Judgment Under Uncertainty: Hueristics and Biases," Cambridge: Cambridge University Press, 1982.

S.R. Watson, A.N.S. Freeling. "Assessing Attribute Weights," *Omega*, Vol. 10, No. 6, 1982, pp. 582-583.

S.R. Watson, A.N.S. Freeling. "Comment on: Assessing Attribute Weights by Ratios," *Omega*, Vol. 11, No. 1, pp. 13-14.

R.L. Winkler, A.H. Murphy. "Good Probability Assessors," *Journal of Applied Meteorology*, Vol. 7, 1968, pp. 751-758.

R.L. Winkler, A.H. Murphy. "On the Generalizability of Experimental Results," in Stacl Von Holstein (Ed.), "The Concept of Probability in Psychological Experiments," Dordrecht, Holland: D. Reidel Publishing, 1974.

D.M. Withers. "The Smarter Way to Make Decisions," *Working Woman*, March 1990, pp. 31-32.

II. THE DELPHI METHOD

This chapter presents an analysis of the Delphi method of forecasting. It is organized into five major subheadings, each followed by a (chronological) listing of reference literature relevant to that particular major subheading. Each of the five papers presented in this study follows the same general outline for ease of reader reference and comparison. Before beginning our dissertation on the Delphi method, we offer a brief historical background of its development.

Analysts at The Rand Corporation (Helmer, 1963, 1967; Dalkey, 1967; Dalkey and Helmer, 1968) originally developed the Delphi method as part of their work on the problem of "using group information more effectively" (Kaplan, Skogstad, and Girshick, 1950; Dalkey, 1969). Since that time, the Delphi method has become closely associated with forecasting and planning studies. Dalkey and Helmer (1968) initially applied the method to develop estimates of nuclear weapons requirements. Subsequently, practitioners have adapted the method to a variety of purposes, including measuring preferences and subjective probability estimates, generating and exploring policy options, and facilitating communication among competing interests. "The Delphi Method, Techniques and Applications" (Linstone and Turoff (eds.), 1975), describes several different types of Delphi applications. Two journals, *Technological Forecasting and Social Change* and *Futures* have become important forums for reporting Delphi methodology research and applications.

A. DESCRIPTION

The Delphi method is a technique for eliciting and refining group opinions. With respect to this purpose, the developers intended the Delphi method to be a rapid and relatively efficient way to "cream the tops of the heads" of a group of knowledgeable people. (Dalkey, 1969, pg. 16) Additional statements (Dalkey, 1969) imply that in designing the method, the developers sought to trade "depth" of group communication for

improved implementation characteristics (e.g., speed of reaching stable group responses, level of effort required from respondents). The argument made in favor of this tradeoff was that

> "...[while] round-table discussions and other psychological interactions tended to produce significantly better predictions than the individuals, equally good results could be obtained by statistically combining the individual responses. The group's psychological interaction thus did not, in itself, lead to improvement of the group's total ability at prediction as defined by the statistically determined response." (Pill, 1971, pp. 58-59).

The loss of communication depth relative to other "methods" however, has, been a focus of critical discussion of the Delphi method.

The theoretical focus of Delphi is on lessening the influence of purely social processes on opinion formation and reporting. It introduced three features for this purpose, which we take to be the method's defining characteristics. They are

Panelist anonymity. Individual panelists and individual responses are kept anonymous in all interactions.

Iterative polling with "statistical feedback." Delphi panelists respond to prepared questions from the process administrators, who provide feedback to the panelists on the group's responses. The feedback consists of, but is not limited to, the central tendency of the group response distribution (e.g., median), or related descriptive statistics (e.g., high-low range, interquartile range, frequency distribution). Process administrators also may provide for more substantive group interaction, which may include responding to needs for information or clarification, reporting comments, requesting justifications for responses, and requesting responses to these justifications. In this way, the administrators moderate an "anonymous debate" among the panelists.[1] Question-feedback cycles are repeated until the group satisfies some criterion; this criterion may be consensus-related, but it also may concern the availability of time, of funds, or of the panelists themselves.

Group response as a statistical aggregate. A measure of central tendency is usually used to represent the group judgment. The group median response has been the most frequently used measure of central tendency.

The rationale given for these three defining characteristics are relatively straightforward, although subject to criticism.

---

[1] In the earliest reported studies, process administrators requested panelists holding statistically extreme opinions to justify their responses. Panelists holding opinions more statistically representative of the group might then be asked to respond to these justifications.

Anonymity is thought to reduce the influence of asymmetries in reputation, authority, prestige, etc., on judgment formation and reporting. It also is thought to facilitate changes of opinion which, when stated publicly, individuals might be more resistant to change.

Selective feedback is thought to maintain the group's focus on the task, diminishing "social noise," such as discussion directed toward "group maintenance" goals. (Dalkey, 1968) It also is thought to prevent individuals from dominating a group discussion because of persistence, charisma, articulateness, etc. Iterative polling allows panelists to reconsider their opinion relative to the feedback received and to refine it with regard to information not considered in the preceding judgments.

Substituting a statistical summary of the panelists' opinions reduces explicit or implicit pressures to reach a consensus. It also forces each panelist's judgment to be explicitly represented in the group response. Thus, panelists who might otherwise reserve their opinion until the direction of the group response becomes obvious are forced to express an opinion on each round.

Researchers have studied the Delphi method as a method and as a context for individuals interacting within a group. This research both describes the dynamics underlying the "Delphi process" and tests ideas for improving the core Delphi methodology.

Some important work describing the "Delphi dynamics" has been reported by Scheib, Skutsch, and Schoffer (1975) and by Dalkey (1969) on changes in the response of Delphi panelists over iterations. Delphi panelists who answered almanac questions were fed back distributional information on the responses of the group. In one case this was a histogram of the responses; in the other, a listing of the quartile boundaries of the distribution. The authors found that the further a response was from the median or mode of the group's responses, the more likely the panelist would change his response on the next Delphi round in the direction of the statistical consensus. Further, the response changes were not completely accounted for by an accompanying movement toward the correct response. In a more illustrative demonstration, Scheib, et al, observed that false feedback could even move response away from the correct answer toward a false consensus value. These results are very suggestive of the statistical bandwagon effect that forms the basis for some criticisms of the Delphi method, which we discuss below.

Illustrative of the "Delphi improvement research" is the work by Brown and Helmer (1964) and by Dalkey, Brown, and Cochran (1969). These authors found that panelist subgroups with high self-ratings about almanac-type questions (e.g., What was the number of telephones in Africa in 1966?) were more accurate than the less confident panelists.

Other research aimed at improving the Delphi method has included looking at alternative criteria for terminating iterative polling (Scheib, Skutsch, and Schoffer, 1975; Dajani, Sincoff, and Talley, 1979; Chaffin and Talley, 1980); eliciting and feeding back distributional responses (Dalkey, Brown, and Cochran, 1969), including relevant facts in feedback (Dalkey, Brown, and Cochran, 1970); and the effects of eliciting and feeding back reasons for opinions (Dalkey, 1969).

The three defining characteristics of the Delphi method allow for considerable variation among applications varying in goals and panelist characteristics. Some applications have been collectively different enough from the original application to forecasting and estimation that they have acquired a distinguishing name (i.e., the Policy Delphi, Turoff, 1970, 1975; The Decision Delphi, Rauch, 1979). In addition, many studies depart from strict adherence to the three defining features while still referring to their methodology as "Delphi" or "modified Delphi."

BIBLIOGRAPHY

A. Kaplan, A.L. Skogstad, M.A. Girshick. "The Prediction of Social and Technological Events," *Public Opinion Quarterly*, Spring 1950, pp. 93-110.

N. Dalkey, O. Helmer. *The Use of Experts for the Estimation of Bombing Requirements, A Project Delphi Experiment*, The Rand Corporation, RM-727-PR, November 1951.

O. Helmer. *The Systematic Use of Expert Judgment in Operations Research*, The Rand Corporation, P-2795, September 1963.

B. Brown, O. Helmer. *Improving the Reliability of Estimates Obtained from a Consensus of Experts*, The Rand Corporation, P-2986, September 1964.

N.C. Dalkey. *Delphi*, The Rand Corporation, P-3704, October 1967.

O. Helmer. *Systematic Use of Expert Opinions*, The Rand Corporation, P-3721, November 1967.

N.C. Dalkey, O. Helmer. "An Experimental Application of the Delphi Method to the Use of Experts," *Management Science*, Vol. 9, 1968, pp. 458-467.

N.C. Dalkey. *The Delphi Method: An Experimental Study of Group Opinion*, The Rand Corporation, M-5888-PR, June 1969.

B. Brown, S. Cochran, N. Dalkey. *The Delphi Method, II: Structure of Experiments*, The Rand Corporation, RM-5957-PR, June 1969.

N. Dalkey, B. Brown, S. Cochran. *The Delphi Method, III: Use of Self-Ratings to Improve Group Estimates*, The Rand Corporation, RM-6115-PR, November 1969.

N. Dalkey, B. Brown, S. Cochran. *The Delphi Method, IV: Effect of Percentile Feedback and Feed-in of Relevant Facts*, The Rand Corporation, RM-6118-PR, March 1970.

M. Turoff. "The Design of a Policy Delphi," *Technological Forecasting and Social Change*, Vol. 2, No. 2, 1970.

H. Sackman. *Delphi Assessment: Expert Opinion, Forecasting, and Group Process*, The Rand Corporation, R-1283-PR, April 1974.

H.A. Linstone, M. Turoff. "The DELPHI Method, Techniques and Applications," Addison-Wesley Publishing Company, Inc.: Reading, MA, 1975.

M. Scheib, M. Skutsch, and J. Schofer. "Experiments in Delphi Methodology." in H.A. Linstone and M. Turoff (eds.), "The DELPHI Method, Techniques and Applications," Addison-Wesley Publishing Company, Inc.: Reading, MA, 1975.

M. Turoff. "The Policy Delphi," in "The DELPHI Method, Techniques and Applications," H.A. Linstone and M. Turoff (eds), Addison-Wesley Publishing Company, Inc.: Reading, MA, 1975.

J.S. Dajani, M.Z. Sincoff, W.K. Talley. "Stability and Agreement Criteria for the Termination of Delphi Studies," *Technological Forecasting and Social Change*, Vol. 13, 1979, pp. 83-90.

W.W. Chaffin, W.K. Talley. "Individual Stability in Delphi Studies," *Technological Forecasting and Social Change*, Vol. 16, 1980, pp. 67-73.

W. Rauch. "The Decision Delphi," *Technological Forecasting and Social Change*, Vol. 15, 1979, pp. 159-169.

W.E. Riggs. "The Delphi Technique, An Experimental Variation," *Technological Forecasting and Social Change*, Vol. 23, 1983, pp. 89-94.

W.G. Rieger. "Directions in Delphi Developments: Dissertations and Their Quality," *Technological Forecasting and Social Change*, Vol. 29, 1986, pp. 195-204.

B. CRITICISMS, CAVEATS, AND REPLIES

A number of studies have compared the Delphi method with other methods for assessing a group opinion. This is a difficult kind of analysis to do well, and many studies of the Delphi method in fact have not been done well; thus we have not extensively researched studies comparing Delphi with other methods. (Many were, in fact, dissertations and thus do not appear in journal publications.) However, these and other studies are referenced in the reviews made by Sackman (1974), Riggs (1983), and Rieger (1986).

The reason such studies are difficult to do is because of the multidimensional nature of the question and the "apples and oranges" nature of the comparisons being made. A study reported by Riggs (1983) is a good example of the difficulties of making a "fair" comparison. Riggs formed groups to predict the point spread of two home-team college football games. One group used the Delphi method and the other used "group discussion." The Delphi method produced forecasts that were more accurate. Riggs therefore concluded that his study lent support to the view that Delphi produces more accurate long-term forecasts than did conference methods. However, we can argue that, for the purpose of

comparing Delphi with group discussion in general, the methodology was woefully inadequate.

Forecasting by group discussion may depend highly on the directive and integrative functions provided by an effective group leader. Running groups well is a skill that some take to be a profession in itself. Rigg's groups were formed *ad hoc* by randomized selection from classroom attendees; thus, there is no guarantee that the discussion groups were run well.

Riggs' conclusion that his Delphi groups outperformed his discussion groups is certainly correct. However, the problems with his methodology preclude the generalization that Delphi groups outperform conference methods.

There also are questions regarding what criteria should be used to compare competing methodologies. Methods may differ in accuracy (i.e., defined by a point estimate, or by a prediction interval), implementation cost, time-to-consensus, marginal improvement in accuracy over time, marginal cost of accuracy differences, etc. Also, the differences between methods in performance may well depend on the composition of the group (e.g., demographic characteristics, interest of members in the question of interest, "status" levels represented, etc.), and the importance of effective leadership if the group cannot direct itself well. Rather than leading to an all-out winner, we hypothesize that different methods may be preferred under different circumstances. Intelligent use of the Delphi method, or any other group problem-solving method, may depend on recognizing what method is best for the current application.

Research reported by Brockhoff (1975) illustrates this point well. In comparing Delphi groups with "natural discussion groups," Brockhoff reported that Delphi groups outperformed discussion groups in answering almanac questions, but were inferior in making short-range forecasts. Brockhoff was not able to explain this result, and we would not want to generalize it to a working principle.

In this review of Delphi we focus on the soundness of the method's defining characteristics and underlying assumptions, and point out methodological issues that may need special attention by the practitioner. We have taken this approach for several reasons. First, for the reasons discussed above, it is difficult to make a good comparison of alternative methods; rather, it is more to the point to know what method is most suited under given circumstances. Second, just as there is considerable freedom in implementing the specifics of a Delphi method, so should there be restraint in adopting only those of the Delphi's defining characteristics that suit the requirements of the study. Finally, our

literature review revealed that sloppy implementation may represent a greater threat to the value of Delphi method results than would the validity of the method's underlying assumptions.

There have been a number of methodological analyses of the Delphi method. One in particular, Sackman's (1974) critique, has attracted the most attention. In it, Sackman broadly condemned the method on the basis of general methodological standards and on the basis of the method's underlying assumptions. Sackman felt that Delphi does not meet accepted scientific standards for questionnaire-type and other social science methods. As a result,

> "Neglect of standard experimental guidelines may lead to uncontrolled variations in results and inability to define, replicate, and validate methods and findings. This neglect may be acceptable for an informal exploratory technique, but it is unacceptable for a rigorous social science experiment." (pg. 12)

However, most of Sackman's comments concern deficiencies he observed in the method's practice (Linstone, 1975, Hill and Fowles, 1975, Goldschmidt, 1975, Rieger, 1986) rather than in the method's inherent characteristics. Nonetheless, his comments are significant. Sackman observed that Delphi investigators *do not*...

> "...Subject numerical results to rigorous statistical controls and analysis. For instance, Delphi estimates should be accompanied by confidence bound-type statements of precision and comparisons between estimates should be accompanied by statements of statistical significance.
>
> "Evaluate the reliability of questions prepared for Delphi panelists.
>
> "Adequately evaluate the validity of the questions prepared for panelists. Furthermore, the questions asked are frequently ambiguous, requiring each panelist to make assumptions in order to answer them. The panelist may then not be answering the same questions in their responses.
>
> "Document the qualifications and experience of their panelists so as to operationally define their expertise. Further, the skills required by the Delphi questions are not explicitly matched to the objectively measurable skills of the panelists.
>
> "Take care to preclude or control for the influence of panelist demographic characteristics or panelist dropout on the study outcomes. Nor do they take care to ensure that Delphi panelists do not have an interest in the outcome of the study."

On the basis of these criticisms, Sackman concluded that "...conventional Delphi neglects virtually every major area of professional standards for questionnaire design, administration, application, and validation." (pg. 27)

As we pointed out above, his criticisms do not regard the inherent characteristics of the method itself, but rather the practice of the method as he observed it. Sackman did, however, raise several criticisms which appear to be inherent to the method:

> "We do not understand well the circumstances under which anonymous polling is superior to face-to-face discussion in formulating a group opinion....
>
> "The alleged superiority of anonymous Delphi opinion over face-to-face opinion, and its converse, are unprovable general propositions. They cannot be proved or disproved, in general, because the propositions are amorphous stereotypes and are not amenable to scientific testing unless they are operationally defined." (pg. 45)[2]

We similarly do not understand well the circumstances under which group opinion is superior to individual opinion.

Delphi deliberately eliminates the adversary process inherent in face-to-face confrontation. In the eyes of Delphi practitioners this is a positive attribute because it mitigates undesired social processes. However, Sackman not only thought otherwise but also thought the "anonymous debate" to be a weak substitute for face-to-face confrontation. He argued that

> "...*authentic consensus* refers to group agreement reached as a result of mutual education through increased information and the adversary process, which leads to improved understanding and insight into the issues." (pg. 45)

Although writing in response to Sackman, Coates (1975) concurred.

> "He [Sackman] ignores the numerous variations, advances [in the Delphi technique] and so forth...that more effectively address the issues of Delphi as a tool for drawing forth ideas, options, alternatives, diagnoses, etc." (pg. 194)

Overall, none of these criticisms is sufficient to determine whether or not to use the Delphi method in a given study. For instance, without information on the circumstances under which anonymous polling is superior to face-to-face discussion, it cannot be used as a general basis for preferring one method over another.

Another of Sackman's comments that the quality of the Delphi consensus is questionable concerns an inherent characteristic of the Delphi method, but deserves a separate treatment.

---

[2]Sackman doesn't define the term "superiority," but his usage suggests it refers to both the quantity of relevant information upon which estimates are made, and the absence of deleterious group social processes.

> "The Delphi procedure arrives at a consensus by feeding back the 'correct' answer, by rewarding conformity and effectively penalizing individuality, ... It [authentic consensus] does not refer to changes of opinion associated primarily or exclusively with bandwagon statistical feedback." (pg. 45)

Ford (1975) has come to a similar conclusion, although his argument refers more to observations of Delphi practice than to inherent characteristics of the method.

> "Delphi is meant to reduce pressure toward conformity and it is claimed that 'there is no pressure to arrive at a consensus.'(Dalkey (1968) p. 4)..Yet the controlled feedback of a typical exercise is designed to influence subsequent estimates in the direction of the whole group while ignoring possible emergent subgroups or cliques. There may not be any overt pressure to reach a consensus, but feedback constitutes an obvious pressure to influence conforming response changes." (p. 139.)

The Delphi method is premised on sampling the opinions formed solely on the basis of "facts" perceived to be directly relevant to the study.[3] However, Sackman's concern about the quality of the Delphi consensus is precisely that statistical feedback leads panelists to conform to the perceived group opinion under the pressure of group social processes. Furthermore, Hill and Fowles (1975) have pointed out that unmotivated Delphi participants may simply conform to the average response in order to bring the process to an end.

Do the hypothesized conformity-inducing aspects of the Delphi method make its accuracy worse than other methods? We have not found unambiguous answers to this question, and we suspect that the answer varies with the specifics of the Delphi study and the individual characteristics of the participants. A particular group of self-confident panelists may be unmoved by the group mean opinion if they believe their opinion to be correct. Others may adjust their opinion closer to the group mean if they are unsure about the quality of their opinion. However, a good process administrator should be able to prevent this from happening.

To Sackman's comments we might add Ford's concern about "public commitment" (1975). He defines public commitment as a social process whereby stating a position publicly makes it difficult to change the position. "...in the worst case, Delphi does not force rethinking of a problem and thus tolerates the same answer over iterations without thought." (pg. 140)

---

[3] What constitutes "relevant facts" is subject to debate. For instance, the opinions of prestigious experts are arguably relevant to forming an opinion. However, the Delphi method is clearly designed to exclude second order information e.g., about the sources of information.

Sackman's critique provoked a number of responses. The journal *Technological Forecasting and Social Change* published a number of them in 1975 (Coates, Goldschmidt, Schieele). However, for our purposes, most did not substantively address Sackman's itemized comments; rather, they argued in more general terms. For instance, Coates and others argued that the Delphi method provides an essential tool for certain objectives, not all of which may require precise quantitative estimation. Coates writes of his use of the Delphi method,

> "[It] is not a scientific tool, nor is it related to a scientific experiment or a scientifically structured activity. (pg.193)....By the same token, *the criteria in evaluating a Delphi are not so much that it is right but that it is useful* [emphasis added]. The value of the Delphi is not in reporting high reliability consensus data, but rather in alerting the participants to the complexity of issues." (pg.194)

In response to his own critique, Sackman (1976) reported an exploratory variant of the Delphi method intended to correct the method's deficiencies while retaining its valuable features and ideas. This method is described in more detail below in Section C, methodological variants and extensions.

Of all the responses to Sackman's critique (Hill and Fowles, 1975; Coates, 1975; Schieele, 1975; Linstone, 1975), Goldschmidt's (1975) was the most substantive and perhaps the most influential (Rieger, 1986). He correctly argued that many of Sackman's comments regarded the practice of the method rather than its inherent characteristics. However, he did not address the substance of many of Sackman's important concerns about the method itself. For instance, he objected to the social science standards to which Sackman referred in his critique. He explained that the standards were developed for standardized tests and not for the "social experimentation and opinion questionnaires." However, he did not explain why the standards are not sensible for controlling Delphi studies, regardless of their origin.

Goldschmidt also addressed Sackman's concern that Delphi panelists shift their responses to the group median as a result of "extraneous" group social processes. Sackman referred to the consensus as specious and unauthentic. In response, Goldschmidt contested Sackman's choice of words. He argued that

> "...an expressed group opinion does represent a consensus and [that] the way in which the consensus is reached is another matter....Any group opinion is subject to the constraints under which the group formed that opinion." (pg. 205)

However, Goldschmidt does not discuss whether a consensus formed on the basis of conformity to an initial group median response is a useful or desirable result, nor does he summon any evidence to dispute Sackman's claim of a "statistical bandwagon effect."

BIBLIOGRAPHY

N.C. Dalkey. *Experiments in Group Prediction*, The Rand Corporation, P-3820, March 1968.

W.T. Weaver. *Delphi as a Method for Studying the Future: Testing Some Underlying Assumptions*, Educational Policy Research Center, Syracuse, New York, 1970.

J. Pill. "The Delphi Method: Substance, Contexts, A Critique and an Annotated Bibliography," *Socio-Economic Planning Sciences*, (5), 1971, pp 57-71.

W.T. Weaver. *Delphi, A Critical Review*, Syracuse University Research Corporation, RR-7, February 1972.

J.C. Derian, F. Morize. "Delphi in the Assessment of Research and Development Projects," *Futures*, October 1973, pp. 469-483.

H. Sackman. *Delphi Assessment: Expert Opinion, Forecasting, and Group Process*, The Rand Corporation, R-1283-PR, April 1974.

K.Q. Hill, J. Fowles. "The Methodological Worth of the Delphi Forecasting Technique," *Technological Forecasting and Social Change*, 7 (2), 1975, pp. 179-192.

J F. Coates. "In Defense of Delphi: A Review of 'Delphi Assessment, Expert Opinion, Forecasting, and Group Process' by H. Sackman," *Technological Forecasting and Social Change*, 7 (2), 1975, pp. 193-194.

D.A. Ford. "Shang Inquiry as an Alternative to Delphi: Some Experimental Findings," *Technological Forecasting and Social Change*, 7, 1975, pp. 139-169.

P.G. Goldschmidt. "Scientific Inquiry or Political Critique? Remarks on 'Delphi Assessment, Expert Opinion, Forecasting, and Group Process' by H. Sackman," *Technological Forecasting and Social Change*, 7 (2), 1975, pp. 195-213.

D.S. Schieele. "Consumerism Comes to Delphi: Comments on 'Delphi Assessment, Expert Opinion, Forecasting, and Group Process' by H. Sackman," *Technological Forecasting and Social Change*, 7 (2), 1975, pp. 215-219.

K. Brockhoff. "Evaluation: Performance of Forecasting Groups," in H. Linstone, M. Turoff (eds), "The DELPHI Method, Techniques and Applications," Addison-Wesley Publishing Company, Inc.: Reading, MA: 1975.

H.A. Linstone, M. Turoff. "The DELPHI Method, Techniques and Applications," Addison-Wesley Publishing Company, Inc.: Reading, MA: 1975.

I. Jillson. "Developing Guidelines for the Delphi Method," *Technological Forecasting and Social Change*, 7 (2), 1975, pp. 221-222.

H.A. Linstone. "Eight Basic Pitfalls: A Checklist," in Harold Linstone, Murray Turoff (eds.), "The DELPHI Method, Techniques and Applications," Addison-Wesley Publishing Company, Inc.: Reading, MA: 1975.

H. Sackman. *Toward More Effective Use of Expert Opinion: Preliminary Investigation of Participatory Polling for Long-Range Planning*, The Rand Corporation, P-5570, October 1976.

S.J. Press. "Qualitative Controlled Feedback for Forming Group Judgments and Making Decisions," *Journal of the American Statistical Association*, September 1978; see also The Rand Corporation, P-6290, January 1979.

W.E. Riggs. "The Delphi Technique, An Experimental Variation," *Technological Forecasting and Social Change*, 23, 1983, pp. 89-94.

W.G. Rieger. "Directions in Delphi Developments: Dissertations and Their Quality," *Technological Forecasting and Social Change*, 29, 1986, pp. 195-204.

C. METHODOLOGICAL VARIANTS

The purpose of this section is to briefly review "Delphi method variants" which have been developed to correct for weaknesses identified in critical reviews. These studies are reported here to illustrate how the core Delphi method may be extended for requirements of specific studies.

We review three major Delphi variants here, "Shang Inquiry" (SI) (Ford, 1975), "Participatory Polling" (PP) (Sackman, 1976), and "Qualitatively Controlled Feedback" (QCF) (Press, 1979a, b).

Ford developed the SI primarily because he felt that the Delphi method did not force panelists to rethink their answers over iterations. In particular, panelists near the group median response may simply repeat their response again and again, knowing that it is near the group average. The SI attempts to address this by modifying the second defining characteristic of the Delphi method, iterative polling with statistical feedback.

The SI is initialized by developing a response range that contains the true value of the quantity to be estimated; the panelists state a preference for the upper or lower half of the range. The majority response then determines a half range to be used for a subsequent round. The process ends when the initial range has been bisected enough to achieve a specified level of accuracy. Ford assumes that at each iteration, the panelists reconsider the unknown quantity with regard to a new reference point. Of course, this may not be true -- If asked whether a quantity is greater or less than 5, I may estimate it to be 8, and respond accordingly on successive iterations. Ford compared the SI to two Delphi variants in an experiment and reported encouraging results. However, he did not address the important question of whether there are particular circumstances under which SI should be preferred to other methods.

Sackman developed the PP method as an exploratory demonstration ("prototype trial") of how to correct Delphi method deficiencies (Sackman, 1974) while retaining its valuable features and ideas. An objective central to his demonstration was to institute a

"balanced adversarial procedure" whereby statistical feedback would be augmented with the full range of reasoning underlying the all of the panelists' responses.

Sackman provided for a "balanced adversarial procedure" by requiring all panelists to provide written justifications for all their first round responses. In the second round, he provided panelists with essentially all distinct justifications, and required them to rate the importance of each justification before making their second round responses. By doing this, he hoped to force examination and consideration of both the substance and the content of the other panelists' first round responses.

The PP method does not actually represent a significant departure from the core definition of the Delphi method. None of the three defining characteristics precludes soliciting justifications and including them in the feedback. In fact, the only departure from the core method is Sackman's tolerance for situation-specific departures from perfect anonymity. However, in his own study, the departure from perfect anonymity was fairly modest. The panelists met in an open briefing prior to the first round of questions in order to define the area of inquiry and establish a common understanding of its content. Further, Sackman stated that the panelists could contact other panelists to compare views on their own initiative, because "...we believed that more discussion would generally lead to more useful and thoughtful opinion." (pg. 14)

Press' QCF method differs from the Delphi method in that subjects do not receive statistical feedback over iterations. After answering the questions of primary interest, the panelists record the reasons for their response. The process administrators combine these reasons, eliminating duplicates, and feed these back to the panelists for the next round. The process ends when when the panelists do not add any new reasons to the composite from prior rounds.

Press developed the QCF to address the "statistical bandwagon" issue. He observed, like Sackman, that

> "...when feedback is quantitative (say, the median response is fed back), there is an *artificially induced pressure towards consensus*, since panelists will often move their subjective judgments towards the median on successive rounds, if they see that their earlier responses are outliers. (1979a, pg.5)

> "When quantitative measures, such as the mean, are fed back, the panelists are psychologically pressured to shift their answer on the next round towards the given mean. Social psychologists are very familiar with this phenomenon of group conformity." (1979b, pg.3)

He argued that because QCF feeds back only the reasons supporting individual judgments, panelists are likely to shift their responses only on the basis of the logic of the presented reasons and those they might have failed to consider on prior rounds.

BIBLIOGRAPHY

M.T. Bedford. "The Value of 'Comments' Analysis and an Analysis of SPRITE as a Planning Tool," in "Delphi: The Bell Canada Experience," Bell Canada, October 1972.

H. Sackman. *Delphi Assessment: Expert Opinion, Forecasting, and Group Process*, The Rand Corporation, R-1283-PR, April 1974.

D.A. Ford. "Shang Inquiry as an Alternative to Delphi: Some Experimental Findings," *Technological Forecasting and Social Change*, 7, 1975, pp. 139-169.

H. Sackman. *Toward More Effective Use of Expert Opinion: Preliminary Investigation of Participatory Polling for Long-Range Planning,"* The Rand Corporation, P-5570, October 1976.

S.J. Press. "Qualitative Controlled Feedback for Forming Group Judgments and Making Decisions," *Journal of the American Statistical Association*, September, 1978; see also The Rand Corporation, P-6290, January 1979a.

S.J. Press. *An Empirical Study of a New Method for Forming Group Judgments: Qualitative Controlled Feedback,* The Rand Corporation, P-6333, May 1979b.

D. SUPPLEMENTARY TECHNIQUES: CROSS IMPACT ANALYSIS

Unlike the other methods we review in this paper, the Delphi technique is exclusively geared toward data collection rather than modelling or data analysis. As a result, many Delphi studies are methodological hybrids, in which the authors combine the Delphi method with other analytical techniques. There are many possible such combinations.

In this section we introduce one kind of analysis which is associated with using Delphi-derived data to model complex systems -- "cross impact analysis."

In modelling complex systems, we frequently need to represent dependency and feedback relationships between events and entities. There are many ways of doing this. Conventional discrete event computer simulation is one such method. Systems dynamics methods and differential equations provide deterministic means for modelling complex systems (Kane, 1975, Roberts, 1976).

With its use for long-range forecasting, Delphi practitioners also have viewed the path to alternative futures as complex systems -- the occurrence of one event may influence the likelihood of another event; several events may influence each other in this way in a feedback relationship. As a result, a number of methods for modeling complex systems

have evolved in the context of long-range forecasting by the Delphi method, all under the idiom "cross impact analysis".

There is no single cross-impact analysis method because the choice of method depends upon the analytical problem. In addition, there have been disagreements over the best way to do probabilistic cross-impact analysis. Dalkey's (1975) approach uses Bayes' Theorem and the laws of probability to update and make consistent the probabilities of interrelated future events. On the other hand, others have rejected this approach (Turoff, 1972, Helmer, 1977). The debate on how to best perform probabilistic cross-impact analysis has largely appeared in the journals *Futures* (Mitchell and Tydeman, 1976; Godet, 1976; McClean 1976; Kelly, 1976; Helmer, 1977; Mitchell, Tydeman, and Curnow, 1977; Enzer and Alter, 1978) and *Technological Forecasting and Social Change* (Turoff, 1972; Jackson and Lawton, 1976; Godet, 1976).

Apart from the probabilistic methods, the systems dynamics approach to cross-impact analysis taken by Kane (1975) also has been associated with the Delphi method. Kane's work concerns the time-dependent magnitude of system characteristics rather than the time-dependent likelihood of events. An example helps to clarify this idea.

Consider the question, "What is the effect, over time, of increased gasoline taxes on 1) road maintenance costs; 2) taxi fares; 3) the cost of suburban housing?"

In the case of road maintenance costs, the answer might be "a decline over time" because of declining road use.

In the case of taxi fares, the answer is less clear. The response, "an increase" might be justified if the demand for transportation remains the same, while private auto use declines. On the other hand, fares might decline if the industry expanded under the perception of increased demand for taxi services. The purpose of Kane's cross impact method is to develop a representation of a complex system that the mathematics of the method can use to "crank out" predictions of this sort.

The methodology is general enough that it could also be used to address questions more distantly related to the "what if" event, such as that about the effect of a gasoline tax on the cost of suburban housing.

Finally, the supermatrix generalization of Saaty's (1978) analytic hierarchy method is also a cross-impact method. We review the analytic hierarchy process method elsewhere in this paper.

In addition to those papers cited above, we list several below that discuss cross-impact analysis, develop other cross-impact methods, or illustrate cross-impact methods with interesting applications.

BIBLIOGRAPHY

T.J. Gordon, H. Hayward. "Initial Experiments with the Cross-Impact Matrix Method of Forecasting," *Futures*, Vol. 1 No. 2, December 1968, pp. 100-116.

R.C. Amara. "A Note on Cross-Impact Analysis," *Futures*, Vol. 4 No. 3, September 1972.

N.C. Dalkey. "An Elementary Cross-Impact Model," in Harold Linstone, Murray Turoff (eds), "The DELPHI Method, Techniques and Applications," Addison-Wesley Publishing Company: Reading, MA: 1975; see also *Technological Forecasting and Social Change*, Vol 3, 1972, pp.341-351.

J.C. Duperin, M. Godet. "SMIC 74 - A Method for Constructing and Ranking Scenarios," *Futures*, August 1975, pp. 302-312.

J. Kane. "A Primer for a New Cross-Impact Language - KSIM," in Harold Linstone, Murray Turoff (eds), "The DELPHI Method, Techniques and Applications," Addison-Wesley Publishing Company: Reading, MA: 1975; see also *Technological Forecasting and Social Change*, Vol. 4, No. 2, 1972, pp. 129-142.

W.B. Rouse, T.B. Sheridan. "Computer-Aided Group Decision Making: Theory and Practice," *Technological Forecasting and Social Change*, Vol. 7, 1975, pp. 113-126.

M. Turoff. "An Alternative Approach to Cross Impact Analysis," in Harold Linstone, Murray Turoff (eds), "The DELPHI Method, Techniques and Applications," Addison-Wesley Publishing Company: Reading, MA: 1975; see also *Technological Forecasting and Social Change*, Vol. 3, 1972, pp. 309-339.

M. Godet. "Scenarios of Air Transport Development to 1990 by SMIC 74 - A New Cross-Impact Method," *Technological Forecasting and Social Change*, Vol. 9, 1976, pp. 279-288.

M. Godet. "SMIC 74 - A Reply from the Authors," *Futures*, Vol. 8, No. 4, August 1976, pp. 336-339.

J.E. Jackson, W. Lawton. "Some Probability Problems Associated with Cross-Impact Analysis," *Technological Forecasting and Social Change*, Vol. 8, 1976, pp. 263-273.

P. Kelly. "Further Comments on Cross-Impact Analysis," *Futures*, August 1976, pp. 341-345.

M. McClean. "Does Cross-Impact Analysis Have a Future?" *Futures*, August 1976, pp. 345-349.

R.B. Mitchell, J.Tydeman. "A Note on SMIC 74," *Futures*, Vol. 8, No. 1, February 1976, pp. 64-67.

F.S. Roberts. "Discrete Mathematical Models," Englewood Cliffs, N.J.: Prentice Hall, Inc., 1976 (see chapter 4 on pulse process cross-impact analysis models).

O. Helmer. "Problems in Futures Research, Delphi and Causal Cross-Impact Analysis," *Futures*, February 1977, pp. 17-31.

R.B. Mitchell, J.Tydeman, R.Curnow. "Scenario Generation: Limitations and Developments in Cross-Impact Analysis," *Futures*, June 1977, pp. 205-215.

D.W. Bunn, M.M. Mustafaoglu. "Forecasting Political Risk," *Management Science*, Vol. 24, No. 15, 1978, pp. 1557-1567.

S. Enzer and S. Alter. "Cross Impact Analysis and Classical Probability," *Futures*, June 1978, pp. 227-239.

T.L. Saaty. "The Analytic Hierarchy Process," New York: McGraw Hill, 1978.

R.K. Sarin. "A Sequential Approach to Cross-Impact Analysis," *Futures*, Vol. 10, 1978, pp. 53-62.

R.E. Jensen. "Reporting of Management Forecasts: An Eigenvector Model for Elicitation and Review of Forecasts," *Decision Sciences*, Vol. 13, 1982, pp. 15-37.

E. DEFENSE-RELATED APPLICATIONS

A search of the Defense Technical Information Center's (DTIC) database for references on "Delphi techniques" yielded the bibliography appearing in Appendix B of this report. We augment the bibliography below with excerpts from this listing that we believe to be especially valuable to our discussion.

BIBLIOGRAPHY

N.C. Dalkey, O.Helmer. "An Experimental Application of the Delphi Method to the Use of Experts," *Management Science*, Vol. 9, pp 458-467, 1968.

R.G. Leahy, N.B. Ohman. *A Method for Determining an Optimum Reconnaissance Sensor Mix*, Air Force Institute of Technology thesis, GSA/SM/70-07, June 1970, DTIC.

L.J. Sebastiani. *The Delphi Technique and its Applicability to Army Systems Analyses*, U.S. Army Materiel Systems Analysis Agency, Aberdeen Proving Ground Technical Memorandum No. 127, June 1972, DTIC

G.F. Elsbernd. *The Use of the Delphi Method Within the Defense Department*, Auburn University thesis, May 1974, DTIC AD 920 545.

J.C. Duperin, M. Godet. "SMIC 74 - A Method for Constructing and Ranking Scenarios," *Futures*, August 1975, pp. 302-312.

O.A. Larson, S.I. Sander. *Development of Unit Performance Measures Using Delphi Procedures*, Navy Personnel Research and Development Center, San Diego, CA, NPRDC TR 76-12, September 1975.

S.W. Peterson. *Numerical Methods for the Evaluation of Potential Research and Development Contractors*, Army Materiel Command, Texarkana, Texas, USAMC-ITC-02-08-75-214, DTIC AD-A009 415, April 1975.

F.W. Ross. *A Cost-Effectiveness Model, Choice Through Preferences*, Army Aviation Systems Command, St. Louis, Missouri, DTIC AD A006 205, February 1975.

W.L. Brockhhaus, J.F. Mickelsen. *The Delphi Method and its Applications: A Bibliography*, Industrial College of the Armed Forces, DTIC AD A035 463, June 1976.

H. Sackman. *Toward More Effective Use of Expert Opinion: Preliminary Investigation of Participatory Polling for Long-Range Planning*, The Rand Corporation, P-5570, October 1976.

R.B. Mitchell, J.Tydeman, R.Curnow. "Scenario Generation: Limitations and Developments in Cross-Impact Analysis," *Futures*, June 1977, pp. 205-215.

D.W. Bunn, M.M. Mustafaoglu. "Forecasting Political Risk," *Management Science*, Vol. 24, No. 15, 1978, pp. 1557-1567.

J.D. Campbell, J.D. Carlin. *A Description of a Logistically Ideal Aircraft*, DTIC AD-A148 425, September 1984.

D.J. Bonney. *A Quantitative Method for Determining Artillery Basic Loads of Ammunition*, DTIC AD-A161 891, August 1985.

F. BIBLIOGRAPHIES OF STUDIES AND APPLICATIONS

W.T. Weaver. *Delphi, A Critical Review*, Syracuse University Research Corporation, RR-7, February 1972.

H. Sackman. *Delphi Assessment: Expert Opinion, Forecasting, and Group Process*, The Rand Corporation, R-1283-PR, April 1974.

W.L. Brockhhaus, J.F. Mickelsen. *The Delphi Method and its Applications: A Bibliography*, Industrial College of the Armed Forces, DTIC AD A035 463, June 1976.

W.G. Rieger. "Directions in Delphi Developments: Dissertations and Their Quality," *Technological Forecasting and Social Change*, 29, pp. 195-204, 1986.

G. APPLICATION SOFTWARE BIBLIOGRAPHY

W.E. Cundiff. "Interactive Software for the Capture, Management, and Analysis of Data in Delphi Inquiries," *Technological Forecasting and Social Change*, Vol. 28, pp. 173-185, 1985.

W.E. Cundiff. "Interactive Software for the Capture, Management, and Analysis of Data in DELPHI Inquiries: Defined Functions in APL," *Technological Forecasting and Social Change*, Vol. 34, pp. 189-195. 988.

H. EVALUATION AND COMMENTS

We have remarked that the three defining characteristics of the Delphi method allow for considerable variation amongst applications. The potential for variation is so great that for the purpose of evaluation, indeed, for all practical purposes, there is no "Delphi method." Commenting on any particular variant omits a great number of related alternatives; similarly, demonstrating that one particular implementation of the Delphi method is superior in some way to a "non-Delphi" method does not generalize to other Delphi variants or to applications of the same methodology differing in purpose and panelist composition.

The broader question, and the important question, that the Delphi method has brought into relief is how to improve the acquisition of information from a group of respondents.

The developers of the Delphi method were interested in making improvements by reducing the occurrence of "undesirable" group social processes. A secondary goal may have been to make the process inexpensive relative to group discussion methods in terms of necessary resource commitments (e.g., time, level of respondent effort). Naturally, designing a system involves tradeoffs, and the Delphi method is no exception. Thus, for instance, the method's developers and most ardent practitioners felt that anonymity and statistical feedback allowed for greater freedom of expression. However, the method's critics (Sackman, 1974, 1976; Press, 1979a,b) have argued that anonymity and strict numerical feedback limit the reasoning and argumentation that add substance to a collective judgment. Dalkey (1969) characterized the Delphi method as a rapid and relatively efficient way to "cream the tops of the heads" of a group of knowledgeable people. This suggests that the accuracy of Delphi results may be less than that achievable by other methods (i.e., commissioning an in-depth study) under some circumstances.

We therefore feel that the appropriate question to ask is not "What about the Delphi method?", but "How should we intelligently design a process for acquiring information from a group of respondents?" The classically defined Delphi method provides three suggestions for doing this -- anonymity, iterative polling, and statistically defined group response. Later research on the Delphi method made additional suggestions concerning confident panelist subgroups, introducing additional information, and other ideas. A recent report issued by Meyer and Booker (1989) identifies a number of threats to the validity of group judgmental data and suggests ways to counter them.

However, many questions about how to design an effective information acquisition process remain. Is anonymity important? If so, under what conditions? Does feedback of reasons and substantive argumentation help? If so, under what conditions? Is statistical feedback of group response a good idea? If so, under what conditions? What kind of accuracy can we expect from these methods? Under what conditions can we expect this level of accuracy? Is statistically defined group judgment worse than requiring an actual consensus? If so, under what conditions?

Our evaluation thus revolves around a recurring theme. For every idea about the acquisition of information from a group of respondents we should ask -- "Does it help? Under what conditions?" We also should ask the corresponding question about criticisms --

Is the avoidance of adversarial discussions a bad idea? If so, under what conditions? Does the classical Delphi method encourage a "statistical bandwagon effect"? If so, under what conditions?"

Unfortunately, we have observed the Delphi literature to contain many unsubstantiated claims and criticisms. Some authors implicitly appeal to common sense and intuition, which are not necessarily reliable bases for reasoning. Further, some claims and criticisms are justified with reference to laboratory experiments and demonstrations made under limiting "laboratory" environments. As a result, there are real questions about whether we can generalize these results to the breadth of interesting applications (see Winkler and Murphy, 1974 for an example from subjective probability assessment).

The consequences of this state of affairs are twofold. First, the practitioner does not have many general guidelines for designing an information acquisition process that have been empirically evaluated in a "realistic" context. Second, the practitioner does not have many specific guidelines differentiating when some design features work well and when they don't. Naturally, this lack of information also makes evaluation of specific methods like the Delphi variants (Ford, 1975, Sackman, 1976, Press, 1979a,b) difficult, if not impossible.

As a result, the best that we can offer at this point are three suggestions.

First, carefully analyze the proposed study and tailor the methodology in a way that is "sensible" for the requirements of the study. The literature on the Delphi method, on group social processes, and on group problem-solving applications can be a good source for ways to tailor a methodology to particular requirements (e.g., Sackman, 1976, Press, 1979a,b, Meyer and Booker, 1990).

How do we then know that a study design choice is "sensible?" Some of the evidence to support "sensibility" will come from the empirical literature on group problem-solving and group processes. However, one must be careful not to make unwarranted generalizations from studies whose findings should be interpreted narrowly (Winkler and Murphy, 1974). The remainder of support for sensibility will have to come from experience and intuition. In particular, one should seek the advice of a person with experience in conducting group process studies. Applied carefully, experience and intuition may well provide a better basis for designing a data collection method than applying a standard method "in cookbook fashion."

Any design choice (e.g., anonymity vs direct confrontation, "statistical consensus" vs instructed consensus) may carry both costs and benefits. For instance, if the panelists

are mutually respectful acquaintances with a history of cooperative efforts (e.g., "an effective team"), then anonymity may be both unnecessary as well as undesirable. Panelists who know each other well also may know the areas in which each has expertise, or experience which suggests additional weight or consideration be given to his judgments. On the other hand, if the panelists differ greatly in reputation or authority, then anonymity might provide benefits that outweigh its costs.

Second, proposed study designs should be pilot-tested to allow for modifications and fine tuning.

Finally, there are specific steps in the areas of panelist selection, questionnaire design, and statistical controls that the analyst can take to assure the quality of his particular study. Sackman (1974, 1976) reviews many of these steps; we discuss related issues earlier in this document.

REFERENCES

S.E. Asch. "Effects of Group Pressure Upon the Modification and Distortion of Judgments," in E. Maccoby (ed), "Readings in Social Psychology," Third Edition, London: Holt, Rheinhardt, and Winston, 1958.

R.L. Winkler, A.H. Murphy. "Generalizability of Experimental Results," in Von Holstein (ed.), "The Concept of Probability in Psychological Experiments," Dordrecht, Holland: D. Reidel Publishing Company, 1974.

H. Sackman. *Delphi Assessment: Expert Opinion, Forecasting, and Group Process*, The Rand Corporation, R-1283-PR, April 1974.

D.A. Ford. "Shang Inquiry as an Alternative to Delphi: Some Experimental Findings," *Technological Forecasting and Social Change*, Vol. 7, pp. 139-169, 1975.

H. Sackman. *Toward More Effective Use of Expert Opinion: Preliminary Investigation of Participatory Polling for Long-Range Planning*, The Rand Corporation, P-5570, October 1976.

S.J. Press. "Qualitative Controlled Feedback for Forming Group Judgments and Making Decisions," *Journal of the American Statistical Association*, September 1978; see also The Rand Corporation, P-6290, January 1979a.

S.J. Press. *An Empirical Study of a New Method for Forming Group Judgments: Qualitative Controlled Feedback*, The Rand Corporation, P-6333, May 1979b.

M.A. Meyer, J.A. Booker. *Eliciting and Analyzing Expert Judgment, A Practical Guide*, Los Alamos National Laboratory LA-11667-MS, NUREG/CR-542, 1989. (also London: Academic Press, in press.)

III. ANALYTIC HIERARCHY PROCESS

This chapter presents an examination of the Analytic Hierarchy Process. It is organized into four major subheadings, each followed by a (chronological) listing of reference literature relevant to that particular major subheading. Each of the five papers presented in this study follows the same general outline for ease of reader reference and comparison. Before beginning our detailed discussion on the analytical hierarchy process, we offer a brief explanation of its methodology.

A. DESCRIPTION

The Analytic Hierarchy Process (AHP) is a measurement technique developed by Thomas Saaty and his colleagues in the early 1970s. The AHP's purpose is to measure quantities for which expertise allows subjective estimates of relative magnitude. Such quantities include those that fall outside the scope of ordinary physical measurement techniques, but do not exclude those that are conventionally measurable.

The measurements are claimed to be ratio-scaled (Saaty, 1980). Ratio-scaled measurements are determined up to a *similarity transformation* -- multiplication of the measurements by a positive constant. Thus we can rescale the measurements to different units (e.g., pounds to kilograms, dollars to yen) and not affect several kinds of conclusions made about the measurements. Specifically, a similarity transformation on ratio-scaled measurements does not change 1) rank orderings (Is *A* or *B* heavier?); 2) differences when corrected for a change in units (How much heavier is *A* than *B*?); or 3) ratios (How many times is *A* heavier than *B*?).[1]

---

[1] Consider the ratio-scaled measurement of *A* in terms of some attribute to be *a* and that of *B* to be *b*, and $a>b$. Multiplication by a positive constant *k* does not change the rank order of *A* and *B* in terms of that attribute ($ka > kb$). That is, *A* is heavier than *B* whether we compare their weights in kilograms or in pounds, where kgs=2.2xlbs. Nor does a similarity transformation change the difference between the measurements of *A* and *B* ($a-b$) after the unit of measurement is taken into consideration ($(ka-kb)/k=a-b$). The measured difference in weight between a 1 kilogram object and a 2 kilogram object does not change when the objects are measured in pounds if the difference in pounds is corrected for the change in units. Finally, multiplication by a positive constant doesn't change the ratio of two measurements ($(ka/kb)=(a/b)$). If *A* weighs twice as much as *B*, then the ratio of the measurements will be the same whether we weigh them in pounds or in kilograms.

Physical temperature, unlike weight, is frequently not a ratio-scaled measurement. As a result, we may lose the invariance of ratios after a similarity transformation. For instance, the Celsius scale, in which water freezes at 0° and boils at 100°, is not a ratio-scaled measurement system. The "warmth" ratio between physical

1. Main Components

The AHP has three main components: hierarchical decomposition, pairwise judgment, and synthesis of overall ratings. In the terminology of the AHP, the ratings are called weights, or *priorities*. We will use the terms "weights" and "priorities" interchangeably here.

a. Decomposition

The purpose of decomposition is to facilitate the analysis of a system by breaking it down into components. Decomposition facilitates the analysis of a system when 1) there is "better" knowledge (i.e., more accurate, easier to obtain) regarding the components and their relationships than that regarding the system as a whole; and 2) there is a method for aggregating the knowledge of the parts that preserves this superiority. The form that decomposition usually takes in the AHP is decomposition into a hierarchy.[2] Decomposition of systems into more general network structures is allowed and may be necessary under some circumstances. Saaty has developed a generalization of the AHP to handle these cases. We discuss these cases below in Section 3.c. on AHP assumptions regarding independence of hierarchy elements.

Within the AHP, an overall goal (e.g., rank the alternatives, determine the expected outcome, derive weights for allocating resources) generally occupies the single node at the top of the hierarchy. The attributes, forces, or criteria (henceforth referred to as "criteria") that bear on accomplishing the goal occupy the next level of the hierarchy. Successively decomposing these elements and their descendents yields the sub-elements through the next to last level of the hierarchy. The alternatives constituting the decision space occupy the lowest level of the hierarchy and are nested underneath the lowest level criteria.

The alternatives may be "alternatives" in the conventional sense of the term (e.g., weapon systems, research and development projects); however, they also may be more general characteristics, or elements. For instance, alternative "probabilities" of an event can be viewed as alternatives (e.g., 0-20 percent, 21-40 percent, etc.). In assessing the amount of resources to allocate to one of several projects, we are interested in ordering or weighting the alternatives, as

---

temperatures of 1° C and .01° C, both of which are within a degree of freezing, is not equivalent to that between 100° C and 1° C. The reason that temperature on the Celsius scale is not a ratio scaled measurement is that the lowest temperature is -273.15° rather than 0°. Correcting for the lowest temperature, the "warmth" ratio between 1° C and .01° is only about 1.003 (274.15/273.25), whereas that between 100° C and 1° C is about 1.361. These ratios are certainly closer to our felt intuitions.

2 We use the term according to its usual meaning. However, for a more precise reference, Saaty (1980, 1986) has given the concept of *hierarchy* a more precise definition.

opposed to choosing one or more from the set. However, for the purpose of this paper, we will generally refer to the elements at the lowest level of the hierarchy as *alternatives* or *options*.

Most hierarchies are *complete* (Saaty, 1980, pg. 42), that is, all alternatives are judged with respect to all criteria at the next higher level in the hierarchy. However, some hierarchies may not be complete. For example, we cannot legitimately compare walking or bicycling to a bus with regard to relative carbon monoxide emission levels -- ratios containing a zero (e.g., 5/0 or 0/5) are undefined. Incomplete hierarchies present a problem -- all things being equa', incomplete comparison sets receive normalized priorities that are larger than complete comparison sets. For instance, alternatives A, B, and C may have weights of 6/10, 3/10, and 1/10 on one criterion, and alternatives A and B may have weights of 1/3 and 2/3 on another, equally important, criterion. Although alternatives A and B each dominate the other by a 2 to 1 ratio on one criterion, the absence of alternative C from one comparison gives alternative B a greater total weight, an artifact of the incomplete hierarchy.

Saaty (1980, pg. 42) suggests weighing the priorities of each comparison set "by the ratio of the number of elements in that set to the total number of all [alternatives]'. (Dr. L. Vargas has suggested in a personal communication that this approach is only intended to apply to absolute measurement in the AHP (see Section C.2)). However, simple counterexamples show this not to be correct. At the time this review was written, this issue had not yet been satisfactorily addressed, although approaches have been suggested (L. Vargas, personal communication; W. Wedley, personal communication).

Figure III-1 is a hierarchy adapted from Tullington, Batcher, and Guess (1985) for evaluating candidate infantry rifles. In the example, rifle effectiveness is decomposed into three criteria -- performance, suitability and supportability. These criteria are themselves decomposed into constituents. Performance, for instance, comprises target acquisition ability, penetration ability, lethality, hit probability, and volume of fire capability. Target acquisition ability itself takes place under conditions of day, night, and smoke/fog. Volume of fire capability is a function of the stowed ammunition load, the number of rounds carried in the rifle, the sustained rate-of-fire, and the rifle's reliability during operation. Figure III-1 does not show the lowest level of the hierarchy, in which the candidate rifles are nested under each element of the lowest level of the hierarchy (e.g., day, night, smoke-fog, lethality, hit probability).

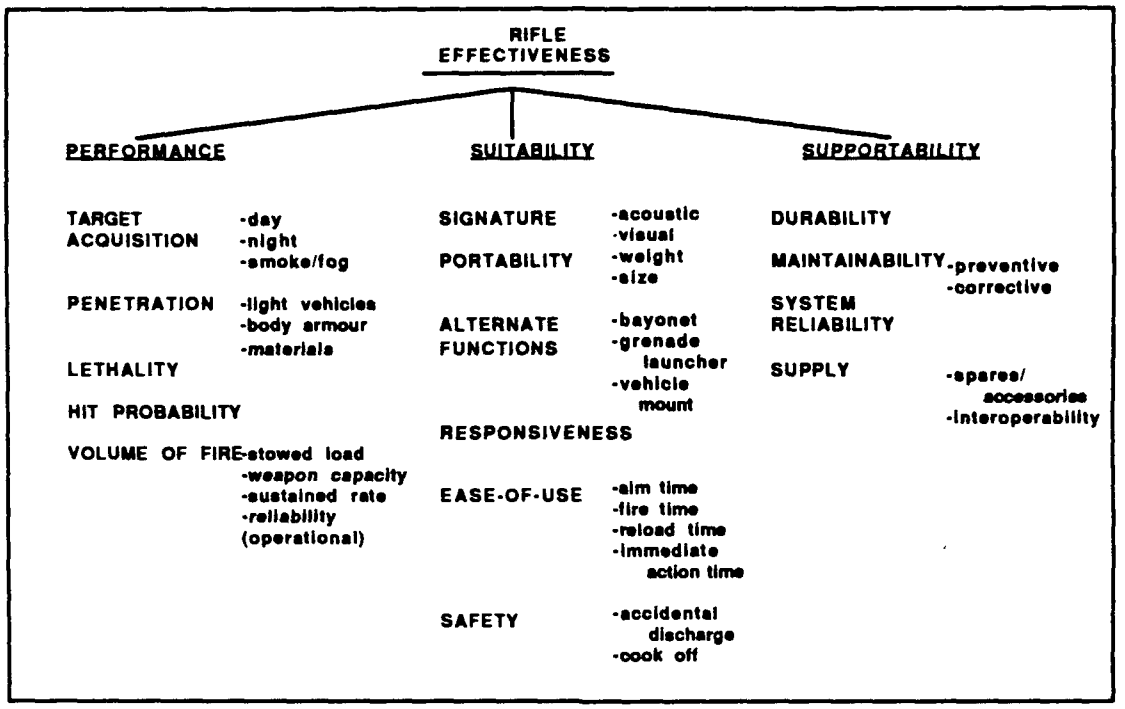


Figure III-1. Rifle Effectiveness Hierarchy (adapted from Tullington, Batcher & Guess, 1985)

b. Pairwise Judgments

The second component of the AHP requires collecting pairwise judgments among all elements descended from a common parent one level higher in the hierarchy. The meaning of the term "priority" varies according to the content of the problem, and alternately is interpreted as: importance, intensity, dominance, priority, preference, weight, value, likelihood of occurrence, etc. The judgments regard the relative "priority" of the elements. In the rifle effectiveness example, this means pairwise comparisons among Performance, Suitability and Supportability to determine their relative importance as "determinants" of rifle effectiveness. Within Performance, we require the relative priorities of target acquisition, penetration, lethality, hit probability, and volume of fire as "determinants" of rifle performance. Within target acquisition, we require the relative priorities of the day, night, and smoke/fog elements as conditions under which target acquisition takes place. Within the signature component we require the relative priorities of its acoustic and visual constituents. At the lowest level of the hierarchy, we require pairwise comparisons among the rifle candidates with respect to each lowest level performance criteria (e.g., target acquisition during day, during night, and in smoke/fog; lethality; acoustic and visual signatures; durability; etc).

Judgments of relative priority are ratio-type judgments. The statement "rifle #1 is twice as good as rifle #2 in terms of system reliability" might be one such judgment. The AHP allows ratio-type judgments to be made in terms of either verbal/categorical scales or numerical scales.

The permitted numerical responses comparing a stronger to a weaker element vary between 1 and 9. Those comparing a weaker to a stronger element vary between 1 and 1/9th.

The responses on the categorical scale are: equal, weakly more, strongly more, very strongly more, or demonstrably more and absolutely more. These categorical responses respectively correspond to the odd-valued numerical responses. The even-valued numerical responses correspond to between-category responses on the categorical scale. For the purposes of the review, we consider only the numerical scale.

The AHP requires forcing the numerical valuation of how any element *I* compares to any element *J* to be the reciprocal of how *I* compares to *J*. For example, if we judge one criteria to be four times as important than another, then the second criterion must be valued as being only a fourth as important as the first. As a special case, the value associated with the comparison of any element with itself is taken to be 1.

c. Synthesis of Overall Priorities

The third component of the method requires determining the overall priorities for the alternatives. This involves two activities: For each of the elements compared among each other, whether alternatives with respect to criteria (e.g., rifles with respect to "hit probability") or criteria with respect to higher order criteria (e.g, the weight given to day, night, and smoke/fog conditions as subcriteria of target acquisition), the procedure estimates the "true" priorities underlying the pairwise judgments. Across the entire hierarchy, we weight the priority estimated for each element by the priority of its parent one level up in the hierarchy. At the end, we compute the priority of a given alternative by summing the priorities computed for it under each criterion with respect to which it has been evaluated.

(1) Eigenvector Prioritization

The methodology for estimating the priorities underlying the comparative judgments is one of the major methodological innovations represented by the AHP. However, Saaty and others continue to discuss the relative merits of alternative estimators. We discuss these alternative methods below in Sections B and C on critiques and extensions of the AHP, respectively. In this section we describe Saaty's original eigenvector prioritization method.

Let $A=[a_{ij}]$ be the nxn matrix of judgments comparing element i with element j (e.g., element i is a_{ij} times as important as element j with respect to some criteria.). We force $a_{ii}=1$ and we typically set $a_{ji}=1/a_{ij}$., although the method does not require forcing equality in this way. As a result, A is a special kind of matrix called a *positive reciprocal matrix*.

Saaty's method for estimating the priorities underlying a positive reciprocal judgment matrix is to determine its principal eigenvector. The rationale for doing this is relatively straightforward and is easily summarized.

Let W_i be the true priority (e.g., weight, intensity, value, importance, etc.) of the ith element. Under perfect consistency in judging relative priorities the responses a_{ij} are assumed to equal the ratio W_i / W_j. As a result, $a_{ik}= (a_{ij})\ (a_{jk})$ for all j because $a_{ik}=W_i / W_k=(W_i / W_j)(W_j / W_k)$). For example, if I is twice as good as J ($a_{ij}=2$) and J is three times as good as K ($a_{jk}=3$), then I is six times as good as K ($a_{ik}=6$). As a result, $(a_{ij})(W_j) = (w_i)$ for all j, and the sum over all j $\{j=1..n\}$ of the product $(a_{ij})(W_j)$ equals $(n)(W_i)$. This is true regardless of the value of i, $\{i=1..n\}$.

We express this summation in matrix notation as $A^iW=n(W)$, where A^i is the *ith* row of A and W is the vector of priorities W_i) underlying the judgments $[a_{ij}]$. We can include all rows of A in the summation with the expression $AW=nW$.

The solutions λ and x to any relationship of the form $Ax=\lambda x$, where x is a vector and λ is a scalar, are respectively the eigenvalues of the matrix A and their associated eigenvectors. When the judgments are *consistent* in a positive reciprocal matrix, that is, when all $(a_{ij})=(W_i)/(W_j)=(a_{ik})(a_{kj})$, all but one of the eigenvalues will equal zero. The nonzero eigenvalue will be n and its associated eigenvector in the solution will thus be the vector of priorities underlying the judgment data.

Most matrices will not be consistent because of inconsistencies in judgment (e.g., measurement error), biases, shifting judgment criteria, and other causes. For instance, suppose a respondent judges A to be twice as important as B and B to be three times as important as C. A consistent respondent would judge A to be six times as important as C. However, we generally do not expect the respondent to make the perfectly consistent response. Further, the consistent response may be less obvious when using the verbal/categorical scale (eg, A is between "equally" as important as B and "weakly more" important than B and B is "weakly more" important than C, implies that A is between strongly and very strongly more important than C).

As a result of inconsistency, a_{ij} generally will not equal W_i/W_j and $(a_{ij})(W_j)$ generally will not yield the same estimate for W_i over all j. For instance, suppose we know the priority of A to

be 6 (W_A=6), that of B to be 4, and that of C to be 2. We then expect the comparison of A to B to yield a response of 1.5 (a_{AB}=6/4) and that of A to C to yield a response of 3 (a_{AC}=6/2). Multiplying a_{AB} by W_B gives W_A (1.5x4=6), as does multiplying a_{AC} by W_C. However, if the respondent is not exact in making a response, the response to A and B might be 2. As a result, $(a_{AB})(W_B)$=8 (2x4).

If the judgment inconsistencies are "scattered" around the true value of (W_i/W_j) then an "average" of several products $(a_{ij})(W_j)$ over j might serve as a reasonable estimate of W_i. In order to identify one such function we refer to three results from linear algebra.

First, theorems due to Perron and Frebonius assure that positive reciprocal matrices have a positive eigenvalue that exceeds all other eigenvalues in absolute value. Associated with this eigenvalue is an eigenvector that is positive in all of its components and is determined up to a similarity transformation (Saaty, 1980).

Second, where there is a solution to the eigenvector problem for positive reciprocal matrices, the n eigenvalues sum to n. In the case of the consistent matrix A, we have seen that there is one nonzero eigenvalue with value n.

Third, the eigenvalues and eigenvectors of a consistent positive reciprocal matrix A change only by a small amount as a result of "small" perturbations to its elements a_{ij}. The model of perturbations underlying inconsistent judgments is multiplicative. These results hold because the eigenvectors and eigenvalues of a matrix depend continuously on the components of the matrix A. That is, $a_{ij}=(W_i/W_j)(1+d_{ij})$, where d_{ij} is the proportion which a_{ij} differs from (W_i/W_j).

As a result of the third assertion, the largest eigenvalue of an inconsistent judgment matrix is close to n and the remaining eigenvalues are close to zero. The eigenvector associated with this largest eigenvalue is similarly close to that which would have been obtained had we not perturbed the matrix of judgments. The *principal eigenvector* thus represents an estimate of the unknown values of the W_j. We denote these estimates in this paper by w_j.

(2) Hierarchic Composition

The AHP forms priorities for each of the alternatives at the bottom of the hierarchy through an "averaging" procedure that Saaty terms *hierarchic composition*. Replace the priority w_j of each element in the third level of the hierarchy by the product of that priority and the priority of the element in the second level to which it is subordinate. Repeat this process through the lowest level of the hierarchy. At the lowest level, sum the weighted priorities over each distinct alternative.

Saaty and others have provided numerous examples of hierarchic composition (see the bibliography of published applications listed at the end of this section.). We have adapted the rifle effectiveness hierarchy illustrated in Figure III-1 to demonstrate hierarchic composition here.

Figure III-2 presents a simplified version of the hierarchy presented in Figure III-1 in which we have omitted some elements of the hierarchy. We have also associated with each element a priority estimated for it.

The hierarchy for rifle effectiveness displayed in Figure III-2 is a four level system. At the top is the main focus -- determine the effectiveness of the four candidate rifles. At the second level, we display the three criteria contributing to rifle effectiveness. Performance, suitability, and supportability have estimated priorities of 0.5, 0.3, and 0.2, respectively. The performance factor consists of two components, target penetration capability and target acquisition capability, weighted .6 and .4, respectively. However, we must weight these priorities by the 0.5 priority of their parent, performance. Thus, we replace the priorities determined as a result of pairwise comparison by 0.3 (0.6x0.5) and 0.2 (0.4x0.5).

We compare three rifles at the lowest levels of the hierarchy with regard to the criteria to which they are subordinate. Thus, the three rifles have been weighted 0.3, 0.3, and 0.4 with regard to their capability to penetrate targets. However, because the weighted priority of penetration is 0.3, we weight the three rifles' priority under penetration by 0.3. Thus, we replace the respective priorities of 0.3, 0.3, and 0.4 by 0.09 (0.3x0.3), 0.09, and 0.12 (0.4x0.3). We similarly weight the priorities of the three rifles under target acquisition, suitability, and supportability.

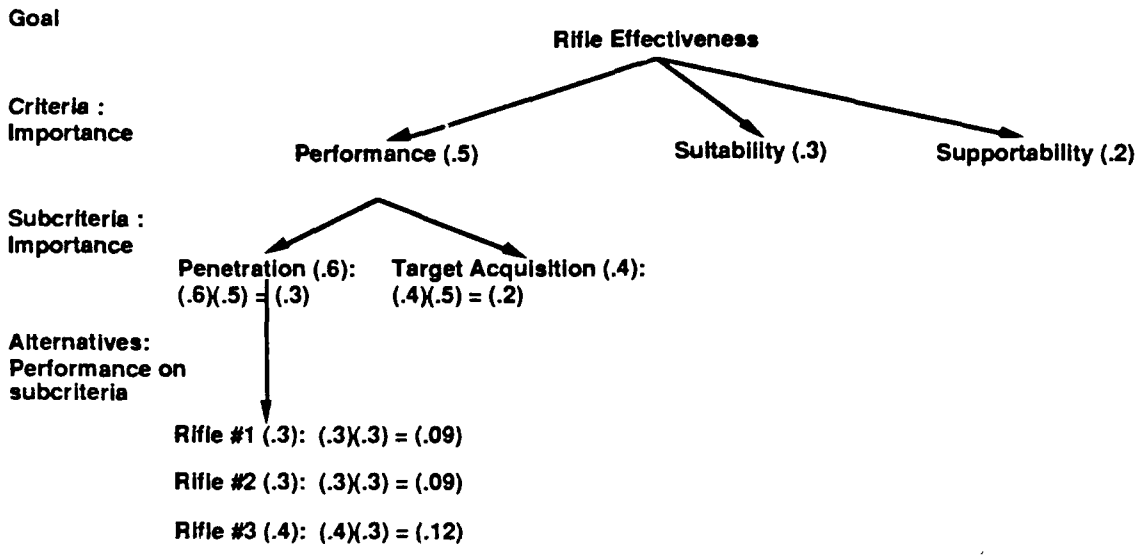


Figure III-2. Rifle Effectiveness (adapted from Tullington, Batcher & Guess, 1985)

Finally, we add the weighted priorities of each rifle wherever they occur in the hierarchy to determine its overall priority. The sum of these overall priorities will sum to 1.0.

2. Indices of Consistency

A second of Saaty's important innovations is a consistency index for assessing the aggregate consistency of pairwise comparisons. Recall that in the case of perfect consistency (e.g., no measurement error), the largest positive eigenvalue associated with the nxn positive reciprocal matrix of comparisons (A) has the value n. However, with perturbations, the maximum eigenvalue increases by a "small amount". Thus, the departure of the maximum eigenvalue from n indexes departures from perfect consistency in the A matrix. Saaty has developed two measures of inconsistency in order to more precisely assess its presence in A (see Belton, 1986; Golden and Wang, 1989; and McCurdy, 1989 for alternative procedures for measuring inconsistency).

The consistency index (CI) measures departure from consistency. CI equals $(\lambda_{max}-n)/(n-1)$, where λ_{max} is the maximum eigenvalue of the matrix of judgments. Saaty (1986a) has demonstrated that this function is convex, achieving its minimum of 0 when the matrix is perfectly consistent. He has also demonstrated that [(2)(CI)] is an estimate of the variance of d_{ij}, the error component underlying inconsistent judgments, where $a_{ij}=(W_i/W_j)(1+d_{ij})$.

The consistency ratio (CR) is the ratio between the CI of a matrix and the average of the CIs of "randomly" generated matrices (RI) of identical size. Saaty (1980, pg. 21) provides RIs for

matrix samples of 500 drawn for sizes up to 11 and for samples of 100 for sizes from 12 to 15. Saaty (1980, pg. 62) also presents the means and variances of RIs for samples of 100 matrices ranging in size from n=2 to n=15.

In judging the relative inconsistency of a judgment matrix, Saaty (1980, pg. 21) suggests that a CR no larger than 0.10 is considered acceptable. Crawford and Williams have remarked that "...because the eigenvector does not fit into any standard statistical framework, there is no readily available device against which deviations from consistency can be measured, (Crawford and Williams, 1984). However, Vargas (1982) lends support to this decision rule, showing that

> "...random consistency follows a truncated normal distribution and that an acceptable upper bound of the ratio between the consistency of a reciprocal matrix and its corresponding average random consistency is 10%. (Vargas, 1982, pg. 80)

In order to provide an empirical basis for judging consistency statistics, Budescu, Zwick and Rapoport (1987) performed a Monte Carlo study of the sampling distribution of CI. They estimated the distribution of the CI for "random" matrices of size 4, 6, 8, 10, and 12 constructed from pseudo-random variates "uniformly distributed" on the interval [1/9,9]. Consistent with AHP practice, they also forced a_{ii}=1 and a_{ji}=1/a_{ij}.

Budescu (et al) present tables listing lower-tailed 1 percent, 5 percent, and 10 percent critical values for CI. That is, when a calculated CI is less than a tabled critical value (e.g., 0.471 for a matrix of size 6 and a significance level of 0.05) we should reject the null hypothesis that the CI was calculated from a randomly generated matrix. For the 0.05 level of protection these critical values range from 0.09 for a matrix of size 4 to 0.804 for a matrix of size 8. Note that these imply CRs that are somewhat more liberal than the "critical value" of 0.10 suggested by Saaty (1980) (eg, (.804/1.41) = .57 > .10, where 1.41 is the consistency index computed for a "random matrix" of order 8. Saaty, 1980, pp. 21). In addition, Kamenentzky (1982) has remarked that the CI may be high when "consistency is high" but certain independence assumptions are not satisfied (pg 711, see section 3.c below).

The authors also estimated regression functions for 1 percent, 5 percent, and 10 percent level critical values with matrix size as a variable. We compared the critical values estimated with these functions against the corresponding tabled critical values and found uncomfortably large differences. In addition, some critical values computed from the regression equations were negative. For a matrix of size 4 and sampling distribution tails of 1 percent, 5 percent, and 10 percent, Budescu's tables give critical values of 0.148, 0.090, and 0.035. However, their regression equations give critical values of -0.082, 0.039, and -0.069. For matrices of size 6, the respective tabled values are 0.596, 0.471, and 0.237. However, the regression function gives

critical values of 1.021, 0.925, and 0.365. We therefore suggest that Budescu's regression equations be used only with great care.

It may be advisable to take remedial steps when a judgment matrix has been determined to be "too inconsistent." Indeed, this is implicit in the act of defining consistency indices and identifying 0.1 as a "critical value." However, there is little clear guidance in the literature on the best procedures for remediating an excessively inconsistent matrix.

In some analyses, CRs that are large by Saaty's standard are reported but not remediated. Saaty (1980) himself has reported such results. On the other hand, some software implementing the AHP suggest reconsidering judgment matrices with excessive CRs. Some software additionally point out matrix entries that contribute the most to the overall inconsistency, implying that reconsideration should be given to these judgments first.

However, the problem of dealing with inconsistency in the AHP has some of the flavor of the problem of handling influential data and outliers in statistics, and in regression analysis in particular. Intervening in the data improves the overall fit, but at the cost of discarding information. Thus, excessive zeal in modifying judgments may tailor the priority estimates to well-fitting data points, but otherwise compromise the integrity of the results.

With regard to the AHP, it also is not clear that judgments contributing the most to overall inconsistency are those that should be changed. If they are not, then the results may be biased away from the underlying priorities.

3. AHP Assumptions

Saaty (1986a; also Harker and Vargas, 1987) has developed four axioms that underlie the AHP. We discuss three of them here as important underlying assumptions (The fourth axiom simply requires all alternatives and criteria to be represented in the problem hierarchy).

a. Reciprocity Assumption

The subjective priority (i.e., importance, strength, value) of an element i compared to an element j (a_{ij}) is the reciprocal of the subjective priority of the elements compared in reverse order, ($a_{ij}=1/a_{ji}$). In practice, we take this assumption as given and force the reciprocal relationship between each of the a_{ij} and a_{ji} pairs. However, we should note that there may be problems with this approach.

Evidence from the research on the psychology of language (psycholinguistics) and the psychology of judgment and decision making suggests that responses to a question like "How

does the weight of the less-valued element compare to that of the more-valued element?" may systematically vary from the reciprocal of those answering questions like "How does the weight of the more-valued element compare to that of the less-valued element?" Treating such a systematic difference as a context effect would require pairwise comparisons to be collected for the entire $[a_{ij}]$, $(i \neq j)$ matrix. The reciprocity requirement for the $[a_{ij}]$ matrix would then require a procedure for reconciling inevitable inconsistencies between. the a_{ij} and a_{ji} elements.

b. Homogeneity Assumption

The relative priorities of the elements being compared should be no greater than some constant K ($\frac{1}{K} \leq (w_i/w_j) \leq K$ for all i,j)). Saaty refers to this condition as *ρ homogeneity*. Saaty (1980) suggests that if any two elements violate this assumption, then the set of elements being be compared should be subdivided so that the relationship is satisfied within the subsets and between the subsets. Although the homogeneity assumption does not require any particular value for K, a value of 9 has been generally used in practice (Saaty, 1980).

Consider the weights of 4 stones respectively weighing 1, 2, 5, and 10 pounds. As it stands, the set violates *ρ homogeneity* for $K = 9$ because $10/1 > 9$. One partition of the stones which remediates this violation is to put the two lightest stones in one set and the two heaviest stones in another. The average weight of the stones in the second set is five times that of those in the first set and the weights of the stones in each set are comfortably within a range of nine to one. Clustering in this way can also be used as a strategy for reducing the number of pairwise comparisons to be made among a large set of criteria or alternatives.

Harker and Vargas (1990) demonstrate that violation of this assumption may not "materially" effect the results. However, systematic research has not be done to suggest under what conditions this may be true.

c. Independence Assumption

The third assumption regards dependence between elements. However, before continuing, we require some terminology. Saaty (1980) and Saaty and Takizawa (1986) have defined several terms for discussing dependence in systems.

> "Functional dependence is what we usually understand by the dependence of one set of elements on another set of attributes or criteria used to compare or score them.
>
> "Functional dependence itself may be between sets [i.e. between elements in adjacent hierarchy levels] or within a set [i.e. among elements in the same hierarchy

level]. The former is called outer dependence of one set on another. The latter is called inner dependence where the elements of a set are on the one hand outer dependent on a second set, and on the other conditionally dependent among themselves with respect to the elements of the second set which serve as attributes." (pg. 230)

The assumption required for the AHP is that there should be neither inner dependence among elements at a given level nor outer dependence *of criteria on alternatives*. Two examples developed by Saaty and Takizawa depict inner and outer dependence and illustrate why the core AHP methodology cannot handle them. For all cases, Saaty and Takizawa also illustrate how to properly collect comparison data and synthesize overall priorities from them.

In the first example, Saaty and Takizawa consider the goal of prioritizing job promotion candidates. The hierarchy has three criteria, research record, teaching ability, and community service.

The problem illustrates outer dependence because the importance of the criteria with regard to evaluating the candidates might differ among the candidates. This could happen when the candidates types are heterogeneous. Hypothetically, teaching ability might be the most important criteria for judging a dance teacher, publication record might be the most important criteria for judging a mathematics teacher, and community service might be the most important criteria for judging a social work teacher.

In order to analyze systems having this kind of outer dependence, Saaty (1980) has developed a generalization of the core AHP methodology called the *supermatrix method*. Harker and Vargas (1987) explain the rationale underlying the method in non-technical language using a network model.

The principal methodological change introduced by the supermatrix methodology is that for all criteria that an alternative K depends on, we require a judgment of relatively how important a criterion A is compared to a criterion B with respect to evaluating alternative K. Relative priorities for alternatives with respect to criteria and for criteria with respect to alternatives can be separately obtained by eigenvector prioritization. Denote the $i + j$ order matrix which contains these priorities as $U = [u_{xy}]$. The element $U_{j,(j+i)}$ contains the priority of the i^{th} alternative with respect to the j^{th} criterion; $U_{(j+i),j}$ contains the priority of the j^{th} criterion with respect to the i^{th} alternative. The limiting value of $\lim_{K \to \infty} U^{2K+1}$ is a matrix which contains revised priorities that take into account the mutual dependence between criteria and alternatives.

In the second example, Saaty and Takizawa illustrate a hierarchy with inner dependence. The purpose of the system is to evaluate the relative importance of several motorcycle subsystems (i.e., brakes, steering, body frame, engine, etc.) to the operation of the motorcycle as a whole. The functions defining motorcycle operation are stopping, turning, running, accelerating. In this example, several of the functions are related to each other. As a result, the priority of any function depends of those of the others.

The analysis requires three steps: 1) prioritizing the functions with respect to their importance to motorcycle operation. Which function should be emphasized more in a motorcycle and by how much more?; 2) prioritizing the subsystems to each other with respect to each function. Which subsystem is more important to this function and by how much?; and 3) comparing the functions to each other with respect to their influence on a given function. This last takes into account dependencies among the functions. For instance, turning requires the ability to stop the wheels from rotating while controlling the acceleration of the engine. With regard to turning, is stopping or accelerating more important, and by how much?

Generalizing the core AHP methodology to this case of inner dependence actually requires no changes to the method if the problem is formulated correctly. The second level of the hierarchy is motorcycle function. The third level of the hierarchy also has motorcycle functions, but subordinate to motorcycle functions in the second level. The third level contains the motorcycle subsystems. We estimate priorities and hierarchically compose overall priorities on this hierarchy in the usual way.

A number of authors (e.g. Dyer and Wendell, 1985; Schoner and Wedley, 1989; Dyer, 1990a, 1990b) have suggested that the assumption of independence of criteria weights from alternatives is never satisfied, and that this violation introduces serious problems. The treatment of this argument in the critical literature is discussed below. Other authors have discussed assumptions underlying the AHP that are less frequently discussed or are taken for granted (eg, that respondents can unambiguously interpret comparison questions and can made ratio-scaled comparisons). We discuss these below in Section B below.

4. AHP AND UTILITY THEORY

As a method for numerically evaluating preferences among alternatives, the AHP would appear to share some underlying features with the multiattribute value (MAV) and multiattribute utility (MAU) approaches to assessing preferences (Kamenetzky, 1982; Dyer and Wendel, 1985; Belton, 1986; Zahedi, 1987; Dyer, 1990a, 1990b). In fact, some researchers argue that the AHP is essentially an additive MAV method . It is claimed that in both cases, the scores (e.g. priorities,

values) of alternatives consist of a linear function (eg, weighted average) of their scores with respect to the decision criteria. The criteria weights constitute the coefficients (eg, the weights of the weighted average) of this linear function.

Dyer has, in part, formulated his critique of the AHP around the relationship he observes between the AHP and multiattribute utility /MAV theory (Section B, below.). In reply, Saaty and others (Saaty, 1986, 1990; Vargas 1987, 1989; Harker and Vargas, 1990; also Vargas 1986) argue that the AHP is distinct from the MAU/MAV approach to measuring preferences.

BIBLIOGRAPHY

T.L. Saaty. "Physics As Decision Theory," Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA, (undated).

T.L. Saaty. *Decisionmaking, Scaling and Number Crunching,* Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA, (undated).

T.L. Saaty. "Dependence and Independence in the Analytic Hierarchy Process," Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA, (undated).

T.L. Saaty."What is the Analytic Hierarchy Process?" Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA, (undated).

T.L. Saaty. "Interaction and Impacts in Hierarchical Systems," in *Decision Information for Tactical Command and Control Workshop*, September 1976.

T.L. Saaty. "A Scaling Method For Priorities in Hierarchical Structures," *Journal of Mathematical Psychology*, Vol. 17, 1977, pp. 234-281.

T.L. Saaty. "Exploring the Interface Between Hierarchies, Multiple Objectives, and Fuzzy Sets," *Fuzzy Sets and Systems*, Vol. 1, 1978, pp. 57-68.

C.R. Johnson, W.B. Beine, and T.J. Wang. "Right-left Asymmetry in An Eigenvector Ranking Procedure," *Journal of Mathematical Psychology*, Vol. 19, 1979, pp. 61-64.

T.L. Saaty. "The Analytic Hierarchy Process," New York: McGraw-Hill, Inc. 1980.

T.L. Saaty. "Priorities in Systems With Feedback," *International Journal of Systems, Measurement and Decisions*, Vol. 1, 1981, pp. 24-38.

T.L. Saaty, J. Alexander. "Thinking with Models," Oxford, England: Pergamon Press, 1981.

R. Kamenetzky. "The Relationship Between the Analytic Hierarchy Process and the Additive Value Function," *Decision Sciences*, Vol. 13, 1982, pp. 702-713.

T.L. Saaty. "Decisionmaking For Leaders: The Analytic Hierarchy Process," Belmont, CA: Lifetime Learning Publications, 1982.

L. Vargas. "Reciprocal Matrices With Random Coefficients," *Mathematical Modelling*, Vol. 3, 1982, pp. 69-81.

L. Vargas and J.J. Dougherty. "The Analytic Hierarchy Process and Multicriterion Decision Making," *American Journal of Mathematical and Management Sciences*, Vol. 2 (1), 1982, pp. 59-92.

J. Aczel, and T.L. Saaty. "Procedures For Synthesizing Ratio Judgments," *Journal of Mathematical Psychology*, Vol. 27, 1983, pp. 93-102.

L. Vargas. "Analysis of Sensitivity of Reciprocal Matrices," *Applied Mathematics and Computation*, Vol. 12, 1983, pp. 301-320.

T.L. Saaty. "The Analytic Hierarchy Process: Decision Making in Complex Environments," in R. Avenhaus and R.K. Huber (eds), *Quantitative Assessments In Arms Control*, New York: Plenum Press, 1984, pp. 285-308.

J. Dyer & Wendell. *A Critique of ihe Analytic Hierarchy Method.* University of Texas Graduate School of Business Working Paper 84/85-4-24, 1985.

T.L. Saaty, K. Kearns. "Analytical Planning: The Organization of Systems," Oxford, England: Pergamon Press, 1985.

T.L. Saaty, L.G. Vargas. "The Logic of Priorities: Applications in Business, Energy, Health, and Transportation," Boston: Kluwer-Nijoff Publishing, 1985.

B. Tullington, R. Batcher, K. Guess. *The Evaluation of Individual Weapon Effectiveness: Part I - The Hierarchical Analysis*, Battelle Columbus Division (DTIC AD-B101 602), September 1985.

J. Aczel, and C. Alsina. "On Synthesis of Judgements," *Socio-Economic Planning Sciences*, Vol. 20, No. 6, 1986, pp. 333-339.

V. Belton. "A Comparison of the Analytic Hierarchy Process and a Simple Multi-attribute Value Function," *European Journal of Operational Research*, Vol. 26, 1986, pp. 7-21.

E. Forman. "In Search of the Right Model," *Telematics and Informatics*, Vol. 3, No. 4, 1986, pp. 229-235.

T.L. Saaty. "Axiomatic Foundation of the Analytic Hierarchy Process," *Management Science*, Vol. 32, No. 7, 1986a, pp. 841-855.

T.L. Saaty. "Exploring Optimization Through Hierarchies and Ratio Scales," *Socio-Economic Planning Sciences*, Vol. 20, No. 6, 1986b, pp. 355-360.

T.L. Saaty. "A Note on the AHP and Expected Value Theory," *Socio-Economic Planning Science*, Vol. 20, No. 6, 1986c, pp. 397-398.

T.L. Saaty, and Takizawa. "Dependence and Independence: From Linear Hierarchies to Nonlinear Networks," *European Journal of Operational Research*, Vol. 26, 1986, pp. 229-237.

L.G. Vargas. "Utility Theory and Reciprocal Pairwise Comparisons: The Eigenvector Method," *Socio-Economic Planning Sciences*, Vol. 20, No. 6, 1986, pp. 387-391.

J. Aczel, and C. Alsina. "Synthesizing Judgments: A Functional Equations Approach," *Mathematical Modelling*, Vol. 9, Nos. 3-5, 1987, pp. 311-320.

E.J. Barbeau. "Reciprocal Matrices of Order 4," *Mathematical Modelling*, Vol. 9, Nos. 3-5, pp. 321-325, 1987.

P.T. Harker. "Derivatives of the Perron Root of a Positive Reciprocal Matrix with Application to the Analytic Hierarchy Process," *Applied Mathematics and Computation*, Vol 22, 1987, pp. 217-232.

P. Harker, and L. Vargas. "The Theory of Ratio Scale Estimation: Saaty's Analytic Hierarchy," *Management Science*, Vol. 33, No. 11, 1987, pp. 1383-1403.

T.L. Saaty. "The Analytic Hierarchy Process - What It is and How It is Used," *Mathematical Modeling*, Vol. 3, No. 5, 1987a, pp. 161-176.

T.L. Saaty. "How to Handle Dependence with the Analytic Hierarchy Process," *Mathematical Modelling*, Vol. 9, Nos. 3-5, 1987b, pp. 369-376.

L.G. Vargas. "Priority Theory and Utility Theory," *Mathematical Modelling*, Vol. 9 (3-5), 1987, pp. 381-385.

F. Zahedi. "A Utility Approach to the Analytic Hierarchy Process," *Mathematical Modelling*, Vol. 9 (3-5), 1987, pp. 387-395.

F. Zahedi. "A Note on Input Consistency in the Application of AHP," *Decision Sciences*. 1988, pp. 708-710.

B.L. Golden, E.A. Wasil, and D.E. Levy. "Applications of the Analytic Hierarchy Process: A Categorized, Annotated Bibliography," in B.L. Golden, E.A. Wasil, and P.T. Harker (eds), "The Analytic Hierarchy Process, Applications and Studies," Springer Verlag: New York, 1989.

J. Dyer. "Why the AHP is Like the MAUT!" presented at the 28th Joint National Meeting of ORSA/TIMS, New York City, 1989.

B.L. Golden, Q. Wang. "An Alternate Measure of Consistency," in B.L. Golden, E.A. Wasil, P.T. Harker (Eds.) "The Analytic Hierarchy Process, Applications and Studies," Berlin: Springer-Verlag, 1989.

M. McCurdy. *Two Enhancements of the Logarithmic Least-squares Method For Analyzing Subjective Comparisons*, Headquarters of the Commander in Chief, U.S. Pacific Command, Strategic Planning and Policy Directorate, Research and Analysis Division Technical Memorandum, 1989.

T.L. Saaty, J. Alexander. "Conflict Resolution: The Analytic Hierarchy Process," New York: Praeger, 1989.

B. Schoner and W.C. Wedley. "Ambiguous Criteria Weights in AHP - Consequences and Solutions," *Decision Sciences*, Summer 1989, pp. 462-475.

L.G. Vargas. "Why the AHP is Not Like the MAUT!" presented at the 28th Joint National Meeting of ORSA/TIMS, New York City, 1989.

J. Dyer. "Remarks on the Analytic Hierarchy Process," *Management Science* Vol. 36, No.3, 1990a, pp. 249-258.

J. Dyer. "A Clarification of 'Remarks on the Analytic Hierarchy Process'," *Management Science*. Vol. 36, No.3, 1990b, pp. 274-275.

P.T. Harker and L.G. Vargas. "Reply to 'Remarks on the Analytic Hierarchy Process' by J.S. Dyer," *Management Science*. Vol. 36, No.3, 1990, pp. 269-273.

T.A. Saaty. "An Exposition of the AHP in Reply to the Paper 'Remarks on the Analytic Hierarchy Process'," *Management Science*, Vol. 36, No.3, 1990, pp. 259-268.

B. CRITICISMS, CAVEATS, AND REPLIES

Critical discussion of the AHP has centered on five major areas, the phenomenon called rank reversal, methods for eliciting judgments, methods for estimating the priorities from the comparative judgments, the assumption that the pairwise comparative judgments reflect a ratio scale of measurement, and the claim that the AHP represents a special case of MAV/MAU theory.

1. Rank Reversal

Rank reversal is the phenomenon wherein adding a new alternative to a set of options or deleting an existing alternative changes the preference ordering among the original set. For the purposes of this discussion, we can divide occurrences of rank reversal into two types, *behavioral* and *methodological*. Behavioral rank reversal occurs when adding or deleting an alternative brings new information to the decision problem; as a result, the decisionmaker may reconsider the original decision problem and reorder the original preferences. Methodological rank reversal occurs when the procedure for estimating the rank ordering of the alternatives itself causes rank reversals among the original options. In this paper we are interested only in methodological rank reversal.

Researchers view behavioral rank reversal to be a genuine characteristic of human behavior. However, many researchers devising methods for identifying preferences consider methodological rank reversal to be an undesirable property of a decision analytic method.[3]

We refer to an example devised by Belton and Gear (1984) to illustrate how the AHP allows rank reversal to occur. Belton and Gear developed the following judgment matrices A_K *for*

---

[3] Bunn (1984) describes a related principal, "independence from irrelevant alternatives" as follows: "Informally stated, this [independence of irrelevant alternatives] implies that we should require of a sensible decision criterion that its ranking of two alternative actions shall not depend upon a third option, which is never preferable to both, is considered." pp. 21-22. The more general criterion for assessing decision analytic methods is *coherence*. For a brief discussion, see Bunn (1984) and Lindley (1985).

three alternatives with regard to three criteria as follows. Each matrix A_K contains the pairwise comparisons among the three alternatives relative to the Kth criterion. Thus, the entry "1/9" in the first row and second column of matrix A_1 means that alternative 1 has 1/9th the priority (e.g., importance, weight) of alternative 2 with respect to criterion 1. However, matrix A_3 shows alternative 1 to have 8/9ths of the priority of alternative 2 with respect to criterion 3.

A_1

1	1/9	1
9	1	9
1	1/9	1

A_2

1	9	9
1/9	1	1
1/9	1	1

A_3

1	8/9	9
9/8	1	9
1/8	1/9	1

The principal eigenvectors of the respective matrices are as follows:[4]

1/11	9/11	8/18
9/11	1/11	9/18
1/11	1/11	1/18

Assuming equal priorities for the three criteria for the sake of computational convenience, the priorities of the three alternatives are 0.45 (1/3x(1/11+9/11+8/18)), 0.47, and 0.08. Thus, the second option is preferred to the first option.

Belton and Gear then add the comparative judgments for a fourth alternative to the judgment matrices with the following results. Note that all entries but those for the fourth rows and fourth columns are unchanged from those above. Thus, the preferences stated among the three original alternatives remain unchanged.

A_1

1	1/9	1	1/9
9	1	9	1
1	1/9	1	1/9
9	1	9	1

A_2

1	9	9	9
1/9	1	1	1
1/9	1	1	1
1/9	1	1	1

A_3

1	8/9	9	8/9
9/8	1	9	1
1/8	1/9	1	1/9
1	8/9	9	1

The principal eigenvectors of the respective matrices are as follows:

1/20	9/12	8/27
9/20	1/12	9/27
1/20	1/12	1/27
9/20	1/12	9/27

[4] We find the principal eigenvector of a consistent matrix by summing the matrix's column vectors and dividing this vector's elements by the grand sum of all of the matrix's elements. In matrix notation this is $(1'A1)^{-1}(A1)$.

We again assume equal priorities for the three criteria for the sake of computational convenience. The priorities on the four options are about 0.37, 0.29, 0.06, and 0.29. Thus, the first option is now preferred to the second option, the order among the first two alternatives having changed solely as a result of the AHP's priority estimation procedure and the addition of an additional alternative.

Careful consideration of the examples shows the immediate reason for rank reversal in this example. The relative priorities among the alternatives do not change when a new option is added. Regardless of the number of alternatives, the second option dominates the first option in a 9-to-1 ratio with respect to the first criterion and in a 9-to-8 ratio with respect to the third criterion. However, the denominators of the priorities do change when alternatives are added.

Where the second alternative initially gets 9/11 of the weight allocated to the first criterion, it only gets 9/20 of the weight in the four alternative case, which constitutes a 45 percent loss. With respect to the third criterion, the proportion of the weight received by the second alternative drops from 9/18 to 9/27, which constitutes a 33 percent loss. With respect to the second criterion, where the first alternative dominates the second by a 9-to-1 ratio, the proportion of the weight declines only from 9/11 to 9/12, which constitutes only an 8 percent loss. Thus, the first alternative suffers less of a penalty overall when a new option was added and so became the preferred option.

Dyer and Wendell (1985) developed a similar example based on the following scores for four alternatives on four criteria:

Table III-1. Scores for Four Alternatives on Four Criteria

Criteria	Alternatives			
	1	2	3	4
A	[illegible]	9	8	4
B	9	1	1	1
C	1	9	4	8
D	3	1	5	5

Pairwise comparison matrices with respect to each criterion are formed by taking pairwise ratios of the scores among the alternatives (e.g., for criterion A, $a_{12}=1/9$, $a_{13}=1/8$, $a_{24}=9/4$). Omitting the fourth alternative respectively results in the ranking 3, 2, 1. However, including the fourth alternative results in the ranking 1, 3 = 4, 2.

Dyer and Wendell then formulated an alternative decision problem for his example data which implied a rank ordering of 2, 3=4, 1. Consider that the four alternatives are mutually exclusive investment opportunities, each yielding returns for four years and each requiring the same capital outlay. The four criteria respectively constitute the returns in years 1 through 4. The above table therefore gives the dollar return for each alternative in each year and the four criteria should be equally important. In order to simplify the problem, Dyer and Wendell assume a zero discount rate. The overall value of each alternative is clearly proportional to its total score over the four criteria, and the second alternative is the best with a score of 20. However, the AHP never selects this alternative. Dyer and Wendell then formulated several variations of the standard AHP procedure which yield the answer consistent with intuition.

Harker and Vargas (1987) responded to the Dyer and Wendell counterexample with the argument that the criteria of the example should not be considered to be equally important because the returns vary for each year. As a result, the relative importance of the years are dependent on the alternatives, and we thus require the supermatrix methodology (Section A.3.C on the independence assumption) to estimate the priorities of the alternatives. Employing this approach yields the correct answer.

Saaty (1987) recognized this asymmetric change in the denominators and defined a new term for the AHP methodology, *structural criterion*, to subsume it within AHP theory. Structural criteria represent an alternative interpretation of the normalization performed to force the priorities of a set of elements to sum to 1. Adding options to a decision problem inevitably enlarges the denominator underlying the priorities of the alternatives. However, as we have seen in the Belton and Gear example, the degree to which the denominator gets larger depends on the degree to which new alternatives dominate the original alternatives.

Belton and Gear suggested an alternative to eigenvector prioritization which precludes the rank reversal problem. Saaty and Vargas (1984) refuted this method with a counterexample demonstrating that Belton and Gear's method is itself subject to rank reversal. However, Belton and Gear (1985) in turn suggested that Saaty and Vargas misinterpreted their method and thereby misapplied it. Vargas (1985) subsequently introduced a second counterexample demonstrating that Belton and Gear's method produced counterintuitive intermediate results in the AHP. However, Belton and Gear's method was not geared to producing intuitive intermediate results, but rather reversal-resistant final results. We thus do not consider Vargas' rejoinder to have been an effective counterargument.

We defer discussion of Belton and Gear's method to the discussion of Schoner and Wedley's (1989) work below. Schoner and Wedley discuss two corrective measures to avoid rank reversal.

Saaty and Vargas (1984a) and Saaty (1990) have argued that rank reversal occurs naturally in human decisionmaking and that...

> "It is important to point out that rank reversal can be a good thing. That is how a new and important attribute can alter previous preferences.
>
> "However, the AHP makes it clear that rank reversal does occur and should be acceptable. In life, people often learn new things which may cause them to reverse previous preferences, and this can lead to rank reversal even under consistency conditions." (Saaty and Vargas, 1984a, pp. 515)

However, Saaty's argument concerns what we have termed behavioral rank reversal and not methodological rank reversal. For the purpose of this review, methodological rank reversal is more serious because it means that preference orderings change even when the preference structure regarding the original problem has not changed. Further, the AHP has never been tested rigorously as a descriptive model of human preference. Thus, even if the AHP allows for rank reversal, there is no evidence that human decisionmakers also change ranks or would want to change ranks in the way that occurs when using the AHP.

A problem related to rank reversal in the AHP is that adding alternatives to the original set of options may change the priorities of the original set of alternatives relative to each other while not yielding discrete changes in rank order. Forman's (1987) example concerning the relative worth of basketball players illustrates a continuous shift in priorities without a change in ranks. Among the original three players in the analysis, the player with strong offensive skills is preferred to the rounded player. Adding an additional player with strong offensive skills preserves this ordering but reduces the magnitude of the difference in preferences. It is only after adding a total of four players with strong offensive skills that the rounded player becomes preferred to the offensive player from the original set of three players.

Saaty and Vargas (1984a) and Saaty (1987a) have taken a more careful look at rank reversal in the core AHP methodology. As a result of their inquiry, they identified conditions in which adding or deleting an alternative in a consistent matrix results in rank reversal. However, these conditions do not hold for inconsistent matrices, which is the expected case in most applications.

Saaty (1986c, 1987a) has shown that the absolute measurement variant of the AHP (see Section C.2 below) is not subject to rank reversal. However, he has pointed out that it is not always appropriate to use this methodology.

A concept that emerged out of the discussion on rank reversal in the AHP is that of the alternative which is a *near copy* of another alternative. Saaty (1987a) defines the term more rigorously. However, for the sake of this discussion, the usual understanding is sufficient.

Consider an example adapted from Saaty (1987a, pp. 173; also see Forman's example above), and represented in Table III-2 below. There are two alternatives, A and B_1, compared with respect to two criteria, C_1 and C_2. Saaty weights the criteria 2/3 and 1/3, respectively. With respect to C_1, the two alternatives respectively have priorities of 1/3 and 2/3. With respect to C_2, the two alternatives respectively have priorities 3/4 and 1/4. Thus, neither alternative dominates the other on both criteria. Overall, A has a priority of 0.47 $(\frac{2}{3} \times 1/3)+(\frac{1}{3} \times 3/4)$ and B_1 has a priority of 0.53.

Table III-2. Priorities for Two Alternatives with Respect to Two Criteria

		Criteria (priority)	
		C_1 (2/3)	C_2 (1/3)
Alternatives	**A**	1/3	3/4
	B_1	2/3	1/4

Now introduce an alternative B_2 identical to B_1. Under criterion C_1, A_1 has the priority 1/5 and B_1 and B_2 each have priority 2/5. Under criterion C_2, A has priority 3/5 and B_1 and B_2 each have priority 1/5. The revised priorities for the alternatives are now 1/3 for A, B_1, and B_2.

In general, if we introduce n-1 identical copies of B_1 to the set of original options, the priorities change in the direction of diluting the original advantage of B_1 with respect to the second criteria. In our example, (see Table III-3) the priority of B_1 with respect to C_1 is (2/(1+2n)) and the priority of B_1 with respect to C_2 is (1/(3+n)). Overall, the B_i will have a greater priority than A as long as 2/3x(1/(1+2n)) +1/3x((3/(3+n)) < 2/3x(2/(1+2n)) + 1/3x(1/(3+n)), or n<2.

Table III-3. Priorities for Two Alternatives with Respect to Two Criteria

		Criteria (priority)	
		C_1 (2/3)	C_2 (1/3)
Alternatives	A	1/(2n+1)	3/(n+3)
	B_i	2/(2n+1)	1/(n+3)

Forman (1987) illustrates the effect of identical copies on priorities with another example, but argues that under some circumstances, retaining the copies may provide useful information. However, note our exceptions to Forman's analysis in Section C.2 on the absolute measurement variant of the AHP.

Saaty (1987a) argues that near copies should be omitted from the analysis under many circumstances and offers a heuristic for identifying which of a set of options are near copies of each other. Harker and Vargas (1987) further argue that the presence of near copies violates the fourth axiom upon which the AHP is based, that the structure of alternatives and criteria representing the prioritization problem is complete, and thus excludes indistiguishable alternatives.

2. Implementation

The AHP requires judgments of the relative priority of criteria or alternatives in terms of higher order criteria. Watson and Freeling (1982), Belton and Gear (undated manuscript), and Dyer (1990a) have noted that while people

> "...appeared to be capable of interpreting this request and providing numerical responses, we maintain that this question is meaningless. In comparing the relative importance of distinct attributes we must ask *how much* of one attribute (in some specified units) is worth a *particular amount* of some other attribute (in some specified units). To ask this question without this specification ought to evoke the response *I cannot answer this question unless you specify the units of measurement.* (Watson and Freeling, 1982, pg. 282, 283)

Thus, instead of the more general question regarding relative importance or value, Watson and Freeling argue that the appropriate question asks for an assessment of tradeoff preferences in terms of specific units. They continue with the following example:

> "Let us illustrate with the example often used in this area, that of evaluating the relative merits of different cars. Two attributes might be comfort and reliability. The approach in the papers referred to above (Saaty, 1980) would be to ask the

> The approach in the papers referred to above (Saaty, 1980) would be to ask the question: 'Which of comfort and reliability is most important, and by how much?' We argue that this is a meaningless question. Instead we should ask: 'Consider the change in comfort between car A and car B (thus defining the unit of measurement for comfort). Compare this with the change in reliability between car C and car D (giving the unit of measurement for reliability). Which is more valuable, and how many increments of the less valuable change is equivalent to the more valuable change?" (Watson and Freeling, 1982, pg. 283)

Watson and Freeling thus argue that the conventional AHP question "Which is more important and by how much?" does not provide a standard for answering the question. They thus also implicitly raise the following questions. If people do not ask for more information in answering this question, then *exactly how are they answering the question?* Are they relying on some subjectively defined standard? Is knowing the subjectively defined standard necessary to interpreting the results? Must the results of the study be adjusted or discounted if the subjectively defined standard differs from that assumed by the analyst? (ie, Has the validity of the results been compromised? See section E.4 in Chapter 1 above.) Not understanding the "assumptions" underlying the responses may confuse the meaning of the results if the results depend on how the question is "interpreted". Yet asking only the general question allows respondents to vary in how they make comparisons.

Watson and Freeling, for instance, suggests that asserting that reliability is five times as important as comfort *could mean* "to have a car as reliable as the car of maximum current reliability is five times as valuable as having a car as comfortable as the car of maximum current comfort." (pp. 283). Saaty, Vargas, and Wendell (1983) argue that when alternatives are measured in common units, eg, dollars, then the mean score of the alternatives under each criterion should be compared. Saaty (1990) suggests two different procedures that respondents might be using use to compare criteria. In practice respondents may yet be doing some else altogether.

Consider the following data measured on a scale with a natural zero point (e.g., dollars).

Table III-4. Scores for Three Alternatives on Two Criteria

		Alternatives		
		1	**2**	**3**
Criteria	**A**	1.0	2.0	3.0
	B	0.5	5.0	3.0

The question "Is the first criterion more or less important than the second criterion, and by how much?" can legitimately produce the following answers: (1) less important with a ratio of 0.2; (2) less important with a ratio of .4; (3) less important with a ratio of 0.6; (4a) less important with a ratio of 0.7; (4b) less important with a ratio of 0.7; (5) equally important; (6) more important with a ratio of 2.

The numerical responses vary over a range of 10 to 1 and either criterion may be the more important one, depending on the basis for forming the response. The first answer is based on the ratio of the variance of the scores under each criteria. The second answer is the ratio between the maximum scores, perhaps reflecting a focus on the best result under each criterion. The third answer reflects a ratio between the best and worst alternative under each criterion (e.g., Dyer and Wendell, 1985, Dyer,1990a). Answer 4a is the ratio of the mean scores under each category and corresponds to the interpretation of the question specified by Saaty, Vargas and Wendell (1983) in their response to Watson and Freeling. Answer 4b is the ratio of the median scores under each category. The fifth answer is based on a perception that there is no reason for believing that one criterion is more important than the other. The sixth answer is the ratio of the minimum scores under each criterion, perhaps reflecting a concern for the worst result under each criterion.

Dyer and Wendell (1985) point out a similar example regarding investment profits of $15,000 and $20,000. One answer to the importance question might be 4/3 (20/15). However, if the respondent judges profits relative to a baseline of $10,000, the response should be 2 ((20-10)/(15-10)).

However, no particular answer or interpretation is "correct" *a priori* because neither AHP theory (eg, Saaty 1980, 1986c; Harker and Vargas, 1987; however Saaty, Vargas, and Wendell, 1983 may be an exception.) nor the eliciting question specify a basis for interpreting the question or forming a response. The analyst, however, may prefer one interpretation over another because it corresponds to a particular study requirement. Further, one person may give different answers at different times, or two individuals may give different responses, even though the beliefs underlying the responses are identical because they make the comparisons in different ways. In addition, it may be difficult to interpret the overall priorities resulting from the AHP when we do not know 1) on what basis the responses are formed; 2) if the basis for response is consistent between responses; or 3)if the basis for response is consistent between respondents in a multiple-respondent analysis.

Dyer and Wendell (1985, also Dyer 1990) argue for a more general problem with the standard AHP question-response format. Ratio comparisons and ratio measurement scales imply the existence of a fixed endpoint with a zero value. Ratio comparisons of weight are possible because weight is measured on a ratio scale, and thus has a zero endpoint, no weight. The Celsius and Fahrenheit temperature scales have lowest temperatures that are negative and thus are not ratio scales. Ratios of temperatures measured on these scales are thus not meaningful. Some concepts may not be easily thought of in terms of a scale with a zero-valued end point. One such concept is "warmth". What level of "warmth" serves as a the lowest value of warmth on a ratio scale of warmth? Belton (1986) raises the same issue when she points out "What does it mean to say that one course of action would contribute twice as much to [say] cultural advancement [as described by Saaty and Rogers, 1976]"? What level of "cultural advancement" serves as the zero-valued end point required for a ratio-scaled comparison of alternatives with respect to the criterion "contribution to cultural advancement"? Does the interpretation of the AHP results depend on knowing what "level" of cultural advancement corresponds to the zero point?

There are two ways of avoiding the problems potentially associated with question ambiguity. The first is to diagnose the basis that respondents use for forming a response. This is a modeling problem that may be difficult and certainly is extraneous to AHP. Furthermore, if respondents use a less preferred basis for forming a response, or are inconsistent, we may be left with discarding the data as unusable. The second method is to instruct the respondent as to the basis for forming a response. This instruction may consist of training sessions, but also should consist of incorporating explicit directions within AHP questions. This, of course, does not guarantee that respondents will follow the directions.

Saaty, Vargas and Wendell (1983) and Harker and Vargas (1987) both have responded to the criticism of ambiguity. Saaty et. al. suggested the following question for comparing criteria in the context of a car comparison example -- "What is the ratio of the average (or total) contribution to cost of attribute *i* to the average (or total) contribution to cost of attribute *j*." Suppose three cars cost \$6,000, \$8,000, and \$10,000 to purchase and \$1,800, \$1,200, and \$600 dollars to maintain. The average purchase cost is \$8,000, and the average maintenance cost is \$1,200. Thus, one valuation of the relative priority of purchase price to maintenance cost is 8000/1200, or 6-2/3.

Harker and Vargas (1987) have argued that

> "The problem of ambiguity is not a flaw of the AHP, but in fact arises out of a fundamental question concerning the frame of reference in which one makes the necessary judgments. The meaning (or lack of meaning) of a question ultimately depends upon the cognitive environment in which one exists. One's beliefs as to the meaning of terms such as "more important" or "more strongly important" is a

function of the cognitive frame of reference in which one currently resides. These definitions will vary from day to day and from individual to individual. While it is true that a poorly worded question yields poor results and that better wording of a question can significantly increase the effectiveness of the methods, no method or no perfect question will ever remove ambiguity completely due to the reliance on the individual's frame of reference. Watson and Freeling (1983) and others have criticized the mode of questioning outlined in Saaty's theory while not fully comprehending the above-mentioned issue or understanding that excessive ambiguity not explicable within the context of the frame of reference is not a failure of the method being used, but rather a failure of the analyst or decisionmaker to fully comprehend the issue at hand and state questions which meaningfully address it. (pp. 1387)

"In assessing classical utility functions numerous experiments by Tversky and others (Tversky and Kahneman 1981; Hershey et al, 1982; Hershey and Shoemaker 1983; Shoemaker and Wait 1982; Krantz et al, 1971) have shown conclusively that one's frame of reference matters. For example, whether one is asked to adjust probabilities (probability equivalence) or the sure amount (certainty equivalence) in eliciting a von Neumann-Morgenstern utility function, although they are theoretically equivalent methods, they lead to different utility measurements (McCord and de Neufville 1984). Thus, the dependence of utility measurements on one's frame of reference is a well-established phenomenon.

"Clear definitions for the criteria, subcriteria and alternatives is essential in all decision aids and should obviously be of major concern to users of AHP or any other methodology." (pp. 1388)

Harker and Vargas appear to be arguing against a criticism that the AHP is flawed because of a dependence on a subjective frame of reference. They argue that such flaws are not inherent to the method, but are naturally part of human b[illegible]vior and may be introduced unnecessarily as a result of flawed practice by AHP practitioners. They do not appear to take issue with Watson and Freeling's argument, or the argument that we state above that the conventional AHP question allows considerable leeway for respondents to subjectively determine what the question means, what data are relevant to answering the question, and how the data should be used to answer the question. Nor do they address the argument that ensuring the interpretability of AHP results may depend on either controlling or knowing the answers to these questions. However, Harker and Vargas (1987) are not against well-formulated AHP questions.

"a poorly worded question yields poor results and that better wording of a question can significantly increase the effectiveness of the methods;

"excessive ambiguity not explicable within the context of the frame of reference is not a failure of the method being used, but rather a failure of the analyst or decision maker to fully comprehend the issue at hand and state questions which meaningfully address it." (pg. 1387)

On a more practical level, Harper and Vargas also argue that the AHP methodology avoids the problem of explicitly defining a fixed zero point for the comparisons. Instead, the AHD treats

the dominated (eg, less preferred, smaller) alternative as the reference point in each pairwise comparison. It is hypothesized that the respondent "divides" the psychological intensity associated with the dominated alternative into that associated with the superior alternative to determine a ratio response (e.g., Saaty, personal communication). However, this idea about how people respond to "ratio comparison instructions" has yet to be adequately tested. Some evidence (Section B.3 below) supports a continuing view about responses to ratio comparison instructions.

In addition, Harker and Vargas' hypothesis requires a scale of subjective measurement that admits of ratio in the first place comparisons, i.e., the "intensities" mentally compared must be ratio-scaled measurements. However, Harker and Vargas do not provide evidence that this is true, either generally or in particular cases.

Finally, Schoner and Wedley (1989), and Dyer (1990a) have also discussed suggested specific question formats to resolve the difficulties raised by Watson and Freeling.

Dyer and Wendell (1985), Belton (1986), and Dyer (1990a) have raised additional questions with regard to AHP implementation. These questions concern the relationship, specified by Saaty (1980), between verbal categorical responses and their equivalent numerical ratios. Saaty (1980) presents empirical evidence that the ratios 1 to 9 provide a reasonable numerical equivalent to the verbal categorical scale. However, the authors argue that there is no necessary basis for this correspondence. Belton (1986) further argues that the significance of the verbal responses as ratio judgments is frequently not explained to respondents, and so may not provide a good basis for a ratio-scaled measurement of priority. She also argues that the evidence supporting the 1-to-9 ratio scale as the best numeric representation of the verbal scale is weak at best and inconclusive at worst. Finally, Belton suggests that individual differences exist in how people interpret the verbal rating scale, thus weakening their usefulness.

The authors also point out that the equivalence between the categorical response "Weak importance of one over another....Experience and judgment slightly favor one activity over another" and a ratio of three is not intuitive. Nor is it intuitive that if *A* is weakly more important than *B* and *B* is weakly more important than *C*, that *C* is absolutely more important than *A* (3x3=9). (We elaborate on this issue in Section D.) Overall, the authors are equivocal about the use of the verbal-categorical scale, with Belton suggesting that if the relationship between the verbal and numeric scales are explained (and they should be), the verbal responses have no additional value over numeric responses.

3. Are AHP Results Ratio-Scaled?

Veit, Callero and Rose (1984), Mellers, Davis and Birnbaum (1984), and Meyer and Booker, 1990) have questioned the claim that the AHP yields ratio-scaled results (eg, an estimated priority of 4 is subjectively valued about twice as much as an estimated priority of 2).

Meyer and Booker (personal communication) argue that responses made on the verbal/categorical scale may not be interpretable as being ratio-scaled, thus precluding a similar interpretation for the overall results. (Recall Belton's (1986) conjecture that the relationship between the verbal/categorical scale and its numeric equivalent is frequently not explained.)

Veit et al argue that the assumption that respondents are actually making ratio responses is empirically untested. Veit (personal communication) has evaluated the predictions of an underlying ratio comparison model with several case example results that Saaty has reported in the literature. Preliminary evaluation does not support Saaty's contention that respondents are making ratio-type comparisons.

This result may sound counterintuitive, especially if respondents are instructed to make relative magnitude (ratio) comparisons (However, Belton (1986) suggests that they frequently are not so instructed.) However, accumulating evidence (e.g., Mellers, Davis, and Birnbaum, 1984; Birnbaum, 1981) suggests that the covert, mental process of making ratio comparisons may frequently involve mentally taking differences on transformed (e.g., logarithmic) subjective scale values.

Furthermore, Veit et al argue that

> "The scale values derived from a 'correct' ratio model are unique only to a power transformation (Krantz and Tversky, 1971). [Thus, even if] the data supported a ratio model, the scale values derived from the ratio model would only be a power transformation of the 'true' values under the model; they would *not* be ratio scales as Saaty suggests. Therefore, it would be inappropriate to make ratio comparisons of the 'weights' (scales derived from the 'ratio' model) within hierarchical levels." (Veit, Callero, and Rose, 1984, pg. 51).

The interested reader is referred to the review of the subjective transfer function method, Chapter IV, for more information and references on these issues.

4. The AHP and Utility Theory: Dyer's Critique

Dyer (1985, 1990a, 1990b) has raised several criticisms regarding the AHP. We have summarized several of them above in the discussions of rank reversal and AHP implementation. However, Dyer's core argument is that the results of an AHP analysis are "arbitrary" because they "...are not governed by a basic principle of rationality [independence from rank reversal]" (pg. 3)

However, Dyer's core argument is that the results of an AHP analysis are "arbitrary" because they "...are not governed by a basic principle of rationality [independence from rank reversal]" (pg. 3)

> "Like any strong statement, this one deserves a caveat or two. Under certain restrictive assumptions it is possible to provide answers to the questions posed by this process in such a way that the results will be consistent with the decision maker's true preferences. However, it is unlikely that a user would be able to respond in a manner consistent with these assumptions. (Dyer and Wendell, 1985, pp. 1-2)
>
> "Rank reversal in the AHP is in fact a symptom that a wrong alternative may be chosen by the process. (Dyer and Wendell, 1985, pg. 19)
>
> "It is well known that the additive model is extremely robust (e.g., Dawes, 1974) and, therefore, the results of the procedure may produce 'reasonable' rankings of alternatives even though parameters of the model are seriously in error." (Dyer and Wendell, 1985, pp. 2)

Dyer's underlying argument is that the AHP produces "arbitrary" rankings the assumptions under which hierarchic composition is a valid operation are generally not satisfied.[5] By "arbitrary" Dyer means that the ranking of alternatives depends on the number of alternatives being considered and their degree of similarity to each other in addition to the preferences among the alternatives and criteria. Thus, rankings and even the estimated relative priorities, are a function of factors that Dyer considers to be irrelevant to the problems of rating and rank ordering the alternatives. Saaty (1990) specifically takes issue with this view, arguing that this kind of criterion is an extension from MAU and MAV theory which does not necessarily apply to the AHP. Rather, the AHP captures aspects of preference modelling which are different from those subsumed under MAU/MAV methods. Thus, the AHP should not be judged on the basis of this criterion. For instance, Saaty (1990) and Saaty and Vargas (1984) have argued that rank reversal and more generally, the "dilution" of priorities when "near-copies" exist in the alternatives set, correspond to some human preference behavior. However, if the instances of rank reversal in question occur under only circumscribed conditions in human preferences, then it may not be desirable for them to occur "indiscriminately" in the AHP.

Dyer (1990a) extends his "arbitrary ranking" criticism to the absolute measurement variant of the AHP (section C.2 below), arguing that although rank reversal does not occur when the absolute measurement technique is used, the assumptions allowing for hierarchic composition are still violated.

---

[5] Schoner and Wedley (1989) also have made this argument recently. We summarize their recent work below in Section C, Extensions.

BIBLIOGRAPHY

T.L. Saaty. *What Do Rank Preservation and Reversal Mean In the Analytic Hierarchy Process?* Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA, (undated).

V. Belton, T. Gear. "Assessing Weights by Means of Pairwise Comparisons," (undated).

T.L. Saaty and P.C. Rogers. "Higher Education in the United States (1985-2000): Scenario Construction Using A Hierarchical Framework With Eigenvector Weighting," *Socio-Economic Planning Sciences*, Vol. 10, 1976, pp. 251-63.

C.R. Johnson. "Constructive Critique of A Hierarchical Prioritization Scheme Employing Pair Comparisons," *IEEE Transactions*, 1980, pp. 373-378.

M.H. Birnbaum. "Controversies in Psychological Measurement," in B. Wegener (ed.), "Social Attitudes and Psychological Measurement," Hillsdale, NJ: Erlbaum, 1981.

T.L. Saaty. "The Analytic Hierarchy Process." New York: McGraw-Hill, Inc., 1980.

R. Kamenetzky. "The Relationship Between the Analytic Hierarchy Process and the Additive Value Function," *Decision Sciences*, Vol. 13, 1982, pp. 702-713.

S.R. Watson and A.N.S. Freeling. "Assessing Attribute Weights," *Omega* Vol. 10, No. 6, 1982, pp. 582-583.

T.L. Saaty, L. Vargas, R. Wendell. "Assessing Attribute Weights By Ratios," *Omega* 11, No. 1 1983, p. 9.

S.R. Watson and A.N.S. Freeling. "Comment On: Assessing Attribute Weight By Ratios," *Omega*, Vol. 11, No. 1, 1983, pg. 13.

V. Belton and T. Gear. "On a Short-coming of Saaty's Method of Analytic Hierarchies," *Omega*, Vol. 11, No. 3, 1984, pp. 228-230.

D. Bunn. "Applied Decision Analysis," New York: McGraw Hill Book Company, 1984.

B.A. Mellers, D. M. Davis, and M.H. Birnbaum. "Weight of Evidence Supports One Operation for 'Ratios' and 'Differences' of Heaviness," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 10, No. 2, 1984, pp. 216-230.

T.L. Saaty and L. Vargas, "The Legitimacy of Rank Reversal," *Omega*, Vol. 12, No. 5, 1984a, pp. 513-516.

T.L. Saaty and L. Vargas. "Comparison of Eigenvalue, Logarithmic Least Squares and Least Squares Estimating Methods in Estimating Ratios," *Mathematical Modelling*, Vol. 6, 1984b, pp. 309-324.

T.L. Saaty and L. Vargas. "Inconsistency and Rank Preservation," *Journal of Mathematical Psychology*, Vol. 28, 1984c, pp. 205-214.

C T. Veit, M. Callero, and B.J. Rose. *Introduction to the Subjective Transfer Function Approach to Analyzing Systems*. Rand R-3021-AF, July 1984.

V. Belton and T. Gear. "The Legitimacy of Rank Reversal: A Comment," *Omega*, Vol. 13, No. 3, 1985, pp. 143-144.

J. Dyer, R.E. Wendell. *A Critique of the Analytic Hierarchy Method*. University of Texas Graduate School of Business Working Paper 84/85-4-24, 1985.

D.V. Lindley. "Making Decisions," New York: John Wiley and Sons, Inc., 1985.

L. G. Vargas, "A Rejoinder," *Omega*, Vol. 13, No. 4, 1985, p. 249.

E. Forman. "In Search of the Right Model," *Telematics and Informatics*. Vol. 3, No. 4, 1986, pp. 229-235.

T.L. Saaty. "Absolute and Relative Measurement With the AHP. The Most Livable Cities In the United States," *Socio-Economic Planning Sciences*. Vol. 20, No. 6, 1986a, pp. 327-331.

T.L. Saaty. "A Note on the AHP and Expected Value Theory," *Socio-Economic Planning Sciences*, Vol. 20, No. 6, 1986b, pp. 397-398.

T.L. Saaty. "Axiomatic Foundation of the Analytic Hierarchy Process," *Management Science*, Vol. 32, No. 7, 1986c, pp. 841-855.

L.G. Vargas. "Utility Theory and Reciprocal Pairwise Comparisons: The Eigenvector Method," *Socio-Economic Planning Sciences*, Vol. 20, No. 6, 1986, pp. 387-391.

F. Zahedi. "A Simulation Study of Estimation Methods in the Analytic Hierarchy Process," *Socio-Economic Planning Sciences*, Vol. 20, No. 6, pp. 387-391, 1986, pp. 347-354.

E.H. Forman. "Relative vs. Absolute Worth," *Mathematical Modelling*, Vol. 9, Nos. 3-5, 1987, pp. 195-202.

P. Harker and L. Vargas. "The Theory of Ratio Scale Estimation: Saaty's Analytic Hierarchy," *Management Science*, Vol. 33, No. 11, 1987, pp. 1383-1403.

T.L. Saaty. "Rank Generation, Preservation, and Reversal in the Analytic Hierarchy Decision Process," *Decision Sciences*, Vol. 18, 1987a, pp. 157-177.

T.L. Saaty. *A Note on Decisionmaking and Number Crunching, Is Normalization the Answer?* Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA, (undated, appears in the appendix of the T.L. Saaty. "The Analytic Hierarchy Process", 1988 edition).

L.G. Vargas. "Priority Theory and Utility Theory," *Mathematical Modelling*, Vol. 9, Nos. 3-5, 1987, pp. 381-385.

J. Dyer. "Why the AHP is Like the MAUT!" presented at the 28th Joint National Meeting of ORSA/TIMS, New York City, 1989.

B. Schoner and W.C. Wedley. "Ambiguous Criteria Weights in AHP - Consequences and Solutions," *Decision Sciences*, Summer 1989, pp. 462-475.

L.G. Vargas. "Why the AHP is Not Like the MAUT!" presented at the 28th Joint National Meeting of ORSA/TIMS, New York City, 1989.

J. Dyer. "Remarks on the Analytic Hierarchy Process," *Management Science* Vol. 36, No.3, 1990a, pp. 249-258.

J. Dyer. "A Clarification of 'Remarks on the Analytic Hierarchy Process'," *Management Science*, Vol. 36, No.3, 1990b, pp. 274-275.

M.M. Meyer, J.M. Booker. *Eliciting and Analyzing Expert Judgment*, Los Alamos National Laboratory, LA-11667-MS (NUREG/CR-5424)

P.T. Harker and L.G. Vargas. "Reply to 'Remarks on the Analytic Hierarchy Process' by J.S. Dyer," *Management Science*. Vol. 36, No.3, 1990, pp. 269-273.

T.A. Saaty. "An Exposition of the AHP in Reply to the Paper 'Remarks on the Analytic Hierarchy Process'," *Management Science*, Vol. 36, No.3, 1990, pp. 259-268.

C. AHP EXTENSIONS

1. Schoner and Wedley's Extension of the AHP

Schoner and Wedley (1989) have formulated an analysis that puts the issues of rank reversal and judgmental ambiguity into a common framework. Similar issues appear in Dyer and Wendell (1985), Belton and Gear (undated), and Dyer (1990a, 1990b). However, because Schoner and Wedley's work appears to be more inclusive, we have chosen to summarize it here separately. Nevertheless, additional extensions of the AHP similar to those proposed by Schoner and Wedley are discussed in Dyer and Wendell (1985) and Dyer (1990a, 1990b).

In general, Schoner and Wedley argue that "there is a necessary correspondence" between the kind of information elicited in pairwise judgments and how the priorities are computed (eg, how weights are normalized; Belton and Gear (undated) make an identical argument). They argue that not matching the method for computing priorities with the form of the pairwise judgment

> "...results in the generation of incorrect weights for the options under consideration regardless of whether or not new options are added or deleted. A rank reversal upon the addition of an option is merely symptomatic of this fact, and such reversals are shown not to occur when the correspondence condition is met." (pg. 1)

They further argue, similarly to Dyer (Dyer and Wendell, 1985; Dyer, 1990), that the priorities of criteria are <u>never</u> independent of the alternatives dependent on them in conventional AHP. Thus, some kind of remediation to Saaty's formulation of the AHP is always necessary.

Schoner and Wedley frame their arguments around a car selection example in which the three criteria are purchase price, annual maintenance cost, and fuel consumption. The authors use illustrative data rather than judgmental data in order to provide a baseline for comparing alternative methods. The data for their example follow:

Table III-4. Schoner and Wedley's Car Purchase Decision Example

	Purchase Price ($)	Maintenance ($/Year)	Fuel (Gals/Mile)
Car 1	14,000	2,000	0.05
Car 2	5,000	4,000	0.03
Car 3	6,000	4,000	0.05

Ownership period: 5 years
Annual mileage: 10,000 miles
Gasoline cost: $1.50 per gallon

Assuming a specific ownership period allows us to scale data for annual maintenance costs to data for lifetime maintenance costs. This procedure puts maintenance costs on a common basis with purchase price, which is also a lifetime cost. Similarly, assuming a gasoline cost and an annual mileage figure provides for scaling fuel consumption data to lifetime fuel costs. Lifetime fuel costs equal the product of fuel efficiency and 75,000 (5 years of ownership x 10,000 miles per year x 1.5 dollars per gallon). The numbers 1, 5, and 75,000 are termed *scale factors* because they convert the data to a common unit, which in this example is lifetime dollar cost.

Let T_{ik} be data on car i relative to criteria k (e.g., T_{11}=14,000, the purchase price for car 1) and let q_k be the scale factor for the kth criterion (q_1=1, q_2=5, q_3=75,000). Without the AHP, we could compare the three cars by comparing their total lifetime cost. The total lifetime cost of each car equals the sum of purchase price, lifetime maintenance cost, and lifetime fuel cost, or

$$\sum_k q_k T_{ik}.$$

We find the relative standing of one car to another by taking ratios of their lifetime costs. Let x_k be the priority of the kth criterion.

Schoner and Wedley show that x_k is proportional to the product $q_k \sum_i T_{ik}$. That is,

> "The relative importance of a criterion must be proportional to the product of its scaling factor and the sum (or average) of the absolute values of the option measurements on that criterion." (pg.5)

In the example, this is nothing more than saying that the relative importance of the criteria should be determined on the basis of their contribution (summed over alternatives) to total lifetime cost. For purchase price, this sum is \$25,000 (14,000+5,000+6,000). For maintenance cost, this is \$50,000 (5x(2,000+4,000+4,000)). For fuel consumption, this is \$9,750. Normalizing these values by dividing each by the sum of the three yields criterion priorities of 0.295, 0.590, and 0.115, respectively. Schoner and Wedley claim that their methodology is functionally equivalent to the supermatrix approach and argue that it is easier to implement.

Saaty and Vargas (1983) made a functionally identical recommendation in response to Belton and Gear's (1983) rank reversal example. In their paper they stated that priorities amongst the criteria should be determined by the average contribution to cost of the attributes. Thus, the judgment comparing the relative importance of purchase price and maintenance cost would follow from the question "*For the given set of cars, consider the average contribution to lifetime cost of purchase price* ((14,000+5,000+6,000)/3) *and the average contribution to lifetime cost of maintenance* ((5x2,000+5x4,000+5x4,000)/3). *How does that for purchase price compare with that for maintenance, and by how much?*" Formulating an answer should involve mental estimation similar to taking the ratio of \$(25,000/3) to \$(50,000/3), or 0.5. The eigenvector of pairwise comparison matrix formed in this way yields the same priorities among the criteria as does Schoner and Wedley's baseline procedure.

Where Schoner and Wedley's method may depart from Saaty, Vargas and Wendell (1983) is in the protection it claims to provide from rank reversal with the removal or addition of any alternatives.

Saaty, Vargas, and Wendell (1983) recognize the dependency of criteria on alternatives that result in rank reversal and which Schoner and Wedley's procedure corrects for. In recognition of this problem, they state:

> "Note that the above AHP approach [eliciting criteria comparison relative to the sum or average of 'scores' on the criteria] is alternative dependent, in that the elements of the matrix may well change if another alternative (e.g., a car D) is added to the problem. This is a case of interdependence (of attributes on alternatives and alternatives on attributes) that can also be approached in a more sophisticated manner through the generalization of the hierarchy to a system with feedback (see Saaty, 1980, Chapter 9 for details) [i.e., the supermatrix approach to handling dependence of criteria on alternatives]. *However, in practice, when the alternative are not known in advance, one may simplify the analysis initially by attempting to elicit priorities on the attributes without knowledge of the particular alternatives.* (italics added, pg. 12)

elicit priorities on the attributes without knowledge of the particular alternatives. (italics added, pg. 12)

Schoner and Wedley prescribe that AHP comparison questions should reference a common basis for the alternatives being compared. In the car selection example, this basis is lifetime dollar cost. This reference makes the standard of comparison explicit compared to simply asking "Which criteria is more important and by how much?" Watson and Freeling, as well as Saaty and Vargas (1983) and Harker and Vargas (1987), have written that it is the total costs to accrue under each category that should be compared in determining priorities between the sources of these costs.

However, Schoner and Wedley also argue that a specific type of AHP question should reference a specific function of the attribute values present in the alternatives. Again, it is not sufficient to simply ask "Which criteria is more important and by how much?" Rather, the question must take a form consistent with the method used for determining the priorities of the criteria. In the example discussed above, such a question takes the form -- *"Consider the average/total contributions of criteria A and B over the current set of alternatives to the problem of choosing a best alternative. Which average/total contribution is greater and by how much?"* This question satisfies Watson and Freeling's argument that AHP questions should reference a standard for making comparisons. It also is exactly the kind of question prescribed by Saaty and Vargas (1983) in response to Watson and Freeling's argument.

In practice, the kind of question prescribed by Schoner and Wedley may not be easy to ask well, or answer easily or accurately. Schoner and Wedley have discussed this problem in their paper, and with these authors in personal communication. Formulation and comparison of average contributions is easy in the car example because the data are given in the table. However, it may be less easy to formulate and accurately answer questions when tabular data are not given or when the criteria are intangible.

Consider an example in which alternatives vary in size and color. Schoner and Wedley's question to compare the color and size criteria would be -- "For the average size and color of the alternatives being considered, is the average color or the average size more important in its contribution to the value of the alternatives, and by how much?" In order to compare color and size in this way, the respondent needs to mentally score the alternatives on a common footing, in the same way that the three cars were scored in terms of lifetime cost. This is the first problem. The analyst might have to explicitly provide a method for doing this rather than leaving it up to the ingenuity of the respondent. If the respondent is not able to do so, then they might improvise an answer not consistent with what the analyst expects from the response. A second problem involves being able to denote an "average color." A method for doing this also might require a

suggestion from the analyst. It is not clear that the mean frequency of the waveforms represented is what we mean by average color.

Schoner and Wedley also analyzed the method of deriving overall priorities of alternatives suggested by Belton and Gear (1983). Recall that Belton and Gear formulated their method to preserve ranks with the addition or removal of alternatives.

In Belton and Gear's method, the priority of an alternative with respect to a criterion, w_{ik}, equals T_{ik}/T^*_k, where T^*_k is the absolute measurement of the largest valued option under criterion k. In addition, the matrix of comparisons among the alternatives is different from that for conventional AHP because the eigenvector priorities are normalized by dividing each by the largest element of the eigenvector. In conventional AHP, the eigenvector priorities are normalized by dividing each by their sum, thus constraining them to sum to 1. In the car selection example, T^*_1=\$14,000 purchase cost, T^*_2=\$4,000 per year maintenance cost, and T^*_3=0.05 gallons per mile in fuel consumption. The criteria weights are calculated as the product of T^*_k and the scaling factor for the kth criterion.

Schoner and Wedley further suggest that to gather judgments consistent with Belton and Gear's method, the respondent must evaluate the alternatives in terms of a common basis (i.e., contribution to lifetime cost in the car selection example) and determine whether the largest valued option on criteria i is greater than that on criteria j, and by how much. The largest scaled-purchase price (e.g., Criterion 1) is \$14,000 and the largest scaled-maintenance cost (e.g., Criterion 2) is \$20,000 (5x4,000). Thus, the a_{21} entry of the matrix of pairwise criteria comparisons among the criteria is 1.43.

Schoner and Wedley remark that the comparison question consistent with Belton and Gear's method is probably easier to answer than that consistent with their own technique. This is because it makes reference to the best option under each criterion rather than a composite of the scores of all of the options under each criterion.[6] For example, in the car selection example, one such question might be *"Among the alternatives, consider the alternative with the greatest purchase price and the alternative with the greatest lifetime maintenance cost. Is the purchase price or the maintenance cost greater, and by how much?"*

---

[6] Schoner and Wedley (1990 and personal conversation) have recently extended Belton and Gear's method. The extension involves a substantial relaxation of the requirement that the "score" of the best alternative under each criterion be used as the standard for judgments. In their extension, the choice of alternative for each criteria is arbitrary. The only constraint in making this choice is that the priorities of the alternatives relative a given criterion would be normalized so that the arbitrary standard has a value of 1.

Schoner and Wedley also describe an extension of their method to Saaty's absolute measurement case (see Section C.2, this paper), in which the evaluation categories (i.e., excellent, average, poor) are viewed as alternatives nested under the categories they are used to rate alternatives on.

Schoner and Wedley state that their extensions to the conventional AHP only apply to the decision alternatives and the categories they depend on; other category levels throughout the hierarchy may be treated according to the method prescribed by conventional AHP. In personal conversation with these authors, Wedley has explained that this is because the method is only required when elements at a given level of the hierarchy are dependent on elements at the next level down. Wedley and Schoner are currently evaluating whether structural dependence exists elsewhere in the hierarchy, other than at the lowest level.

However, recalling our discussion above, we feel that questions explicitly directing the respondents are necessary in making all comparisons, because the general question leaves too much leeway for respondents to interpret what the question means and what should be done to answer it. The attendant loss of experimental control may compromise the interpretability of the results as a consequence.

2. *Absolute Measurement Scales*

An important extension of the core AHP methodology involves making absolute rather than relative judgments on the decision alternatives. Saaty (1986, 1990) summarizes this extension of the AHP to absolute measurement; the current version of the "Expert Choice" software implementation of the AHP includes a module for absolute measurement.

The absolute measurement technique requires two steps different from the standard AHP methodology: 1) rating each decision option with regard to each criteria (in one particular example, Saaty (1986d) suggests using a seven category rating scale bounded by "excellent" on the high side, "average" in the middle, and "very poor" on the low side); and 2) pairwise ratio comparisons of the rating categories for each criteria to determine their relative priorities. We then replace the alternatives' ratings with the priorities estimated for the respective rating categories.

When applicable, the absolute measurement has several advantages over relative measurement. If there are a very large number of alternatives, pairwise comparison with respect to each criterion may be prohibitive. Further, Saaty reports that the results of the absolute measurement technique are not subject to rank reversal (see Section B, above, on criticisms).

Finally, prioritizing options by relative comparison may not be sufficient when the best alternative is not good enough in an absolute sense.

Saaty (1986) provides examples of the absolute measurement technique in the areas of student admissions and rating cities according to their livability.

Forman (1987) provides a hypothetical example regarding the rating of basketball players with different mixes of offensive and defensive capabilities. As part of his paper, Forman concludes that the absolute measurement approach fails to capture the intuition that the relative value of a resource varies with its scarcity. However, the reasoning underlying this conclusion is incorrect. As part of his model, Forman assigns equal value to offensive and defense skills. However, as he points out, with the increased availability of great offensive players, we may not be willing to pay as much for a great offensive player. Thus the question to be asked regarding offensive and defensive skills should not be "How important are offensive and defensive ability to team management? (Forman, 1987, pg. 195), but "How much would you be willing to pay for offensive ability relative to defensive ability?". Asking the correct question would "correct" for the increased availability of offensive skill *as perceived by team management.*

3. Prioritization Methods

Several authors have suggested alternative methods to Saaty's (1980) eigenvector method for estimating the priorities underlying the matrix of pairwise comparisons. We will review some of the arguments made on behalf of the major alternatives, and Saaty's replies. The interested reader is referred to the research papers referenced in the bibliography below.

The two major alternatives to eigenvector prioritization have been a logarithmic least squares (LLS) method and a least squares (LS) method. In addition, McCurdy (1989) has developed a robust regression (RR) method for estimating underlying priorities and Zahedi (1985) has developed what she terms the "mean transformation method."

The LLS method entails minimizing $\sum_{i,j}\left(ln\left(a_{ij}\right)-ln\left(u_i/u_j\right)\right)^2$, where the u are the estimated priorities.

The solution to this minimization problem is given by $u_i=\prod_{j=1..n}\left(a_{ij}\right)^{1/n}$, $i=1, 2,...n$. In a consistent matrix, $a_{ij}=u_i/u_j$, where a_{ij} is the relative priority judgment comparing hierarchy elements i and j and u_i is the priority of element i. In an inconsistent matrix, the two quantities will

differ. The LLS estimator for u_i is the geometric mean of the n elements of the ith row of the matrix of pairwise comparisons. The geometric mean of n numbers is the nth root of their product.

The LS method is similar to the LLS, except that the function to be minimized is

$$\sum_{i,j}\left[\left(a_{ij}\right) - \left(u_i / u_j\right)\right]^2.$$

However, unlike the LLS method, there is no closed form solution to the problem of estimating the u_i. Jensen (1984) has argued that the LS method yields priority estimates that have a number of important advantages over estimates produced by eigenvector prioritization. A disadvantage to implementing the LS method, however, is that the minimization problem does not have a closed form solution. As a result, the priority estimates must be made using numerical methods. In addition, there is no guarantee of a unique solution to the minimization problem (Saaty and Vargas, 1984).

All prioritization methods are techniques for estimating the unknown weights underlying the noisy but otherwise consistent judgments of relative priority. Whereas the LLS, the LS and the RR methods are developed from explicit loss functions to be minimized, eigenvector prioritization is not.

Thus based on different assumptions and loss functions, we should not expect the four methods to have identical properties. The methods could be judged on any number of criteria, including accuracy (i.e., statistical bias, efficiency). ability to preserve known rankings, robustness under the violation of assumptions, ease of implementation, the availability of auxiliary information associated with the method (e.g., the eigenvector method's consistency-indexing statistics), generality to special problem conditions, etc. Further, different methods could be reasonably adopted under different circumstances.

Crawford and Williams have reported analyses of the LLS method (1984, 1985, and Crawford, 1987) and have suggested several criteria upon which it [LLS] may be preferred to eigenvector prioritization. Among their important arguments, they conclude that the LLS method

> "...is statistically superior. The geometric mean is the maximum likelihood estimator and is thus also the least squares estimator of the priorities underlying the comparison judgments. As a result, it has all of the "usual desirable properties of least squares estimates," such as unbiasedness and minimum variance. Further, the LLS method can be adapted to cases of missing or multiple judgments;
>
> is easier to calculate;

gives rise to a measure of consistency with known statistical properties. Crawford and Williams show that S^2, the sum of squared differences between $ln\ a_{ij}$ and $ln(w_i/w_j)$, divided by $(.5)(n-1)(n-2)$, is an unbiased estimator of the variance of the disturbances, and "hence is a natural measure of consistency of A" (Crawford and Williams, 1984, pg. 19). Recall that inconsistency will be directly related to the magnitude of the disturbances."

As part of their Monte Carlo analysis of random "judgment" matrices, Budescu et al (1987) also reported critical values for S^2. Their tables include these values for matrices of sizes 4, 6, and 8 at three levels of "significance." However, their regression equations for estimating the critical values of other-sized matrices do not perform well in estimating the tabled values.

As an alternative criterion, Crawford (1987) developed a table that provides values for S^2 that correspond to Saaty's (1980) CI=0.10 criterion for consistency.

Saaty and Vargas (1984b) criticize the LLS method, in part on the ground that for $n>3$, the LLS method uses only the data in the ith row of the matrix of comparisons to determine the priority of the ith comparison element. As a result, it produces rankings insensitive to inconsistencies between the rows.

In concluding their paper, Saaty and Vargas show that the LLS method and the LS method yield priority vectors different from eigenvector prioritization when judgments are inconsistent (in addition to the aforementioned uniqueness problem of the LS method). Further, the vectors imply different rank orders of the elements compared, implying that eigenvector prioritization more adequately preserves the rank order of the preferences underlying the judgments. However, Crawford and Williams (1984, 1985) have shown in Monte Carlo simulations that the LLS method performs no worse and frequently performs better than eigenvector prioritization on several criteria, including rank preservation.

The study was based on pseudo-randomly perturbed matrices that Crawford and Williams constructed from consistent matrices. For this purpose they used the error model $a_{ij}=(W_i/W_j)(1+d_{ij})$, which all researchers assume for AHP judgment matrices. Using this error model, Crawford and Williams constructed perturbed matrices from consistent baseline judgment matrices of dimension 5, 7, or 10. They pseudo-randomly drew the d_{ij} disturbances from either lognormal and from uniform distributions. They also varied the variance of these distributions between 0.01 and 1.0, and thereby varied the resulting average CIs.

Crawford and Williams (1984, 1985) compared eigenvector prioritization with the LLS method on several criteria, including the summed squared differences between estimated and actual weights, and the summed squared differences between estimated and actual log weights. They also reported the number of rank reversals and the sum of squared differences in ranks that

occurred. Finally, they reported the percentage of trials in which the LLS method outperformed eigenvector prioritization on each of the aforementioned criteria.

As Crawford and Williams did not report the matrices used as consistent baselines, we can only assume that the matrices were "different" under the different size conditions and were representative of an interesting class of judgment matrices. As a result, we cannot determine if their results are independent over different matrix sizes and whether their results are generalizable to all "types" of judgment matrices. However, given these caveats, they found that the LLS method was essentially equal to or better than eigenvector prioritization with regard to squared differences.

The LLS method similarly dominated eigenvector prioritization with regard to rank preservation, although less strongly and less clearly so in the case of uniformly distributed disturbances. What is less clear is the practical significance of the absolute differences between the methods. For instance, consider the difference between the LLS method and eigenvector prioritization with respect to the sum of squared differences of logs. For disturbance term variances between 0.01 and 1.0, this statistic varied between 0.2 percent and 8.9 percent for matrices of size 5, and between 0.2 percent and 15.8 percent for matrices of size 7.

Crawford and Williams also reported that the superiority of the LLS method over eigenvector prioritization increased with the size of the comparison matrix and with the variance of the disturbances. They reported that both these results were expected on theoretical grounds. They thus concluded that for the conditions they explored in their study, the LLS method is superior to eigenvector prioritization with respect to accuracy measures.

Barzilay, Cook, and Golany (1987) also discuss criteria that favor the LLS method over eigenvector prioritization. In their paper they showed that the "only solution satisfying consistency axioms for the problem of retrieving weights from inconsistent judgements matrices whose entries are the relative importance ratios of alternatives is the geometric mean." (Barzilay et al, 1987, pg. 1).

Zahedi (1986) also conducted a Monte Carlo study of alternative estimators of the priorities underlying inconsistent judgment matrices. However, the results of the study are limited with regard to practical applications because Zahedi allowed "judgments" on the interval [.00001, 100,000]. Zahedi's results were that her mean transformation method was no worse or better than either the LLS method or eigenvector prioritization when the disturbances were gamma distributed. Otherwise, all methods performed about equally well when the disturbances were either lognormally distributed or uniformly distributed.

Zahedi's mean transformation method estimates the priority w_j as the mean of a statistic b_{ij} with respect to the index i. The statistic b_{ij} is found by transposing the comparison matrix and dividing each row element by the sum of the elements in that row.

An interesting suggestion made by Zahedi was to collect data for all off-diagonal elements of the comparison matrix, rather than forcing a_{ij} to equal $1/a_{ji}$. She observed that the accuracy of the estimates could be substantially improved under some conditions of disturbance distribution and matrix size by requiring simulated judgments for all off-diagonal elements. However, note that this condition violates the AHP's reciprocity assumption and does not leave the matrix of pairwise judgments a positive reciprocal matrix. It remains to be seen whether this result holds for real judgment data.

4. Aggregating Group Opinion and Missing Data

The AHP can be used both for individual and group decisionmakers. When used with a group, Saaty (1980) recommends that the group reach consensus on the judgments rather than averaging their responses, particularly if the group is well-informed on the subject domain. However, if the disparate responses need to be averaged, Saaty suggests using their geometric mean because it is compatible with the constraints that $a_{ij}=1/a_{ji}$. (Aczel and Saaty, 1983).

As an alternative, Crawford and Williams (1985) demonstrate that an extension of the LLS method also can be easily adapted to handling the responses of multiple judges. McCurdy (1989) has further enhanced the LLS method by developing a robust regression variant of it that also handles multiple respondents.

A second area in which extensions to the AHP have added to its versatility is that of missing observations. This area is important because the number of comparisons required in a given study may be excessive from the point of view of data collection. In addition, the practitioner may have to handle the very practical case of missing responses.

Crawford and Williams (1985) and McCurdy (1989) show how the LLS method and its robust variant can be extended to the case of missing observations. In addition, Harker (1987a) has discussed the issues of incomplete data and of reduction in the number of questions that must be asked in an AHP analysis. His paper developed a method based on gradient of the right Perron vector of the judgment matrix to handle the cases of missing observations. He then used Monte Carlo simulation analysis to evaluate the adequacy of this approach. Harker (1987b) refined this approach and provided further simulation evidence on its adequacy.

BIBLIOGRAPHY (Extensions)

K.O. Bowman, C.S. Cheng, L.J. Gray, W.L. Lever, T.J. Mitchell, L.R. Shenton, and V.R.R. Uppuluri. *Logarithmic Least-squares Approach to Saaty's Decision Problems*. Oak Ridge National Lab. Mathematics and Statistics Research Dept., (undated).

O.L. Smith. *Draft Alternative Methodology For Pairwise Comparisons in Sotaca*. Oak Ridge National Laboratory Working Paper., (undated).

V. Belton, T. Gear. "Assessing Weights by Means of Pairwise Comparisons," (undated).

R.E. Jensen. "Aggregation (Composition) Schema For Eigenvector Scaling of Criteria Priorities in Hierarchical Structures," *Multivariate Behavioral Research*, January 1983, pp. 63-84.

V.R.R. Uppuluri. *Expert Opinion and Ranking Methods*. Oak Ridge National Laboratory NUREG/CR-31, February 1983.

R.E. Jensen. "An Alternative Scaling Method for Priorities In Hierarchical Structures," *Journal of Mathematical Psychology*, Vol. 28, 1984, pp. 317-332.

T.L. Saaty and L. Vargas. "Comparison of Eigenvalue, Logarithmic Least Squares and Least Squares Estimating Methods In Estimating Ratios," *Mathematical Modelling*, Vol. 6, 1984, pp. 309-324.

J.J. Buckley. "Fuzzy Hierarchical Analysis," *Fuzzy Sets and Systems*, Vol 17, 1985, pp. 233-247.

G. Crawford and C. Williams. *The Analysis of Subjective Judgment Matrices*, The Rand Corporation, R-2572-1-AF (1985); also appears in "A Note on the Analysis of Subjective Judgment Matrices," *Journal of Mathematical Psychology*, Vol. 29, 1985, pp. 387-405.

J. Fichtner. "On Deriving Priority Vectors from Matrices of Pairwise Comparisons," *Socio-Economic Planning Sciences*. Vol. 20, No.6, 1986, pp. 341-345.

T.L. Saaty. "Absolute and Relative Measurement With the AHP. The Most Livable Cities in the United States," *Socio-Economic Planning Sciences*, Vol. 20, No. 6, 1986, pp. 327-331.

F. Zahedi. "A Simulation Study of Estimation Methods in the Analytic Hierarchy Process," *Socio-Economic Planning Sciences*, Vol. 20, No. 6, 1986, pp. 347-354.

J. Barzilay, W.D. Cook, and B. Golany. "Consistent Weights for Judgments Matrices of the Relative Importance of Alternatives," *Operations Research Letters*, Vol. 6, No. 3, 1987, pp. 131-134.

J.J. Buckley and V.R.R. Uppuluri. "Fuzzy Hierarchical Analysis, " in V.T. Covello, L.B. Lave, A. Moghissi, and V.R.R. Uppuluri (Eds.) "Uncertainty in Risk Assessment, Risk Management, and Decision Making," New York: Plenum Press, 1987.

D.V. Budescu, R. Zwick, and A. Rapoport. *A Comparison of the Analytic Hierarchy Process and the Geometric Mean Procedure for Ratio Scaling*, U.S. Army Research Institute for the Behavioral and Social Sciences Research Note 87-70, December 1987.

G.B. Crawford. "The Geometric Mean Procedure for Estimating the Scale of a Judgment Matrix," *Mathematical Modelling*, Vol. 9, Nos. 3-5, 1987, pp. 327-344.

S.Y. Dennis. "A Probabilistic Model for the Assignment of Priorities in Hierarchically Structured Decision Problems," *Mathematical Modelling*, Vol. 9, Nos. 3-5, 1987, pp. 335-343.

D.M. DeTurck. "The Approach to Consistency in the Analytic Hierarchy Process," *Mathematical Modelling*, Vol. 9, Nos. 3-5, pp. 345-352, 1987.

P.T. Harker. "Alternative Modes of Questioning in the Analytic Hierarchy Process," *Mathematical Modelling*, Vol. 9, Nos. 3-5, 1987a, pp. 353-360.

P.T. Harker. "Incomplete Pairwise Comparisons in the Analytic Hierarchy Process," *Mathematical Modelling*, Vol. 9, No.11, 1987b, pp. 837-848.

T.L. Saaty. "How to Handle Dependence with the Analytic Hierarchy Process," *Mathematical Modelling*, Vol. 9, Nos. 3-5, 1987, pp. 369-376.

T.L. Saaty. *A Note on Decisionmaking and Number Crunching, Is Normalization the Answer?* Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA, (undated, appears in the appendix of the T.L. Saaty. "The Analytic Hierarchy Process", 1988 edition).

B.L. Golden, Q. Wang. "An Alternate Measure of Consistency," in B.L. Golden, E.A. Wasil, P.T. Harker (Eds.) "The Analytic Hierarchy Process, Applications and Studies," Berlin: Springer-Verlag, 1989.

R.E. Jensen. "Ordinal Data AHP Analysis: Block Clustering Comparisons of Multiple Respondents," Department of Business Administration, Trinity University, San Antonio, TX, Working Paper 180, 1989a.

R.E. Jensen. "Consistency Adjustment Surrogacy of Human Responses in AHP Analysis: Issues, Controversies, and State of the Art," Department of Business Administration, Trinity University, San Antonio, TX, Working Paper 182, 1989b.

R.E. Jensen. "Ordinal Data AHP Analysis: Measures of Rank Association and Consistency," Department of Business Administration, Trinity University, San Antonio, TX, Working Paper 183, 1989c.

M. McCurdy. *Two Enhancements of the Logarithmic Least-squares Method For Analyzing Subjective Comparisons*, Headquarters of the Commander in Chief, U.S. Pacific Command, Strategic Planning and Policy Directorate, Research and Analysis Division Technical Memorandum, 1989.

J. Prillaman. "Identifying Expert Inadequacies," presented at the 1989 Annual Washington Operations Research/Management Science Council Symposium, 1989.

B. Schoner and W.C. Wedley. "Alternative Scales in AHP," in "Springer Lecture Notes in Economics and Mathematical Decisions," forthcoming, 1989.

W.C. Wedley. "Consistency Prediction with Incomplete AHP Matrices," Faculty of Business Administration, Simon Fraser University, Burnaby, British Columbia, paper presented at the International Symposium of the Analytic Hierarchy Process, Tianjin, China, 1988b.

B. Schoner and W.C. Wedley. "Ambiguous Criteria Weights in AHP - Consequences and Solutions," *Decision Sciences*, Summer 1989, pp. 462-475.

W.C. Wedley. "Consistency Tests for Incomplete AHP Matrices: A Comparison of Two Methods," Faculty of Business Administration, Simon Fraser University, Burnaby, British Columbia. paper presented at the ASAC 1989 Conference, Montreal, Quebec, 1989.

T.L. Saaty. "Eigenvector and Logarithmic Least Squares," Joseph M. Katz Graduate School of Business, University of Pittsburgh, submitted for publication, 1990.

B. Schoner, W.C. Wedley, and E.U. Choo. "A Unified Approach to AHP with Linking Pins," Faculty of Business Administration, Simon Fraser University, submitted for publication, 1990.

D. BIBLIOGRAPHY: (APPLICATIONS)

D. Kocaoglu. "A Participative Approach To Program Evaluation," *IEEE Transactions On Engineering Management.*, (undated).

T.L. Saaty. "Measuring the Fuzziness of Sets," *Journal of Cybernetics*, Vol. 4, No. 4, 1974, pp. 53-61.

T.L. Saaty and M. Khouja. "A Measure of World Influence," *Journal of Peace Science*, Spring 1976, pp. 31-47.

T.L. Saaty and P.C. Rogers. "Higher Education in the United States (1985-2000): Scenario Construction Using A Hierarchical Framework With Eigenvector Weighting," *Socio-Economic Planning Sciences*, Vol. 10, 1976, pp. 251-63.

J.M. Alexander and T.L. Saaty. "The Forward and Backward Processes of Conflict Analysis," *Behavioral Science*, Vol. 22, No.2, 1977a, pp. 87-98.

J.M. Alexander and T.L. Saaty. "Stability Analysis of the Forward-Backward Process: Northern Ireland," *Behavioral Science*, Vol. 22, No.6, 1977b, pp. 375-382.

T.L. Saaty. "Scenarios and Priorities in Transport Planning: Application To the Sudan," *Transportation Research*, Vol. 11, 1979, pp. 343-350.

T.L. Saaty and L. Vargas. "Estimating Technological Coefficients By the Analytic Hierarchy," *Socio-Economic Planning Sciences*, Vol. 13, No.6, 1979, pp. 333-336.

R.R. Yager. "An Eigenvalue Method of Obtaining Subjective Probabilities," *Behavioral Science*, Vol. 24, No.6, 1979, pp. 382-387.

T.L. Saaty. "The Analytic Hierarchy Process," New York: McGraw-Hill, Inc., 1980 (republished 1988).

T.L. Saaty, M. H. Beltran. "Architectural Design by the Analytic Hierarchy Process," *Design Methods and Theories*, Vol. 14., No. 3/4, 1980, pp. 124-134.

T.L. Saaty, P.C. Rogers, R. Pell. "Portfolio Selection Through Hierarchies," *The Journal of Portfolio Management*, Spring 1980, pp. 16-21.

D.S. Tarbell and T.L. Saaty. "The Conflict In South Africa: Directed Or Chaotic," *Journal of Peace Science*, Spring 1980, Vol. 4, No.2, pp. 151-168.

Y. Wind and T.L. Saaty. "Marketing Applications of the Analytic Hierarchy Process," *Management Science*, Vol. 26, No. 7, 1980, pp. 641-658.

R.E. Jensen. "Scenario Probability Scaling: An Eigenvector Analysis of Elicited Scenario Odds Ratios," *Futures*, Vol. 13, No.6, 1981, pp. 489-498.

R.E. Jensen. "Reporting of Management Forecasts: An Eigenvector Model For Elicitation and Review of Forecasts," *Decision Sciences*, Vol. 13, 1982, pp. 15-37.

T.L. Saaty, J. Alexander. "Thinking with Models," Oxford, England: Pergamon Press, 1981.

T.L. Saaty. "The Analytic Hierarchy Process: A New Approach To Deal With Fuzziness in Architecture," *Architectural Science Review*, Vol. 25, No.3, 1982a, pp. 64-69.

T.L. Saaty. "Decisionmaking For Leaders: The Analytic Hierarchy Process," Belmont, CA: Lifetime Learning Publications, 1982b.

T.L. Saaty, L. Vargas, and A. Barzilay. "High-level Decisions: A Lesson From the Iran Hostage Rescue Operation," *Decision Sciences*, Vol. 13, 1982, pp. 185-205.

K.H. Mitchell and M.P. Soye. "Measuring the Intangibles In Social Decisions: Assessing Benefits and Costs of Energy Policy Options," *Mathematics and Computers in Simulation*, Vol. 25, 1983, pp. 135-145.

T.L. Saaty. "Conflict Resolution and Falklands Islands Invasions," *Interfaces*, Vol.13, No. 6, Dec. 1983, pp. 68-83.

T.L. Saaty, M.M. Wong. "Projecting Average Family Size in Rural India by the Analytic Hierarchy Process," *Journal of Mathematical Sociology*, Vol. 9, No. 3, 1983, pp. 181-209.

L.G. Vargas. "Prospects for the Middle East: Is a Peaceful Settlement Attainable?" *European Journal of Operations Research*, Vol. 14, No. 2, 1983, pp. 169-192.

T.L. Saaty. "Impact of Disarmament Nuclear Package Reductions," in R. Avenhaus and R.K. Huber, "Quantitative Assessments In Arms Control," New York: Plenum Press, 1984, pp. 309-333.

A. Arbel and N. Novik. "U.S. Pressure on Israel," *Journal of Conflict Resolution*, Vol. 29, No. 2, 1985, pp. 253-282.

T.L. Saaty, L. Vargas. "Modelling Behavior In Competition: The Analytic Hierarchy Process," *Applied Mathematics and Computation*, Vol 16, No. 1, 1985, pp. 49-92.

T.L. Saaty, K. Kearns. "Analytical Planning: The Organization of Systems," Oxford, England: Pergamon Press, 1985.

T.L. Saaty and L.G. Vargas. *The Logic of Priorities*, Boston: Kluwer-Nijoff Publishing, 1985.

A. Arbel and S.S. Oren. "Generating Search Directions in Multiobjective Linear Programming Using the Analytic Hierarchy Process," *Socio-Economic Planning Sciences*, Vol. 20, No. 6, 1986, pp. 369-373.

C.J. Debeljak, Y.Y. Haimes, and M. Leach. "Integration of the Surrogate Worth Tradeoff Method and the Analytic Hierarchy Process," *Socio-Economic Planning Sciences*, Vol. 20, No. 6, 1986, pp. 375-385.

R.P. Hamalainen and T.O. Seppalainen. "The Analytic Network Process in Energy Policy Planning," *Socio-Economic Planning Sciences*, Vol. 20, No. 6, 1986, pp. 399-405.

W.R. Hughes. "Deriving Utilities using the Analytic Hierarchy Process," *Socio-Economic Planning Sciences*, Vol. 20, No. 6, 1986, pp. 393-395.

D.L. Olson, M. Venkataramanan, and J.L. Mote. "A Technique Using Analytical Hierarchy Process in Multiobjective Planning Models," *Socio-Economic Planning Sciences*, Vol. 20, No. 6, 1986, pp 361-368.

T.L. Saaty. "Absolute and Relative Measurement With the AHP. The Most Livable Cities In the United States," *Socio-Economic Planning Sciences*, Vol. 20, No.6, 1986a, pp. 327-331.

T.L. Saaty. "Conflict Resolution With the Analytic Hierarchy Process" Lauder Distinguished Lecturer Talk 3/86, The Wharton School, 1986b.

F. Zahedi. "The Analytic Hierarchy Process -- A Survey of the Method and Its Applications," *Interfaces*, Vol. 16, No. 4, 1986, pp. 96-108.

A. Arbel. "Venturing into New Technological Markets," *Mathematical Modelling*, Vol. 9, Nos. 3-5, 1987, pp. 299-308.

N. Bahmani and H. Blumberg. "Consumer Preference and Reactive Adaptation to a Corporate Solution of the Over-the-Counter Medication Dilemma - An Analytic Hierarchy Process Analysis," *Journal of Accounting and Public Policy*, Vol. 1, No. 2, 1987, pp. 293-298.

V.P. Dorweiller. "Legal Case Planning Via the Analytic Hierarchy Process," *Mathematical Modelling*, Vol. 9, Nos. 3-5, 1987, pp. 251-261.

G.A. Grizzle. "Pay for Performance: Can the Analytic Hierarchy Process Hasten the Day in the Public Sector?," *Mathematical Modelling*, Vol. 9, Nos. 3-5, 1987, pp. 245-250.

R.E. Jensen. "International Investment Risk Analysis: Extensions for Multinational Corporation Capital Budget Models," *Mathematical Modelling*, Vol. 9, Nos. 3-5, 1987, pp. 265-284.

Y. Kathawala, H. Gholamnezhad. "New Approach To Facility Location Decisions," *International Journal of Systems Science*, Vol.18, No. 2, 1987, pp. 389-402.

P. Korhonen. "The Specification of a Reference Direction using the Analytic Hierarchy Process," *Mathematical Modelling*, Vol. 9, Nos. 3-5, 1987, pp.361-368.

M.J. Liberatore. "An Extension of the Analytic Hierarchy Process For Industrial R & D Project Selection and Resource Allocation," *IEEE Transactions on Engineering Management*, Vol. EM-34, No. 1, 1987, pp. 12-18.

P.F. Nelson, W.E. Kastenberg, and K.A. Soloman. "A Value-Impact Approach For Regulatory Decision Making: An Application To Nuclear Power," in L. Lave (ed), "Risk Assessment and Management," New York: Plenum Press, 1987.

P.S. Pak, K. Tsuji, and Y. Suzuki. "Comprehensive Evaluation of New Urban Transportation Systems by AHP," *International Journal of Systems Science*, Vol. 18, No. 6, 1987, pp. 1179-1190.

T.L. Saaty. "Risk -- Its Priority and Probability: The Analytic Hierarchy Process," *Risk Analysis*, Vol. 7, No. 2, 1987a, pp. 89.

T.L. Saaty. "A New Macroeconomic Forecasting and Policy Evaluation Method," *Journal of Mathematical Modelling*, Vol. 9, No. 3-5, 1987b, pp. 219-231.

E.N. Weiss. "Using the Analytic Hierarchy Process In A Dynamic Environment," *Mathematical Modelling*, Vol 9, Nos. 3-5, 1987, pp. 211-216.

T.L. Saaty, W.C. Wedley. "Free Trade Discussions Between Canada and the United States," *RCSA/CJAS*, June 1988, pp. 67-76.

M. Toshtzar. "Multi-Criteria Decision Making Approach To Computer Software Evaluation: Application of the Analytical Hierarchy Process," *Mathematical Computer Modelling*, Vol. 11, 1988, pp. 276-281.

J.M. Alexander. "An Analysis of Conflict in Northern Ireland," in B.L. Golden, E.A. Wasil, P.T. Harker (Eds.) "The Analytic Hierarchy Process, Applications and Studies," Berlin: Springer-Verlag, 1989.

T.L. Saaty, J. Alexander. "Conflict Resolution: The Analytic Hierarchy Process," New York: Praeger, 1989.

E. BIBLIOGRAPHY: (DEFENSE APPLICATIONS)

T.L. Saaty and M. Khouja. "A Measure of World Influence," *Journal of Peace Science*, Spring 1976, pp. 31-47.

J.M. Alexander and T.L. Saaty. "The Forward and Backward Processes of Conflict Analysis," *Behavioral Science*, Vol. 22, No. 2, 1977a, pp. 87-98.

J.M. Alexander and T.L. Saaty. "Stability Analysis of the Forward-Backward Process: Northern Ireland," *Behavioral Science*, Vol. 22, No. 6, 1977b, pp. 375-382.

D.S. Tarbell and T.L. Saaty. "The Conflict In South Africa: Directed Or Chaotic," *Journal of Peace Science*, Spring 1980, Vol. 4, No.2, pp. 151-168.

T.L. Saaty, L. Vargas, and A. Barzilay. "High-Level Decisions: A Lesson From the Iran Hostage Rescue Operation," *Decision Sciences*, Vol. 13, 1982, pp. 185-205.

T.L. Saaty. "Conflict Resolution and Falklands Islands Invasions," *Interfaces*, Vol. 13, No. 6, Dec. 1983, pp. 68-83.

L.G. Vargas. "Prospects for the Middle East: Is a Peaceful Settlement Attainable?" *European Journal of Operations Research*, Vol. 14, No. 2, 1983, pp. 169-192.

T.L. Saaty. "Impact of Disarmament Nuclear Package Reductions," in R. Avenhaus and R.K. Huber, "Quantitative Assessments In Arms Control," New York: Plenum Press, 1984, pp. 309-333.

A. Arbel and N. Novik. "U.S. Pressure on Israel," *Journal of Conflict Resolution*, Vol. 29, No. 2, 1985, pp. 253-282.

T.L. Saaty and L. Vargas. "Modelling Behavior In Competition: The Analytic Hierarchy Process," *Applied Mathematics and Computation* , Vol 16, No. 1, 1985, pp. 49-92.

B. Tullington, R. Batcher, K. Guess. *The Evaluation of Individual Weapon Effectiveness: Part I - The Hierarchical Analysis*, Battelle Columbus Division (DTIC AD-B101 602), September 1985.

S.I. Gass. "A Process for Determining Priorities and Weights for Large-Scale Linear Goal Programming Programmes," *OR: The Journal of the Operational Research Society*, Vol. 37, No. 8, 1986, pp. 779-785.

K.H. Mitchell and G. Bingham. "Maximizing the Benefits of Canadian Forces Equipment Overhaul Programs Using Multiobjective Optimization," *INFOR: The Canadian Journal of Operational Research*, Vol. 24, No. 4, 1986, pp. 251-264.

T.L. Saaty. "Conflict Resolution With the Analytic Hierarchy Process: How To Negotiate the Conflict in South Africa," Lauder Distinguished Lecturer Talk 3/86, The Wharton School, 1986.

S.M. Anderson. "A Goal Programming R&D Project Funding Model of the U.S. Army Strategic Defense Command Using the Analytic Hierarchy Process," Naval Postgraduate School thesis, Department of the Navy (DTIC AD-A187 099), September 1987.

A. Arbel, T.L. Saaty, and L.G. Vargas. "Nuclear Balance and the Parity Index: The Role of Intangibles in Decisions," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-17, No. 5, September/October 1987, pp. 821-828.

K.H. Darko. "Validation of the Tactical Air Force's Decision Making Process to Prioritize Modification Using the Analytic Hierarchy Process," Air Force Institute of Technology thesis, Department of the Air Force, Air University (AFIT/GOR/MA/87D-3, DTIC AD-A188 854), December 1987.

G.A. Luethkey. "An Application of the Analytic Hierarchy Process to Evaluate Candidate Locations for Building Military Housing," Air Force Institute of Technology thesis, Department of the Air Force, Air University (AFIT/GOR/MA/87D-10, DTIC AD-A189 780), December 1987.

M.R. Anderson and C. Corley. *Deficiency Prioritization In the Combined Arms Mission Area Analysis*. 57th Military Operations Research Symposium, Fort Leavenworth, 1989.

J.C. Cole. "Silo Basing and Rail Mobile: Determining the Best Basing Mix for the Peacekeeper ICBM," Air Force Institute of Technology thesis, Department of the Air Force, Air University (AFIT/GST/ENS/89M-04, DTIC AD-B130 686), March 1989.

D. Graham. "Applying the Analytic Hierarchy Process to Economic Analyses of MAISRC Programs," U.S. Air Force Cost Center, Arlington, VA, presented at the 23rd DoD Cost Analysis Symposium, 1989.

T. M. Lazenby. "A Practical Approach to Utilizing Expert Opinion in Costing New Technology," U.S. Army Tank-Automotive Command, Warren, MI, unpublished manuscript, 1989.

J.M. Alexander. "An Analysis of Conflict in Northern Ireland," in B.L. Golden, E.A. Wasil, P.T. Harker (Eds.) "The Analytic Hierarchy Process, Applications and Studies," Berlin: Springer-Verlag, 1989.

R.J. Might. "Decision Support for War Games," in B.L. Golden, E.A. Wasil, P.T. Harker (Eds.), "The Analytic Hierarchy Process," Berlin:Springer-Verlay, 1989.

J.G. Vlahakis, W.R. Partridge. "Assessment of Security at Facilities that Produce Nuclear Weapons," in B.L. Golden, E.A. Wasil, P.T. Harker (Eds.) "The Analytic Hierarchy Process, Applications and Studies," Berlin: Springer-Verlag, 1989.

T.L. Saaty, J. Alexander. "Conflict Resolution: The Analytic Hierarchy Process," New York: Praeger, 1989.

T.L. Saaty. "The Analytic Hierarchy Process in Conflict Managment," *The International Journal of Conflict Management*, Vol. 1, No. 1, 1990, pp. 47-68.

F. PUBLISHED AHP BIBLIOGRAPHIES

F. Zahedi. "The Analytic Hierarchy Process -- A Survey of the Method and Its Applications," *Interfaces*, Vol. 16, No. 4, 1986, pp. 96-108.

B.L. Golden, E.A. Wasil, and D.E. Levy. "Applications of the Analytic Hierarchy Process: A Categorized, Annotated Bibliography," in B.L. Golden, E.A. Wasil, and P.T. Harker (eds), "The Analytic Hierarchy Process, Applications and Studies," Springer Verlag: New York, 1989.

G. EVALUATION

Our evaluation of the AHP falls into two general areas. The first regards implementation issues; that is, how should an AHP analysis be performed. The second involves theoretical questions raised by the critical literature and by our own observations.

1. AHP Implementation Issues

A number of authors have questioned whether the assumption of independence of criteria on alternatives (Saaty, 1980) is ever satisfied in AHP hierarchies (Schoner and Wedley, 1989; Dyer and Wendell, 1985; Dyer, 1990a). However, the practitioner does not have to take this extreme position to understand that assumptions should be tested whenever possible, rather than being taken for granted. AHP users should carefully evaluate their hierarchies for dependence

between and within hierarchy levels, noting that this is an empirical assessment as well as a logical assessment. Thus, for instance, judges should make pairwise assessments of the relative priority of criteria with regard to each lower order element.

If the priorities among the criteria differ over alternatives, then remedial measures need to be taken. One difficulty with doing this at present is that there are no formal guidelines for estimating the "amount" of dependence present, and given a loss function, for deciding that the conventional AHP is not appropriate for a given problem. Therefore, a conservative analyst may want to assume dependence and either use the supermatrix method suggested by Saaty (1980) or the equivalent methods recommended by Schoner and Wedley (1989) or by Dyer (1990a; Dyer and Wendell, 1985).

AHP practitioners should also consider tailoring their judgment-eliciting questions to the requirements of their study, as suggested by Watson and Freeling (1983), Saaty, Vargas, and Wendell (1983), Schoner and Wedley (1989), and Dyer (Dyer and Wendell, 1985; Dyer, 1990a). This means providing standards and ranges upon which the respondents will base their judgments. In addition to targeting the questions to the requirements of the study (and there well may be no particular requirements), this step standardizes the procedure so that the results may be interpreted in the context of the questions asked and the results obtained from different respondents may be at least comparable in terms of the questions asked. In addition, Schoner and Wedley suggest that the kinds of questions asked have implications for the procedure used to determine the relative priorities among the elements being compared. If Schoner and Wedley are correct, the questions asked and the prioritization procedure used should be kept compatible with each other.

All of the above comments, as well as other issues raised in the critical literature, require that AHP studies be pilot-tested. Pilot testing allows the practitioner to determine usable question formats, detect near copies, diagnose dependencies, and take remedial measures when necessary.

Questions remain with regard to the kinds of conclusions that can be drawn from AHP results. The criticisms of Veit and Birnbaum et al from the psychological measurement area and of Meyer and Booker suggest that results from the AHP are not ratio-scaled. Dyer and Schoner and Wedley come to similar conclusions, although from a different perspective. In practical terms, this means that, at minimum, practitioners may not want to give a meaningful interpretation to the ratios between priorities estimated for AHP alternatives. If the results show option A with a priority of 0.6 and option B with a priority of 0.3, analysts may not want to conclude that option A is preferred twice as much as option B, but only that option A is preferred to option B. Note that Dyer's criticism suggests that even this kind of conclusion may be too much, and that a more careful consideration of Veit and Birnbaum's position may agree with Dyer.

Finally, Saaty (1987a) and Harker and Vargas (1987) argue that the AHP is not suitable for handling cases in which the alternatives include "near copies subsets" (Dyer (1990a) has criticized the method for this limitation). However, this argument creates a problem for the practitioner. There is a need for a method that does not require repetitive application of the AHP to empirically identify near copies using Saaty's heuristic. In practice we have observed studies wherein the authors did not have the time to iteratively query their respondents, identify and remove near copies, and then requestion their respondents. Conventional AHP might not be suitable for these studies, although Schoner and Wedley's variation, which is claimed to be insensitive to near copies, might be.

The bottom line with regard to implementation issues is that the AHP is not an "easy" method to use. Existing software and the examples of AHP analysis reported in the literature certainly give the appearance of ease of application. However, very few of these cases report rigorous empirical analysis of dependency relations within the hierarchy, concern for careful design of judgment-eliciting questions, or post-analysis search for near-copy alternatives.

2. Theoretical Issues

At the heart of the theoretical evaluation are unanswered questions regarding assumptions underlying the AHP methodology.

Saaty's psychological model of preference judgments states that respondents form ratios of impression when asked to do so. However, accumulating evidence from the psychophysics and judgment literature suggests that this may not be true. How people respond to AHP-type questions is an empirical question and should be addressed empirically as a critical assumption of the methodology. Answering this question may thus be important for those interested in using the AHP, whether to determine preference orderings or to take advantage of the ratio-scale properties claimed for it. It may turn out in fact that the AHP might be robust enough because of some yet unidentified result to provide "good enough" approximations to ratio-scale results. However, this is only a speculation at present.

Dyer and others have claimed that the AHP is nothing other than an additive MAV/MAU method, and thus is underpinned by the same set of independence assumptions and should be judged in terms of the same criteria as this class of methods. However, Saaty and this colleagues disagree, particularly on the significance of rank reversal to the validity of the method. Thus, the whole issue of whether the AHP requires remediation depends on the resolution of this issue in the mind of the user.

Given a belief that the AHP requires remediation, Schoner and Wedley's and Dyer's suggestions regarding corrective measures to the AHP need to be given more careful thought and discussion to determine their utility. However, their work shows promise as a unified approach to framing specific AHP comparison questions and correcting for the possibility of rank reversal and related problems.

Finally, while recognizing inconsistency in human judgment, AHP theory has no theory for judging differences between weights -- if two alternatives respectively have priorities of 0.21 and 0.20, should we treat the first as being more important than the second? Intuition suggests that the answer is no, first because the difference is small and probably is not significant in a practical sense, and second, because the inaccuracy and low reliability of human judgment and perception make the difference appear more to be part of the "noise than the signal." However, what if the two priorities were 0.23 and 0.20?, or 0.25 and 0.20?, or 0.30 and 0.20?, or 0.40 and 0.20? When does the difference in priorities become credible? What are the criteria that should be used in assessing the credibility of a numerical difference in priorities? The literature does not yet provide any guidance with regard to judging the results of an AHP analysis. However, this is not a criticism of the AHP as much as a suggestion for extending the AHP in a way that would make it more useful to the practitioner.

IV. SUBJECTIVE TRANSFER FUNCTION METHOD

This chapter presents an analysis of the subjective transfer function method of forecasting. It is organized into five major subheadings, each followed by a (chronological) listing of reference literature relevant to that particular major subheading, if applicable. Each of the five papers presented in this study follows the same general outline for ease of reader reference and comparison. Before beginning our dissertation on the subjective transfer function (STF) method, we offer a brief historical background of its development.

The STF method is a procedure for assessing and representing expert knowledge of complex systems in which well-defined outcomes are related to factors that influence them. Veit and Callero developed the STF method at the Rand Corporation and with their colleagues, have modelled several systems, including tactical air command and control in a Korean-based land battle (Callero, Naslund, and Veit, 1981b, c) and immediate targeting systems in battlefield-air interdiction (Veit, Callero, and Rose, 1984). More broadly, the STF method could be used to analyze any systems that could be expressed in terms of well-defined outcomes and factors that influence them, including political/economic systems, strategies, organizational structures, and multi-objective preference systems.

The result of an STF analysis is a model that takes the form of hierarchically linked mathematical functions that relate system characteristics to system outcomes. This structure then allows the user to 1) identify tradeoff relationships between factors, 2) assess the impacts of modifying the level of factors, and 3) compare alternative systems in terms of the outcome.

A. DESCRIPTION

The subjective transfer function method has three main "steps." First, the user develops hierarchical models of the system of interest. These models take the form of a hierarchy of outcomes (e.g., targets killed, materiel transported) and the factors that influence them (e.g., target availability, munitions availability, delivery systems' accuracy, target location accuracy, weather, etc.). Figure IV-1 is a hierarchical representation developed by Veit, Callero, and Rose (1982c) to model the battlefield air interdiction

immediate targeting task. Table IV-1 defines each of the outcomes and factors appearing in Figure IV-1.

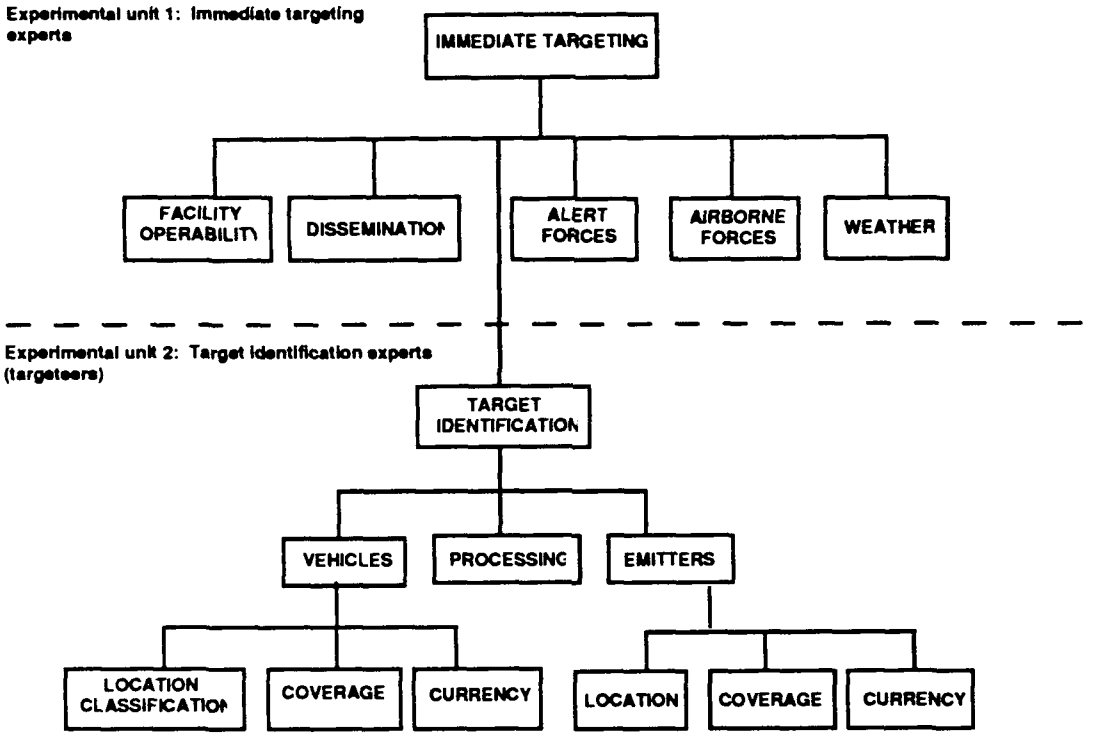


Figure IV-1. Hypothesized Immediate Targeting Structure

The hierarchical form of the model allows for treating some outcomes as factors which influence higher order outcomes. This hierarchical form also allows the analyst to partition the model into outcome-factors units ("experimental units"), which respondents can judge separately according to their expertise.

In Figure IV-1, the immediate targeting outcome is influenced by the six factors subordinate to it in the model. Furthermore, one of the factors, "Target Identification," is itself a sub-outcome dependent on subordinate factors. The two outcomes, "Immediate Targeting" and "Target Identification," and their subordinate factors represent two separate judgment experiments that can involve the knowledge of different respondent groups, Air Force immediate targeting specialists ("Those who put bombs on targets") and Air Force targeteers.

Given one or more system models, respondents next assess outcomes as a function of varying levels of the factors. The STF methodology[1] requires study designs that

[1]The STF methodology itself incorporates an approach to the measurement of subjective values termed "algebraic modelling." This paper will generally not distinguish between those aspects of an STF analysis

simultaneously manipulate the levels of all factors (a "fully-crossed" factorial design), as well as the number of factors simultaneously considered. Each description consisting of a set of factors set to specific levels is a *scenario*. The following is one such scenario formulated with respect to immediate targeting outcomes with the first experimental unit of Figure IV-1.

> "Thirty percent of the important second echelon forces are identified in a timely fashion. Facilities can support 60 percent of the necessary immediate targeting activities. Tasking can be correctly communicated to 60 percent of the forces in time. There is timely access to the status of 10 percent of the Alert and Airborne forces. Weather data are three hours old. What percent of force application opportunities can be exploited in a timely manner?" (Veit, Callero, and Rose, 1982c, p. 17)

Table IV-1. Immediate Targeting Task: Definitions

IMMEDIATE TARGETING: the percent of important force application opportunities that can be exploited by matching a proper tactical air weapon system with an important enemy target at an appropriate time.
FACILITY OPERABILITY: the percentage of all necessary immediate targeting activities the facility can perform if all effective force applications were to be exploited.
DISSEMINATION: the percentage of forces to which tasking can be correctly communicated in a timely fashion.
TARGET IDENTIFICATION: the percentage of important second echelon force elements that can be identified in time for the sake of the immediate targeting function.
ALERT FORCES: the percentage of the designated alert forces for which there is timely access to status information.
AIRBORNE FORCES: the percentage of the tactical air forces that can be airborne in time to be used by the immediate targeting function for which there is timely access to status information.
WEATHER: the age of reliable weather data in the second echelon area and at the tactical air bases.
VEHICLE LOCATION CLASSIFICATION: the ability of and conditions under which sensors can locate and classify enemy vehicles (locate, locate and classify, all weather, clear weather only).
VEHICLE COVERAGE: the percentage of second echelon vehicles observed.

unique to the STF method and those inherited from algebraic modelling. However, the interested reader is referred to the following literature for more information on algebraic modelling: Anderson 1974a, 1974b, 1979, 1981.

Table IV-1. Immediate Targeting Task: Definitions (continued)

VEHICLE CURRENCY: the time interval between the observation of vehicles in the second echelon and the data's availability for processing in the command and control system.
PROCESSING: the process by which enemy vehicle and emitter information is processed and interpreted (fully human processing of hard copy textual information and human interpretation, computer processing to sort textual information/human interpretation, computer generation of graphical displays/human interpretation, fully computerized data processing and interpretation).
EMITTER LOCATION CLASSIFICATION: the accuracy with which sensors can locate enemy emitters in the second echelon.
EMITTER COVERAGE: the percentage of enemy echelon emitters in the second echelon that have been observed.
EMITTER CURRENCY: the time interval between the observation of enemy emitters in the second echelon and the data's availability for processing in the command and control system.

In practice, a fully crossed factorial design may be unnecessary, and a "partial" design may be substituted. Veit, Callero, and Rose discuss guidelines for doing this appropriately. (1984, pg. 32)

Using the responses to the scenario questions, the user next assesses the respondents' covert judgment process for using the questionnaire information to form responses. This assessment produces a mathematical function that models responses as a function of factor levels. Callero and Veit term these functions *STFs* because they link two experimental units via the common entity that is an outcome in one unit and a factor in another. Together with the structural model of the system, these functions represent the respondents' knowledge of how the system functions.

Given the final STF model of the system, we can evaluate the performance of alternative systems by comparing the outcomes for different factor level inputs. Since the STF models consist of mathematical functions, factor levels not specifically evaluated in the assessment study can be examined when they refer to physically-measurable variables. We can readily identify tradeoff relationships by plotting the effects of several factors on outcomes.

We expand on the procedures of the STF method below, following an introduction to the theory that underlies it and that motivates it.

1. Defining Characteristics of The STF Method

The STF method has several defining characteristics. First, it models algebraically the process by which information is covertly combined into a response. Second, it provides entities which serve as outcomes in one experimental unit and as factors, and therefore inputs, into a higher order experimental unit. Thus, the hierarchical levels of the STF model are functionally interlinked without requiring the assumption of an arbitrary aggregation rule.

We introduced the concept of psychological representation processes in Chapter I of this document and demonstrated the significance of the psychological processes that mediate between the statement of a task and the response to the statement. Veit and Callero extend this argument by asserting that correctly modelling the psychological processes underlying respondents' judgments is essential to giving the judgments a meaningful and valid interpretation.

The psychological processes that underlie judgments dictate the properties that we may attribute to the resulting data. Most important for our purposes, these properties include the meaningfulness of various kinds of relationships between numerical judgments, such as differences and ratios -- Is the difference indicated by ratings of 9 and 10 interpretable as equalling that between ratings of 1 and 2? Does a rating of 4 indicate twice the magnitude of a rating of 2? The answer to both these questions is -- the numerical relationships between judgments are <u>not</u> directly interpretable. Numerical scale values, which are interpretable, are derived from models of the judgment process underlying the ratings. Then, the numerical interpretability of the scale values depends on the properties of the scale.[2]

Veit and Callero argue that the psychological judgment processes, and thereby the properties of the numerical judgments, should be assessed rather than assumed. An invalid assumption about the properties of judgment data can invalidate the conclusions that depend on them, while the assessments are conceptually easy to make.

Careful attention to questionnaire designs in the STF methodology permits the discrimination of alternative models of the the psychological processes underlying the

---

[2]One example wherein a simple numerical relationship is not meaningful involves temperature measurement on the Fahrenheit scale, which does not allow for meaningful statements of ratios between temperature measurements. A temperature of 100 degrees Fahrenheit is not twice as "hot" as a temperature of 50 degrees Fahrenheit. However, measured on the correct scale (i.e., Kelvin), such a relationship between temperatures of 100 degrees and 50 degrees is meaningful.

responses. Each model makes a variety of predictions regarding how judgments of outcomes vary with different values of the factors that influence them. It is by collecting these judgment data according to well-planned study designs that the unique footprints of alternative judgment models can be discerned in the data.

a. A Psychological Model of Judgment

The central theory of judgment underlying the STF method is outlined in Figure IV-2. The outline depicts that data are internalized, combined to form a perception or "impression," and then "externalized" in the form of a response.

The outline suggests that three kinds of psychological processes mediate between data and response. The respondent represents data, whether explicitly numerical or not, as *internal subjective values*. The respondent then combines the subjective values to form an integrated perception, or "impression," which [still] is represented internally as a subjective value. Finally, the respondent "externalizes" the internal impression into a response that conforms to the requested response format (i.e., yes/no, numerical rating, a verbal description, etc.). We expand on model stages below.

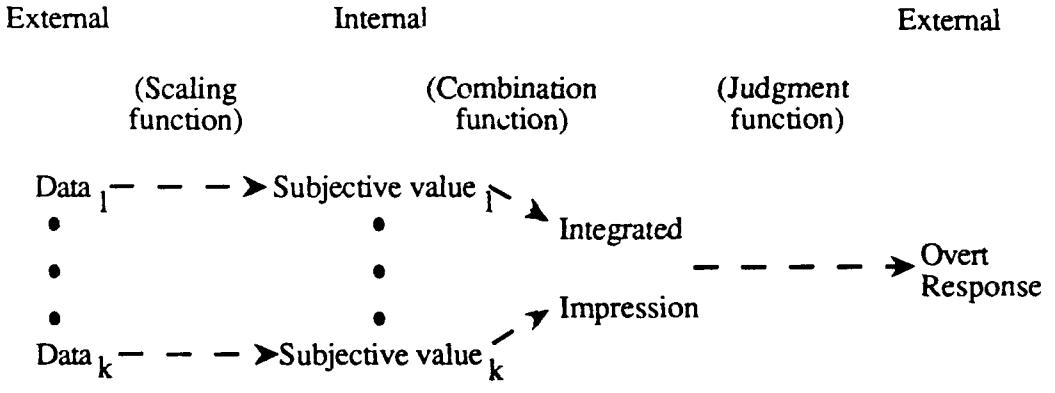


Figure IV-2. Transformation of Data to Judgments

According to the conception of psychological judgment processes depicted in Figure IV-2, respondents first transform data, whether numerical or non-numerical, into a subjective scale representation. The transformation is not necessarily a conscious or overt process and the subjective scale values should not be thought of as being numbers; rather, subjective scale values may be thought of as psychological impressions or as perceptions of magnitude.

b. Combination Rules (Functions)

Without overtly using mathematical formulas, people nonetheless are able to combine magnitude information to form an overall impression or response in ways that researchers have been able to model with arithmetic functions (Birnbaum, 1974a, Veit and Callero, 1982d). The functions vary from simple additive and multiplicative functions to more complex weighting functions. Veit, Callero, and Rose (1984) display four such functions which they have found to be useful in their analyses. While there are an infinite number of rules that could be used to model respondents' combination processes, only a small number have been found to be necessary in practice.

c. Judgment Function

Just as subjective scale values are impression-like rather than overtly numerical in the outline of the judgment process, so too is the product of combining or comparing subjective scale information. Therefore, in order to make a numerical response, the transformation from internal impression to external numerical form must occur in the model.

d. Implications of Psychological Processes: "Ratios" vs. "Differences"

When asking a judge to form a ratio of two quantities -- Judge how many times stronger A is than B?[3] -- the forgoing discussion indicates that we must be interested in the processes people go through when instructed to make judgments of this sort. Asking for this ratio directly as part of the judgment would appear to be the most sensible way to get that information. However, Veit and Birnbaum (Veit, 1978, Birnbaum, 1978, 1981) argue that when instructed to make ratio comparisons of magnitudes, respondents take differences and instead "exponentially transform the difference" of the subjective scale values (i.e., $R_{ab}=e^{a-b} = e^a/e^b$) so that the responses obey the predictions of a ratio model of the judgment process. Responses, however, would not represent the ratio of the respondents' subjective scale values, but rather the ratio of exponentially transformed subjective differences. That is, from the response we should not infer that the respondents' subjective value for *a* is (e^a/e^b) times that of *b*. The assumption that respondents follow

---

[3]Let us denote "objective" data with upper case letters, subjective evaluations of these data in lower case letters, and the response comparing the ratio of subjective values *a* and *b* as R_{ab}.

task instructions to judge ratios implies a model with a ratio combination process and leads us to this erroneous conclusion.

Birnbaum argues further that even if the judgments actually made were ratios, interpreting the results as the ratios of the respondents' subjective evaluations of the data still may not be valid.

In judging the ratios of subjective scale values for *A*, *B*, and *C*, Birnbaum argues that respondents report $R_{ab}=a^n/b^n$, $R_{bc}=b^n/c^n$, and $R_{ac}=a^n/c^n$, where n is an arbitrary, real constant. The ratios stated are consistent amongst each other as a group ($R_{ab}=(R_{cb})(R_{ac})$) supporting an assumption of ratio comparative judgments. (However, except when the exponent has the value of 1, the ratios do not correspond to the ratios of the underlying subjective scale values.) Yet it is the underlying scale values that we will be interested in rather than the exponentiated transforms. Simply asking the respondent to make ratio judgments does not permit the determination of the value of n and the subjective value of *A* is not generally a^n/b^n times that of *B*. Thus, the subjective scale values derivable from these ratio comparisons, and the ratio comparisons themselves, are unique only up to a power transformation (or a log-linear transformation) of the underlying scale values and "true" ratios.

Veit, Callero, and Rose argue that the STF methodology of collecting judgment data with carefully thought out study designs allows the analyst to determine the appropriate model of the judgment data and thereby the properties of the resulting numerical scale.

2. Theoretical Background

The reasoning behind Veit and Callero's arguments comes from algebraic modelling approach to the measurement of subjective judgment. We provide some theoretical background to this area in Appendix A with material from Veit's paper (1978) "Ratio and Subtractive Processes in Psychophysical Judgment." Other helpful references, especially Anderson (1974a, 1974b, 1979, 1981) and Birnbaum (1981), are provided in the "Supplementary Techniques Bibliography." provided in this Chapter.

3. Steps of the STF Method: Function Assessment

Subjective transfer functions are assessed for each experimental unit using the respondents' estimates of the outcomes given specified levels of the factors. Given these data, the STF methodology allows identification of 1) subjective values for the "objective" questionnaire input data (further analyses of these subjective values may permit the

identification of systematic relationships between "objective" input and subjective values); 2) combination functions for modelling how respondents create an overall perception or "impression" of the data in response to the requirements of the task given them; and 3) determination of information about the function that transforms the internal "impression" into an external response.

Having discussed the procedures involved in the STF methodology and some of its underlying theory (Figure IV-2), we can expand upon the role of varied questionnaire designs that we touched on above.

In their several papers, Veit and Callero illustrate interactive and noninteractive combination functions for explaining how people integrate subjective scale values to internal perceptions. Six functions are illustrated in Table IV-2, three each of noninteractive and interactive types. In the Table, all equations are specified for three factors: A, B, and C. The s_{ij} variables represent the subjective values for the j^{th} level of the i^{th} factor (i.e., s_{A1} represents the subjective value of the first level of factor A, where factor A might be currency of weather information and the first level is 1 hour. [See Figure IV-1]); the w variables represent the weights associated with the subscripted factor. The "r" variables are the internal perceptions or "impressions" that result from combining the subjective values; the "a" term is a constant for the additive and multiplicative models. The ω parameter is the weight of the range term s_{MAX}-s_{MIN}, which is simply the difference between the largest and smallest subjective value placed on information presented in a questionnaire item. Finally, w_o and s_o correspond to indexes subjectives and weights on them corresponding to "initial impressions." Initial impressions are "what the response would be in the absence of specific information" (Veit, Callero, and Rose, 1984, pp. 16) stated in the questionnaire items.

Table IV-2. Possible Subjective Transfer Functions

A. Noninteractive functions	
$r = s_{A_i} + s_{B_j} + s_{C_k} + a$	Additive
$r = \frac{w_A s_{A_i} + w_B s_{B_j} + w_C s_{C_k}}{w_A + w_B + w_C}$	Averaging
$r = \frac{w_O s_O + w_A s_{A_i} + w_B s_{B_j} + w_C s_{C_k}}{w_O + w_A + w_B + w_C}$	Relative-weight (averaging with initial impression)
B. Interactive functions	
$r = s_{A_i} s_{B_j} s_{C_k} + a$	Multiplicative
$r = \frac{w_O s_O + w_A s_{A_i} + w_B s_{B_j} + w_C s_{C_k}}{w_O + w_A + w_B + w_C} + \omega\left(s_{MAX} - s_{MIN}\right)$	Range
$r = \frac{w_O s_O + w_{A_i} s_{A_i} + w_{B_j} s_{B_j} + w_{C_k} s_{C_k}}{w_O + w_{A_i} + w_{B_j} + w_{C_k}}$	Differential-weight

Source: Adapted from Veit, Callero, and Rose (1984)

Each model makes a number of predictions about the pattern of judgments that result when the factors are manipulated across levels in factorial designs. The plots of predicted judgments as a function of factor levels and scenario designs (e.g., the number and identity of factors simultaneously manipulated in scenario questions) depict footprints

against which corresponding plots of judgment data can be compared. Critical attributes of the judgment data plots are then diagnostic of one or more combination models which could have generated them. For instance, judgment responses can be plotted as a function of two factors, one factor on the abscissa (S_{Ai}) and one factor (S_{Bj}) as a family of separate curves for each level of the factor. If an additive model is the appropriate way to relate responses to scale values for the levels of these factors, $r = S_{Ai} + S_{Bj}$, then the separate curves for r as a function S_{Bj} will be parallel. More generally, all three of the noninteractive functions displayed in Figure IV-2 will produce parallel curves when predicted judgments are plotted against the levels of two or more factors.[4] Thus parallelism is only partially diagnostic of the most appropriate judgment model for the data. Additional graphical tests are required to tease apart the remaining hypothesized models further; these tests are only possible because the factional experimental designs for contracting scenario questions make the pertinent judgment data available.[5]

Veit et al. (1984) describe several other graphical footprints which are diagnostic of alternative combination functions (pp. 20, 24-26). Veit et al. (1982) demonstrate the diagnosis of judgment models using graphical footprints.

4. Steps of The STF method: Analysis

Once having collected the judgment data and having assessed information about the underlying psychological judgment processes that produced it, the final "step" in the STF methodology is the estimation of the unknowns of the combination functions. The data for this analysis consist of the responses to the questionnaire items, all other variables displayed in the models of Figure IV-2 are unknown coefficients to be estimated. These include the additive constants for the additive and multiplicative models, the factor weights, the subjective values and weights for the initial impressions, the range factor weight, and the subjective values themselves. Thus, all data regarding the covert, psychological processes underlying the observed responses proceed from the combination function selected to account for the judgment data. Veit, Callero, and Rose (1984) used a technique for estimating the unknown values by minimizing the least square error between the observed data and that which would be predicted by the fitted model. The level of

---

[4]The three interactive combination functions yield corresponding plots which converge or diverge in a fan-like pattern.

[5]More complex study designs allow for hypotheses to be tested about the judgment function, which "externalizes" the integrated impression into a response (Veit et al., 1984, pp. 27-30).

"badness of fit," that is, the sum of the squared error remaining after fitting the model, is then used as a diagnostic to compare alternative hierarchical models of the system under study.

5. Steps of The STF method: Additional Considerations

Veit, Callero, and Rose (1984) discuss strategies for selecting experimental designs. This represents an important area of practical concern because a fully factorial design for an experimental unit with several factors at several levels might yield an impractically large questionnaire. In this regard, pilot studies to collect preliminary information are essential. For the full study, it is possible to avoid implementing a fully crossed factorial design; instead, the study could be designed to provide *sufficiently* strong assessments of the functions under consideration. Thus, if two functions make identical predictions for a subset of the complete set of questions, then these questions become candidates for not being included in the final study. Results from pilot studies also can contribute to designing the study inasmuch as useful information is collected about the effects of factors and how they interact. Veit, Callero, and Rose (1984) discuss these strategies in some greater detail in their paper.

Finally, Veit and Callero have discussed a number of other subsidiary issues that are important to understand in order to be able to implement an STF study. Among them are calibrating initial input scale values, and 2) defining alternative structural models of the system. These are discussed at some length in Veit and Callero (1981a).

BIBLIOGRAPHY

N.H. Anderson. "Algebraic Models in Perception," in E.C. Carterette and M.P. Friedman (Eds.), "Handbook of Perception" (Vol. 2). New York: Academic Press, 1974a.

N.H. Anderson. "Cognitive Algebra," in L. Berkowitz (Ed.), "Advances in Experimental Social Psychology" (Vol. 7). New York: Academic Press, 1974b.

M.H. Birnbaum. "The Nonadditivity of Personality Impressions," *Journal of Experimental Psychology*, 102 (monograph), 1974a, pp 543-561.

M.H. Birnbaum, C.T. Veit. "Scale Convergence as a Criterion for Rescaling: Information Integration with Difference, Ratio, and Averaging Tasks," *Perception and Psychophysics*, 15, 1974a, pp. 7-15.

M.H. Birnbaum, C.T. Veit. "Scale-free Tests of an Additive Model for the Size-Weight Illusion," *Perception and Psychophysics*, 16, 1974b, pp 276-282.

C.T. Veit. "Ratio and Subtractive Processes in Psychophysical Judgment," *Journal of Experimental Psychology: General*, 107 (1), 1978, pp 81-107.

N.H.Anderson. "Algebraic Rules in Psychological Measurement," *American Scientist*, Vol. 67, 1979, pp.. 555-563.

N.H. Anderson. "Foundations of Information Integration Theory," New York: Academic Press, 1981.

M.H. Birnbaum. "Controversies in Psychological Measurement," in B. Wegener (ed.), "Social Attitudes and Psychophysical Measurement," Hillsdale, N.J.: Erlbaum, 1981.

C.T. Veit and M. Callero. *Subjective Transfer Function Approach to Complex System Analysis*, The Rand Corporation, R-2719-AF, March 1981a.

M. Callero. W. Naslund, C.T. Veit. *Subjective Measurement of Tactical Air Command and Control* -- Vol. I: *Background and Approach*, The Rand Corporation, N-1671/1-AF,.March 1981b.

M. Callero, W. Naslund, C.T. Veit. *Subjective Measurement of Tactical Air Command and Control* -- Vol. II: *The Initial Representation*, The Rand Corporation, N-1671/2-AF, March 1981c.

C.T. Veit, B.J. Rose, M. Callero. *Subjective Measurement of Tactical Air Command and Control* -- Vol. III: *Preliminary Investigation of Enemy Information Components*, The Rand Corporation, N-1671/3-AF, March 1981d.

C.T. Veit, M. Callero, B.J. Rose. *Demonstration of the Subjective Transfer Function Approach Applied to Air-Force-Wide Mission Area Analysis*, The Rand Corporation, R-1831-AF, February 1982.

C.T. Veit, B.J. Rose, J.E. Ware. "Effects of Physical and Mental Health on Health-state Preferences," *Medical Care*, 20 (4), 1982, pp 386-401.

C.T. Veit, M. Callero. *The Subjective Transfer Function Approach For Analyzing Systems*, The Rand Corporation, P-6893, June 1983.

C.T. Veit, M. Callero, B.J. Rose. *Introduction to the Subjective Transfer Function Approach to Analyzing Systems*, The Rand Corporation, R-3021-AF, July 1984.

B. CRITICISMS, CAVEATS, AND REPLIES

There have been no published critiques of the subjective transfer function method.

C. METHODOLOGICAL VARIANTS

There are no known methodological variants to the subjective transfer function method.

D. BIBLIOGRAPHY (BACKGROUND LITERATURE)

E.C. Poulton. "The New Psychophysics: Six Models for Magnitude Estimation," *Psychological Bulletin*, 69, 1968, pp 1-19.

D.H. Krantz, R.D. Luce, P. Suppes, A. Tversky. "Foundations of Measurement," New York: Academic Press, 1971.

S.S. Stevens. "Issues in Psychophysical Measurement," *Psychological Review*, 78, 1971, pp 426-450.

N.H. Anderson. "Algebraic Models in Perception," in E.C. Carterette and M.P. Friedman (Eds.), "Handbook of Perception" (Vol. 2). New York: Academic Press, 1974a.

N.H. Anderson. "Cognitive Algebra," in L. Berkowitz (Ed.), "Advances in Experimental Social Psychology" (Vol. 7). New York: Academic Press, 1974b.

M.H. Birnbaum. "The Nonadditivity of Personality Impressions," *Journal of Experimental Psychology*, 102 (monograph), 1974a, pp 543-561.

M.H. Birnbaum, C.T. Veit. "Scale Convergence as a Criterion for Rescaling: Information Integration with Difference, Ratio, and Averaging Tasks," *Perception and Psychophysics*, 15, 1974a, pp 7-15.

M.H. Birnbaum, C.T. Veit. "Scale-free Tests of an Additive Model for the Size-Weight Illusion," *Perception and Psychophysics*, 16, 1974b, pp 276-282.

R.N. Shepard. "On the Status of 'Direct' Psychological Measurement," in Savage (ed.), "Minnesota Studies in the Philosophy of Science," Vol. IX, Minneapolis: University of Minnesota Press, 1976.

C.T. Veit. "Ratio and Subtractive Processes in Psychophysical Judgment," *Journal of Experimental Psychology: General*, 107 (1), 1978, pp 81-107.

M.H. Birnbaum. "Differences and Ratios in Psychological Measurement," in N.J. Castellan, F. Restle (eds.), "Cognitive Theory" (Vol 2). Hillsdale, N.J.: Erlbaum, 1978.

N.H.Anderson. "Algebraic Rules in Psychological Measurement," *American Scientist*, Vol. 67, 1979, pp.. 555-563.

N.H. Anderson. "Foundations of Information Integration Theory," New York: Academic Press, 1981.

M.H. Birnbaum. "Controversies in Psychological Measurement," in B. Wegener (ed.), "Social Attitudes and Psychophysical Measurement," Hillsdale, N.J.: Erlbaum, 1981.

E.C. Poulton. "Bias in Quantifying Judgments," Hillsdale, N.J.: Erlbaum, 1989.

E. BIBLIOGRAPHY (REPORTED APPLICATIONS TO DEFENSE PROBLEMS)

C.T. Veit and M. Callero. *Subjective Transfer Function Approach to Complex System Analysis*, The Rand Corporation, R-2719-AF,.March 1981a.

M. Callero, W. Naslund, C.T. Veit. *Subjective Measurement of Tactical Air Command and Control* -- Vol. I: *Background and Approach*, The Rand Corporation, N-1671/1-AF,.March 1981b.

M. Callero, W. Naslund, C.T. Veit. *Subjective Measurement of Tactical Air Command and Control* -- Vol. II: *The Initial Representation*, The Rand Corporation, N-1671/2-AF, March 1981c.

C.T. Veit, B.J. Rose, M. Callero. *Subjective Measurement of Tactical Air Command and Control* -- Vol. III: *Preliminary Investigation of Enemy Information Components*, The Rand Corporation, N-1671/3-AF, March 1981d.

C.T. Veit, M. Callero, B.J. Rose. *Demonstration of the Subjective Transfer Function Approach Applied to Air-Force-Wide Mission Area Analysis*, The Rand Corporation, R-1831-AF, February 1982a.

C.T. Veit, M. Callero. *The Subjective Transfer Function Approach For Analyzing Systems*, The Rand Corporation, P-6893, June 1983.

C.T. Veit, M. Callero, B.J. Rose. *Introduction to the Subjective Transfer Function Approach to Analyzing Systems*, The Rand Corporation, R-3021-AF, July 1984.

F. APPLICATION SOFTWARE

Software to implement aspects of the STF currently is under development at The Rand Corporation. The interested reader should contact Dr. Clairice Veit, at the Rand Corporation, for additional information.

G. EVALUATION AND COMMENTS

The STF is an interesting methodology that has not yet received the kind of critical attention necessary to evaluate its theoretical or practical value. On its face, it seems to require more resources to implement than do many other methods. However, some of these methods have not been applied with the level of effort required to ensure satisfaction of underlying assumptions, and therefore accurate results. Thus, the difference in ease of application *may* be illusory. Furthermore, as a straightforward methodological development of psychological measurement theory, the results of an STF study *may* have a firmer theoretical foundation than do those resulting from comparable methods (i.e., utility assessment, the Analytic Hierarchy Process).

We stress the word *may* because, while Veit and Callero make compelling arguments in favor of the STF, the method has not been rigorously compared with other

comparable methods. This situation may exist for several reasons. First, reports describing the method and its application have been limited to Rand Corporation research reports, which has limited the audience to whom the work has been exposed. Second, the Rand research reports on the STF may be inaccessible to many because the authors describe their work in language and concepts unfamiliar to most readers. Existing documents and tutorials on the STF may not even adequately motivate the value of the method in language accessible to the layman. Finally, the STF lacks a "handbook" that thoroughly documents the method. Such documentation would provide a comprehensive treatment of relevant concepts of experimental design, pretest design strategy, graphical model assessment, model fitting, and the statistical assessment of models for those without a sufficient statistical background. It also would report a case study in enough detail to allow the interested practitioner to learn the method by example.

We generally are not in favor of practioners applying methods "in cookbook fashion" without a good understanding of the underlying theory and assumptions. However, detailing the procedures of the method as steps to be followed in a well-defined algorithm may go a long way toward introducing the methodology to the analytical community.

Apart from its methodological value, the work on the STF is important in that it raises important issues of psychological measurement that some subjective judgment methods have not made contact with. Thus, beyond the question of whether STF itself should be widely adopted by the analytical community, the issues that motivated its development need to be thoroughly discussed by those interested in improving the quality of analyses dependent on subjective judgment data.

In practical terms, we suggest that steps be taken to encourage existing practitioners of the STF to submit their papers for publication to widely read journals (e.g., *Operations Research, Management Science, Interfaces, Naval Logistics Research*) and to present their work at forums widely attended by members of the analytical community.

V. UTILITY THEORY

This chapter presents an examination of the utility theory. It is organized into five major subheadings. Each of the five papers presented in this study follows the same general outline for ease of reader reference and comparison. Before beginning our detailed discussion on utility theory, we offer a brief explanation of its origins.

Utility theory and its extensions are often traced back to the axiomatized theory of utility articulated by Von Neumann and Morgenstern.[1] By assuming that an individual could state his preferences between lotteries of different alternatives, and that these preferences satisfied certain axioms,[2] Von Neumann and Morgenstern showed that a numerical function existed that mapped alternatives into real numbers and that was unique up to a linear transformation. Other authors posited utility axioms somewhat different from those of Von Neumann and Morgenstern,[3] but the objective was the same -- to assign consistent (according to the axioms) numerical values to human preferences. Once determined, utility functions could be used in a variety of mathematical constructs, such as games (as defined by Von Neumann and Morgenstern, and others since) or optimizations. Obvious continuous analogues of this discrete structure have been developed.

Multiattribute utility theory employs vectors of utility functions, each component of which corresponds to an "attribute" of the problem in hand. These vectors then can be used to create an overarching objective function or to identify non-dominated (Pareto-optimal) sets of alternatives.[4]

---

[1] Von Neumann, John, and Oskar Morgenstern, "Theory of Games and Economic Behavior," New York: John Wiley & Sons, 1944.

[2] The original axioms of Von Neumann and Morgenstern were three. The first stated that the set of alternatives under consideration is completely ordered. The second dealt with ordering and combining lotteries. For example, if v is preferred to u, then the certainty of v is preferred to any lottery of v and u, and if u is preferred to w, which is preferred to v, then there is a lottery involving u and v which is preferred to the certainty of w. The third axiom dealt with the algebra of combining lotteries. Many basic references discuss the Von Neumann and Morganstern axioms and alternatives to them.

[3] See the discussion of utility in Luce, R. Duncan and Howard Raiffa, "Games and Decisions," New York: John Wiley & Sons, 1957.

[4] Keeney, Ralph L. and Howard Raiffa, "Decisions with Multiple Objectives: Preferences and Value Tradeoffs," New York: John Wiley & Sons, 1976.

Assuming that utility functions can be determined has proven very useful for application to economics, decisionmaking, mathematical psychology, and other disciplines. Our interest, however, focuses on the development of the functions themselves, i.e., going from subjective preferences to numerical measures.

A. DISCUSSION AND IMPLEMENTATION

Utility functions generally are evaluated by presenting an individual a set of lotteries; components of these lotteries are adjusted until some terminal criterion is reached. Farquhar[5] provides an excellent overview of a number of approaches, as well as a description of the steps involved in a utility assessment activity. In one approach, the probabilities describing the lotteries are varied until indifference between outcomes is reached. Consider the following small example.

Suppose that for some situation there are four alternatives: a_1, a_2, a_3, a_4, with a_1 being a least preferred alternative, a_4 being a most preferred alternative, and the remaining alternatives ranked in preference order (the axioms of utility theory require that such a preference ordering is possible). Denote the corresponding utility values by u_1, u_2, u_3, and u_4. One can arbitrarily assign $u_1=0$ and $u_4=1$. The individual whose utility is being assessed then is asked to choose between the certainty of alternative a_2 and the expected outcome of a lottery among, say, a_1 and a_3 with probabilities p and (1-p). The individual chooses a value for p for which he is indifferent between these two outcomes. This implies $u_2=pu_1+(1-p)u_3$. One similarly could elicit a probability q so that the individual is indifferent between a_3 and a lottery between a_2 and a_4. This probability q would satisfy $u_3=qu_2+(1-q)u_4$ (other lotteries among the alternatives are possible of course -- the axioms ensure that the results will be the same no matter what lotteries are used). Assuming, as we have, the values of u_1 and u_4, we can obtain $u_2=(1-p)(1-q)/(1-q+pq)$ and $u_3=(1-q)/(1-q-pq)$. This method sometimes is called the "direct assessment" method.[6] Keeney recommends that where there are many alternatives with a natural underlying ordering, applying direct assessment on a subset of the alternatives and interpolating for those remaining sometimes may be preferred.

Of course, one can vary more than the probabilities. Farquhar provides a taxonomy. Let [x,a,y] denote the lottery with outcome x with probability a and outcome y

---

[5] Farquhar, Peter H., "Utility Assessment Methods," Management Science, Vol. 30, No. 11, November, 1984.

[6] Keeney, op. cit.

with probability (1-a). Let R be the relation < (less preferred than), ~ (indifferent to), or > (more preferred than). Then Farquhar lists seven possible approaches to evaluating utilities based on what the individual whose utility is being evaluated specifies. These are divided into three categories, as follows. The underlined element indicates the judgments specified by the individual.

1. Standard Gamble Methods

Standard gamble methods compare lotteries to certainties, as was done in our example above. First, "preference comparisons" are of the form [x,a,y] $\underline{R}$ w, where the individual specifies the relationships between lotteries and certainties. Second, "probability equivalences" are of the form [x,$\underline{a}$,y] ~ w, where w is ranked between x and y. Our example was a probability equivalence approach. "Value equivalences" are of the form [$\underline{x}$,a,y] ~ w, and "certainty equivalences" are of the form [x,a,y] ~ $\underline{w}$.

2. Paired Gamble Methods

Paired gamble approaches compare lotteries to lotteries, but parallel the standard gamble methods. "Preference comparisons" are of the form [x,a,y] $\underline{R}$ [w,b,z]. "Probability equivalences" are of the form [x,$\underline{a}$,y] ~ [w,b,z], where both w and z are between x and y; "value equivalences" are of the form [$\underline{x}$,a,y] ~ [w,b,z].

Farquhar discusses details of each of these approaches and the advantages and disadvantages inherent in each. Variations exist, but Farquhar's taxonomy is comprehensive. An approach described by Keeney is a form of "certainty equivalence" approach. This procedure, intended for continuous options over a bounded region, requires that one first choose least preferred and most preferred options, call them a_0 and a_1, and then attempt to find that option in the range, denoted $a_{.5}$, for which the decisionmaker is indifferent between the certainty of $a_{.5}$ and a lottery consisting of a_0 with probability 0.5 and a_1 with probability 0.5. The corresponding utility values are $u(a_0)=0$ and $u(a_1)=1$, arbitrarily, and $u(a_{.5})=.5$. One then proceeds to identify by the same type interval halving technique $a_{.25}$, $a_{.75}$, and so on. One then can fit a function through these points to generate a utility function. The choice of function will depend to some extent on whether the decisionmaker is risk averse, risk neutral, or risk prone. Generally speaking, a decisionmaker who is risk averse will tend to prefer certainties to lotteries with the same expected outcome, and his utility function will be concave. A decisionmaker who is risk prone will prefer lotteries to certainties, and his utility function will be convex. A

decisionmaker who is risk neutral will be indifferent between certainties and lotteries with the same expected outcome, and his utility function will be linear.

The overall approach can be as structured in a variety of ways. Keeney and Raiffa[7] offer five formal steps: structuring the decision problem; identification of characteristics such as continuity and risk properties (concavity); quantitative evaluation (using any of the procedures outlined above); selection of a utility function fitting the quantitative results; and evaluation of consistency.

3. Non-Gamble Methods

There have been a number of approaches proposed for specifying utility functions without employing lotteries. For example, one can *assume* a strictly linear utility function for a particular characteristic, assign a utility of 0 to the minimum plausible value that characteristic can take on, and assign a utility of 1 to the maximum plausible value that characteristic can take on[8]. This, of course, assumes that the characteristic can take on numerical values. In a similar fashion, one can assume *a priori* some other functional form, such as exponential, and use minimum and maximum plausible values to determine parameter values. Some practitioners have even proposed allowing the decisionmaker to draw freehand functions to be used in utility evaluations.

While such approaches avoid lotteries, it is not evident that the results are true utility functions. Assuming a functional form without consulting with the decisionmaker may ignore crucial attitudes toward risk that the decisionmaker may subconsciously hold. For similar reasons, allowing a decisionmaker to simply create a function may not satisfy the underlying axioms needed to justify the use of utility functions, thereby rendering subsequent analyses using that function suspect on theoretical grounds.

B. CRITICISMS, CAVEATS, REPLIES

The attraction of utility theory is that it allows the economist, the game theorist, the operations researcher and other quantitative analysts to presume that utility functions exist and to develop theory based on that presumption. In some cases, monetary attributes of a particular problem or other preexisting quantitative measure provide a good proxy for utility. However, when this is not the case, utility functions must be determined, as

[7] Keeney, Ralph L. and Howard Raiffa, "Decisions with Multiple Objectives: Preferences and Value Tradeoffs," New York: John Wiley & Sons, 1976.

sketched above, by having individuals declare preferences among lotteries. Not only is this cumbersome, but we find that the results of any such exercise depend strongly on how the lotteries are presented, i.e., individuals will give different responses to essentially the same questions depending on how the questions are phrased. For example, experiments have shown that if medical treatment options are presented in which outcomes are expressed in terms of how many people will be saved, measured utilities will be different from the identical options for which outcomes are expressed in terms of how many people will die[9]. This type of difficulty, which is termed "inconsistency," requires extreme care in presenting lotteries. Inconsistency is an example of a "context effect," which has been defined as "influences on preferences that are without normative basis."[10] Hershey, Kunreuther and Schoemaker cite a number of experiments that demonstrate problems in which identical lotteries are interpreted differently. In one example, a 1 in 100 chance of losing $1,000 is compared to the certainty of losing $10. When this was presented to experimental subjects, 56 percent preferred the certain loss. When the certainty was phrased as "insurance" to protect against the loss of $1,000 in the lottery, an identical situation in terms of expected value, 81 percent of the subjects preferred the "insurance" option.

Even if context effects could be eliminated, and it is not clear that this is possible,[11] problems remain. Individuals may have a difficult time determining when they are indifferent to lotteries. As Shafer[12] points out, there is a difference between indifference, wherein one knows one is equally happy with two sets of outcomes, and indecision, in which one cannot state preference orderings. Often, further information will help to eliminate indecision, but that is not always the case.

---

[8] Examples of this sort of utility function evaluation are provided in Edwards, Ward, and J. Robert Newman, *Multiattribute Evaluation*, Sage University Paper 26, Sage Publications, Beverly Hills, 1982.

[9] See, for example, Shafer, Glenn, "Savage Revisited" in Bell, David E., Howard Raiffa, and Amos Tversky (eds.), "Decision Making: Descriptive, Normative, and Prescriptive Interactions," Cambridge, MA: Cambridge University Press, 1988.

[10] Hershey, John C., Howard C. Kunreuther and Paul J. H. Schoemaker, "Sources of Bias in Assessment Procedures for Utility Functions," in Bell, Raiffa, and Tversky, op. cit.

[11] We may consider that judgments are always made in some context. However, it has been argued (C.T. Veit, personal communication) that systematically manipulating contexts provides a more complete description of the bases for judgments and may lead to identification of invariant scales (e.g., utility functions) underlying context-varying judgments (see M.H. Birnbaum and S.E. Sutton. "Scale Convergence and Decision Making," *Organizational Behavior and Human Decision Processes*, in press.

[12] Shafer, ibid.

Although utility axioms may seem *a priori* intuitive, they in fact often violate empirical evidence. For example, the Von Neumann-Morgenstern axioms imply that preferences are transitive: if a is preferred to b, and b is preferred to c, then a is preferred to c. This is sometimes not the case, and a variety of examples have been introduced into the literature over the past twenty or thirty years verifying this. One example[13] presents a job seeker with three offers, with characteristics and rankings as follows:

	Salary	**Location**	**Work Quality**
x:	excellent	satisfactory	good
y:	good	excellent	satisfactory
z:	satisfactory	good	excellent

If the job seeker compares any two offers, and makes his evaluation on the basis of pairwise comparisons wherein he prefers one job to another if the former surpasses the latter in two or more characteristics, it is clear that he prefers x to y (since x dominates y with respect to two characteristics), y to z, and z to x. Each pairwise comparison seems rational, but the result is a set of intransitive comparisons.

Some researchers, in response to such problems, have proposed alternative axioms that relax transitivity requirements.

While distinctions among axiomatic structures are important for the theories that are built upon them, the basic difficulties inherent in converting preferences into utility functions through the use of lottery comparisons remain.

The difficulties of application that are cited above do not necessarily rule out the use of utility theory for specific applications. Indeed, the use of an approach that has sound theoretical underpinnings, even if difficult, often is to be preferred to an *ad hoc* approach which, while more easily implemented, may have inherent characteristics that are not well understood. For example, a common approach to comparing alternatives with multiple attributes that employs subjective scores for each attribute of each alternative and weight by which to compare attributes to each other can be viewed as a degenerate application of multiattribute utility theory,[14] wherein it is implicitly assumed that the decisionmaker's utility function is linear and hence, risk neutral. This may or may not be the case. If not,

---

[13] From Fishburn, P. C., "Normative Theories of Decisionmaking Under Risk and Under Uncertainty, in Bell, Raiffa, and Tversky, *op. cit.*

[14] Bunn, Derek W., "Applied Decision Analysis," New York: McGraw-Hill, 1984.

then the simplified approach is not really justified. Orthodox approaches to determining utility functions, while cumbersome, would provide information concerning the risk attitudes of the decisionmaker and therefore might produce more correct results.

C. VARIANTS AND EXTENSIONS

The best known extension of utility theory is "multiattribute utility theory," which recognizes that many decision problems comprise many issues (or attributes), all of which must be assessed in evaluating options. It is no small matter to define the attributes of any given problem, and the literature discusses useful characteristics of multiple attributes. Among these are[15] *completeness*: if the set of attributes is adequate to describe the problem (it is also desirable to keep the number of attributes to the smallest number possible that satisfies this and other characteristics); *operationality*: if each attribute is meaningful to the decisionmaker and to those to whom the decision will have to be explained; *decomposability*: a property that allows reduction of the preference quantification from one multidimensional task to a number of tasks of smaller dimensionality; and *nonredundancy*: elimination of attributes that are subsumed by other attributes. Once attributes are identified that (to the extent possible) satisfy the above desiderata, the task of developing the multiattribute utility function must be faced.

Let us introduce some notation. Suppose, for a given problem, n attributes have been defined. Let X_1 denote the set (possibly infinite) of options for the first attribute, X_2 the set of options for the second attribute, and so on. Let x_i denote a particular option within X_i. It is necessary, obviously, to develop a multiattribute utility function $u(x_1,x_2,...,x_n)$. One can imagine the difficulties involved in presenting a decisionmaker with multiattribute lotteries among which to chose. Therefore, it would simplify matters greatly if the multiattribute utility function could be reduced to unidimensional utility functions; i.e., if one could express $u(x_1,x_2,...,x_n)$ as $f(u_1(x_1),u_2(x_2),...,u_n(x_n))$ where each u_i was a utility function applying to attribute i. If this were the case, each u_i could be evaluated using simple lottery comparisons. Of course, all the problems involved in lottery comparisons for unidimensional utility theory would be compounded n times.

---

[15] These characteristics have been drawn from Keeney and Raiffa, op. cit.

In fact, conditions have been found that allow reduction of multiattribute problems in this manner. Under various conditions of independence[16] it can be shown that

$u(x_1,...,x_n)=\sum_i k_i u_i(x_i)$ (the "linear multiattribute utility function"), or

$1+ku(x_1,...,x_n)=\prod_i [1+kk_i u_i(x_i)]$ (the "multiplicative multiattribute utility function"), or

$u(x_1,...,x_n)=\sum_i k_i u_i(x_i)+\sum_i \sum_j k_{ij} u_i(x_i) u_j(x_j)$ (the "multilinear multiattribute utility function"). Once the unidimensional utility functions have been evaluated, the scaling constants k_i and k_{ij} can be determined by asking the decisionmaker to identify indifferences among attributes. Enough such identifications produce systems of equations that can be solved for the k_i and the k_{ij}. The constant k in the second equation is determined from the other k_i.

Much of the literature in multiattribute utility functions focuses on one of the above multiattribute utility function forms. There is very little in the literature on multiattribute utility functions that are not separable into unidimensional utility functions.

Computer programs have been developed over time to assist with the construction of utility functions. One such, MUFCAP[17] (Multiattribute Utility Function Calculation and Assessment Package), allows a user to construct a multiattribute utility function of the additive or multiplicative types described above. The component unidimensional utility functions are assumed to be adequately approximated by linear, exponential, or piecewise-linear functions. The various utilities of this package allow one to specify attributes and to calculate utility functions and scaling factors through the comparison of lotteries. Similar computer programs have been developed for research purposes. However, the number of commercial software packages for direct application of utility theory is limited.

One such package is "Logical Decision,"[18] a program that allows a user to structure a decision problem, develop single utility functions and multiple attribute tradeoffs, rank

---

[16] The two most prominent concepts are *preferential utility*, wherein preference among sets of attributes are independent of the remaining attributes; and *utility independence*, wherein preferences among lotteries for any single attribute are independent of the options chosen from all other attributes. See Keeney and Raiffa, op. cit., for more complete and more rigorous definitions.

[17] Keeney, Ralph L., and Alan Sicherman, "Assessing and Analyzing Preferences Concerning Multiple Objectives: An Interactive Computer Program," *Behavioral Science*, Vol. 21 (1976), pp. 173-182.

[18] Copyright 1989, Logical Decisions, Point Richmond, CA.

options and perform sensitivity analyses. The package is available for IBM-PC and compatible computers and appears straightforward to use. No applications of either of these software packages to military applications is known, however.

D. APPLICATIONS

The notion of quantifiable utility is attractive since it allows the development of theories that would be much less robust without it. However, the actual exercise of evaluating utility functions is not frequently done. For example, game theory, whose roots are closely linked with the origins of utility theory, depends crucially on the availability of utility functions to quantify preferences to the outcomes of instances of conflict and cooperation. As a general rule, game theorists do very little evaluation of utility functions. Instead, they substitute quantifiable attributes (such as dollar gains) for utility functions or provide illustrative examples whose numerical inputs have not been derived from the preferences of actual decisionmakers. While this can result in valuable theory, the fact remains that straightforward military applications of utility theory itself are relatively scarce. Most of the applications that do exist deal with civilian economic issues (a classic is a case study done to identify alternatives for development of the Mexico City airport[19]). Military applications are hard to find. A search of the Defense Technical Information Center database using the key phrases "utility theory" and "utility function" turned up fewer than twenty citations, of which perhaps half relate to military issues; in general, however, those tended to be ones done at military graduate schools that discussed utility theory as a basis for decision making but did not actually quantify an actual decisionmaker's utility functions.

E. SUMMARY

Utility theory, on the surface, has direct applications to the problems faced in military analysis. In particular, it can be used to aggregate quantitative assessments for decisionmaker evaluation and to deal with those qualitative issues not amenable to straightforward quantitative analysis. However, it is been around a relatively long time and has not seen much direct application. We feel that the main reason for this is that a great deal of interaction is needed between the analyst and the decisionmaker in order to specify utility functions, and the success of this interaction depends on a decisionmaker's ability to identify his own indifference among certainties and lotteries, which, to many decision-

---

[19] Keeney and Raiffa, op. cit.

makers, may be a difficult task. Ensuring that the various postulates that apply to utility theory are met also may be onerous. Hence, there are significant obstacles to the direct application of utility theory. Even if all that could be done, problems of context will remain, as well as the difficulty of separating indifference from indecision. These may be inherent problems that cannot be overcome in general, but which may be minimal in a particular application. There is no sure way of determining if that is the case, however.

These inherent problems, combined with the level of effort needed to create utility functions, make utility theory cumbersome and uncertain; therefore, it probably will remain applicable only in very limited cases.

VI. VOTING AND PAIRED COMPARISONS

A. INTRODUCTION

Many decisionmaking techniques require, desire or (in some meaningful sense) "work better" if they are provided with consistent inputs. For example, if there is one decisionmaker, then decisionmaking techniques function more reasonably if that person is consistent in always preferring option A to option C whenever there is another option, say B, such that the decisionmaker both prefers option A to option B and prefers option B to option C. If there are multiple decisionmakers, then these techniques function more reasonably if the decisionmakers can reach a consistent consensus concerning their preferences.

Voting theory and the theory of paired comparison are quite different from such techniques in that both are designed to reach decisions when there are conflicts or certain types of inconsistencies among the inputs provided.

In particular, voting theory generally requires each voter to have internally consistent preferences, but it does not require all of the voters to agree with each other on one set of preferences. That is, while the voters may have to agree to accept the outcome of an election, they do not have to agree on one set of preferences concerning the candidates in order to hold the election. Thus, voting theory is specifically designed to address cases wherein there are multiple decisionmakers (voters) who have distinctly conflicting preferences concerning the relevant options (candidates).

Paired comparison theory is directly concerned with developing a ranking (e.g., best to worst) of a set of items based on the outcomes of a set of comparisons involving pairs of those items, where a "better item" is not necessarily selected over "worse item" in any given comparison. An example involves a set of teams that play various numbers of games against each other, and each game results in one team winning and one team losing that game. A paired comparison technique could then be used to rank those teams based on the outcomes of those games. Extensions that allow for ties are frequently (but not always) considered. Extensions that determine ranks based on magnitudes associated with the

comparisons (e.g., on the scores of the games) rather than just on who (or what) won and who (or what) lost occasionally are considered.

In terms of ranking teams by game results, if each team plays each other team the same number of times, and if one team (say team A) wins all of its games, while a second team (say team B) wins all of the rest of its games (i.e., all except its games against team A), and so forth, then developing a suitable ranking is trivial. However, there are many cases wherein the schedules or outcomes of the games do not satisfy this simple structure. To be useful, a paired comparison technique must be able to rank teams in more general cases. For example, it must be able to produce reasonable rankings in (many) cases in which every team has won some of its games, but has lost others. With a finite number of teams, this means that there is at least one subset of the teams, say T_1, ..., T_n, such that team T_i has defeated team T_{i+1} for i=1,...,n–1, yet team T_n has defeated team T_1. Thus, paired comparison theory is designed to address cases in which the outcomes (in terms of wins and loses) of individual paired comparisons are not consistent in that there can be at least one set of options, say T_1, ..., T_n, such that option T_1 is preferred to option T_2, T_2 is preferred to T_3, and so on through T_n, yet option T_n is preferred to option T_1.

Some general points to note here are as follows. If each decisionmaker involved has internally consistent preferences over the relevant options, and if these preferences are in agreement with those of all other decisionmakers involved, then neither voting theory nor the theory of paired comparisons have anything very special to offer. Other decisionmaking techniques, which may have been designed to exploit such situations, would be much more appropriate. However, if these conditions do not hold, and major incompatibilities exist among the preferences of the decisionmakers involved, then it may be quite useful to cast the relevant options and decisions into either a voting context or a paired comparison context (or both) for resolution.

Relatively brief descriptions of voting theory and the theory of paired comparisons are given in Sections B and C, below.

B. VOTING THEORY

At the outset it should be noted that the goal of this section is to provide a general description of several major aspects of voting theory. Precise descriptions, which would require precise (and correspondingly complex) notation, are not given, and not all aspects of voting theory are discussed.

1. Background and Motivation

Consider a situation in which v voters are presented with a alternatives, where v and a are strictly positive (finite) integers.[1] Since the cases in which a=1 (for any v) and v=1 (for any a) are clearly degenerate, assume that $a \geq 2$ and $v \geq 2$.[2] The goal of the vote is either to produce a ranking from 1 (most preferred) through a (least preferred) of these alternatives, perhaps with ties being allowed, or to select a′ alternatives as being "winners" (where a′ is an integer between 1 and a–1, inclusive). Frequently, a′=1.[3]

Many different schemes have been proposed to elicit preferences from voters and to determine rankings of alternatives based on those preferences. Even in the case in which a=2, there are several theoretically possible voting schemes that may reasonably apply in certain special cases. However, when a=2, there is one generally applicable voting scheme -- that of simple majority. To wit, each voter selects (votes for) exactly one of the two alternatives. The alternative receiving the majority of these votes is ranked ahead of the other, with a tie being declared if each alternative receives exactly the same number of votes. Two limitations could be raised concerning this simple majority voting scheme: First, it does not provide a method for breaking ties. Second, it applies only when a=2.

For good reasons, voting theory is not directly concerned with what to do about ties in cases like this (e.g., cases in which v is even, and exactly v/2 of the voters prefer each of the two alternatives over the other). Such ties are not major theoretical problems in the sense that there is little that further theoretical development can contribute here. In large-scale elections (i.e., voting situations with large numbers of voters), such ties are unlikely to occur and thus should not present major practical problems, provided some plausible tie-

---

[1] A generalization here would be to assign weights to each voter (i.e., voter i would be assigned weight w_i where $0 < w_i < \infty$ for i = 1,...,v), and the outcome of the vote would depend on these weights as well as on the voters' preferences. Unless explicitly stated otherwise, in the discussions below all voters are assumed to have equal voting power (i.e., $w_i = 1$ for all relevant i).

[2] Note that an election that considers one "candidate," and allows each voter to either "accept" or "reject" that candidate, constitutes a vote in which a=2.

[3] There are somewhat subtle differences between attempting to select a set of winners (even if a′=1) and attempting to develop a full ranking. If the set of alternatives being considered is finite (which is the case addressed here), then the latter gives the former. (This is not true when there are infinitely many alternatives.) However, there are particular voting structures that can select winners (ties must be allowed in these structures) but cannot be reasonably extended to produce what are normally considered as being full rankings. A careful discussion of these matters is given Sen (1970). However, Sen essentially shows that, given a finite number of alternatives, any characteristic of one of these goals (either selecting winners or developing a ranking) for all practical purposes, can be converted into an equivalent characteristic of the other. Given this practical equivalence, no distinction is made in this overview between the theoretical properties of methods designed to develop full rankings and those designed only to select one or more winners.

breaking procedure is agreed upon in advance. In elections involving small numbers of voters, such ties may occur relatively frequently, but the small numbers involved may allow a flexibility (or an inherent tie-breaking procedure) not present in large elections. In any event, there are several practical ways to break such ties, the choice of which may depend on the order of magnitude of the number of voters and the details of the situation.

While voting theory is not directly concerned with exact ties as just described, it is directly concerned with voting situations involving three or more alternatives. Simple majority rule, which is eminently appropriate for the a=2 case, has no natural analog when $a \geq 3$. As a result, several plausible structures for obtaining voters preferences and several desirable properties for converting these preferences into a ranking of the alternatives have been postulated, and many different voting schemes have been devised. The specification and analysis of voting methods when $a \geq 3$ has produced meaningful problems that, in general, do not have obvious or trivially appropriate solutions. The essence of voting theory is to address these problems, and the rich structure of this problem area has led to many research papers and books on voting theory.

Of course, rich theory does not imply significant practical importance, and one aspect of the practical importance of voting theory is debatable. This aspect concerns the likelihood of changing current voting methods.

On one side, the people most directly involved are interested in winning elections, not in changing voting methods. Accordingly, there can be little motivation for people in power to change an existing voting system to a new (and perhaps somewhat more complex) voting system based on what are essentially altruistic grounds, such as attempting to improve general fairness and reasonability. This is true especially when such changes are quite likely to affect the outcomes of future elections, but in unknown directions. In addition, only recently have computational capabilities become sufficiently powerful enough to calculate results of large-scale elections when using all but the simplest of voting techniques; although it had been argued that the computational aspects of converting to more sophisticated voting methods would make such changes far too expensive to seriously consider. Finally, theoretical work has shown that all voting methods are flawed in some sense when $a \geq 3$, and no one voting method is singularly less flawed than all of the others.

Of course, this is not to say that no changes should be made in voting methods; it does say, however, that for any proposed change, opponents can argue that the flaws in the proposed method will make that change not worthwhile. Also, proponents of some

changes, but not of the particular change in question, can always argue that, while changes should be made, the flaws of the proposed change means that it is not the right one to make. (In general, such arguments can be made about any change to any system.) Our point is that theoretical work on voting already has established many relevant arguments and corresponding flaws, and so relatively little effort is needed to prepare the specifics of such arguments against any particular change proposed for any particular voting system.

On the other side, the rationale for taking a vote in the first place is related to the desire to reach what is, in some sense, a fair decision based on the preferences of the voters. Thus, there should be some motivation to examine the relative strengths and flaws of whatever voting method is being used in any given voting situation and, if a different voting system is found to be more appropriate and change is practical, then the voting method in question should be changed. In any given voting situation, some responsible set of officials should be interested in such matters. Also, since inexpensive but extremely powerful computers are now readily available, the computational complexity argument would appear no longer valid. Thus, the selection of a voting method for any particular voting situation now can depend on the inherent strengths and flaws of the relevant possible methods for the situation question, not on historical limitations concerning the capability to collect and process voters' preferences.

While the likelihood of changing voting methods in any particular voting situation is debatable, there is little disagreement that the impacts of such changes (if made) often would be quite significant. Roughly speaking, in order for the choice of voting system to be significant, there must be more than two alternatives with no single alternative being preferred by a strict majority of the voters.

Real decision processes certainly involve cases in which there are more than two meaningfully different alternatives. However, to structure the voting in such cases, one could propose converting a vote involving a alternatives into a set of a–1 pairwise votes. The winner of each pairwise vote (according to the simple majority rule discussed above) would remain viable, while the loser would be eliminated. After a–1 such pairwise votes, one overall winner would remain. This proposal suffers from two flaws. First, it may be hard to implement in many voting situations; that is, while such a structure might be relatively easily implemented in some situations (e.g., committees voting on amending and passing bills), it can be practically unimplementable in other situations (such as in large-scale elections). Second, voting theory shows that the agenda by which these pairwise comparison are made can strongly affect (to the extent of completely determining) the

result. There may be special cases in which this characteristic is not considered to be a major flaw; for example, if a strict majority of the voters agree on a particular agenda, then the characteristic that the agenda determines the winner may be quite reasonable. However, it may be exactly for those cases that the voters would not agree on the agenda. Also, if (for some reason) the only other voting methods available for use in a particular situation are believed to have even worse flaws, then this agenda-based pairwise voting method might be useful, at least until a more desirable method can be made available. (This pairwise method does have some positive characteristics; the major one will be discussed later.) As stated above, the basic rationale for voting concerns the desire to reach a decision based on the preferences of the voters. Using a voting method that (potentially) allows the agenda-maker to determine the outcome clearly violates this rationale.

In voting situations, there frequently are more than two alternatives available with no single alternative being preferred by a strict majority of the voters; thus the choice of the voting system to be used can significantly affect the outcome of those voting situations. This applies to a vote by a social committee concerning what to serve at an upcoming banquet as well as to a vote by the citizens of a country concerning the election of a President. In terms of the likelihood of changing the current system, the former may precede the later. However, the latter example better exemplifies the potential importance of the selection of voting systems to be used.

A frequently cited example concerning the importance of the voting system used is the 1970 election for Senator from New York. In that election, James R. Buckley received about 40 percent of the vote while his two opponents, Charles E. Goodell and Richard L. Ottinger, split the remaining 60 percent about evenly. As a result, Buckley was elected according to the plurality rules (i.e., each voter votes for exactly one candidate and whichever candidate receives the most votes wins, even if that candidate does not receive a strict majority of the votes cast). However, it is widely believed that Buckley would have lost to Goodell and he would have lost to Ottinger had he faced either in a head-to-head contest. The same observation with the same results applies to the 1983 Democratic primary for Mayor of Chicago; in that election, Harold Washington beat Jane Byrne and Richard M. Daley in a plurality contest, yet probably would have lost to either in a head-to-head election. Other examples abound. Perhaps the most important concerns the 1966 Democratic primary for Governor of Maryland. In that race, George P. Mahoney received about 40 percent of the vote while his two opponents, Thomas Finan and Carlton Sickles, each received about 30 percent, and so Mahoney was declared the winner under the plurality system being used. Both Finan and Sickles are relatively liberal, and Maryland is

a relatively liberal state. Mahoney is an unabashed ultraconservative, and it is extremely unlikely that he could have beaten either Finan or Sickles in a one-on-one contest. Maryland is a heavily Democratic state. However, in the main election, many Democrats could not support the ultraconservative Mahoney, and sufficiently many crossed over to vote Republican that the Republican candidate won. It is widely believed that, had either Finan or Sickles won the Democratic primary, then he would have beaten the (at that time, relatively obscure) Republican candidate, Spiro T. Agnew, in the main race. Agnew won, was later elected Vice President, and then resigned under pressure. Richard Nixon nominated Gerald Ford in Agnew's place and, when Nixon resigned, Ford became President. While this certainly is not a formal proof of importance, it is reasonable to believe that, had any of several different voting systems been used in that Maryland gubernatorial primary, then either Finan or Sickles would have been nominated instead of Mahoney, Agnew would never have been Governor much less Vice President, and the Presidency of the United States would have been different.

2. A Numerical Example

The following example is intended to help explain some of the general claims made above and to provide a setting for the following discussions of the specifics of several particular voting methods. Consider an election in which there are three candidates, A, B, and C, one of which is to be elected. Let [x,y,z] denote a preference for x over y, y over z, and x over z, where x, y, z is any particular permutation of A, B, C. Suppose that the voters preferences can be described as follows:

Preferences	Percentage of Voters with this Preference
[A,B,C]	20
[A,C,B]	15
[B,C,A]	40
[C,A,B]	25
all others	0

With these preferences, note that all of the voters are consistent in that, if a particular voter prefers x to y and y to z, then that voter necessary prefers x to z. Note, however, that a strict majority of the voters (60 percent) prefers A to B, a strict majority (60

percent) prefers B to C, and a strict majority (65 percent) prefers C to A. The property that consistent individual preferences can lead to inconsistent group preferences is a fundamental characteristic of voting theory (and of the study of group decisionmaking in general). This property is frequently called the Condorcet paradox in the voting theory literature. The simplest example that exhibits this paradox is when there are three voters, one of which has preference [A,B,C], one has preference [B,C,A], and one has preference [C,A,B]. However, this simple example is a special case in the sense that it involves an exact tie (exactly one third of the voters have each of three different and incompatible preferences). In the example above, each of the voters' preferences is individually consistent and no ties (or other "gimmicks") are involved.

This example also serves to demonstrate several other dilemmas concerning the choice of voting systems. For instance, it was stated above that if a particular voting situation were to be resolved by a set of pairwise voters according to a particular agenda, then it is possible that the agenda-maker has the power to determine the overall winner. In the example above, if the agenda-maker matches A against B with that winner facing C, then C is elected; if instead B is first matched against C with that winner facing A, then A is elected; and if C is first matched against A with the winner facing B, then B is elected. Thus, three different agendas produce three different winners even though the voters preferences remain the same throughout.

It was also stated above that, given a particular set of preferences, different voting systems could produce different winners. Three commonly proposed voting systems are as follows: Plurality -- elect whoever receives the most first place votes; Runoff -- match the top two candidates as measured by the number of first place votes received, and elect the majority rule winner of that pairwise match; Borda (named after its inventor) -- award each candidate one point for each opposing candidate ranked below that candidate on each voters preference list, and elect the candidate that receives the greatest number of these points. For simplicity, assume that there are 100 voters in the example above. Then the plurality method would count 35 first place votes for A, 40 for B and 25 for C. Accordingly, the plurality method would elect candidate B. The runoff method would match B (with 40 firsts) against A (with 35 firsts) in a two-candidate majority rule runoff. Since 60 voters prefer A to B while only 40 prefer B to A, the runoff would elect candidate A. The Borda method would give 95 points to A (2 points for each of A's 35 first place rankings plus 1 point for each of A's 25 second place rankings), it would give 100 points to B (2 points for each of 40 firsts plus 1 for each of 20 seconds), and it would give 105

points to C (2 points for 25 firsts plus 1 point for 55 seconds). Accordingly, the Borda method would elect candidate C.

3. Some Alternative Voting Methods

As stated above, many different voting methods have been devised. Several of the more common are described after the following brief discussion concerning ties.

Roughly speaking, two types of ties can occur in any of the voting methods described here, and a third type can occur in some (but not all) of them. One type is an input tie. This occurs when any given voter is absolutely indifferent over a subset of the available alternatives. All but one of the voting methods described here can be modified, at the cost of some increased complexity but with no additional theoretical difficulties, to handle such input ties; further, the one exception (approval voting) automatically handles such ties and so needs no such modification. For simplicity only, it is assumed here that no such input ties occur.

The second type of tie that can occur in any of these methods is an exactly tied outcome, namely, a tie in the results of the voting that could be broken (or changed) by any one voter if that voter (and no others) changed preferences. For example, an exactly tied vote would occur in plurality voting if the number of voters, v, divided by the number of alternatives, a, is an integer, and exactly v/a of the voters vote for each of the a alternatives. As stated above, reasonable methods exist to break such ties for each of the voting methods discussed here, where the specific details of the tie-breaking procedures can depend on the voting method and on the application involved. For simplicity, the discussions below implicitly assume that suitable tie-breaking procedures are used to break any exact ties that might occur, and, with two exceptions, no specific mention of such ties is made in these discussions. (The two exceptions concern methods that are also subject to a third type of tie.)

The third type of tie is a methodological tie in that it is created by the details of the voting method being used. Such ties can occur in voting methods that, in some sense, group the voters together. These methodological ties have the property that, in general, no single voter (acting alone) can break the tie by changing preference. Some specifics concerning methodological ties are discussed in conjunction with the descriptions of those voting methods subject to such ties.

The first three voting methods described below have commonly used descriptive names. The remaining methods are most frequently referred to by the name of their (best known) developer, and this convention is followed here.

a. Plurality Voting

Suppose that, for a given election, one alternative is to be declared the winner (i.e., $a' = 1$). Plurality voting asks each voter to select exactly one alternative from the set of available alternatives and to cast a vote for that alternative (and only that alternative). The alternative that receives the most votes wins, whether or not it receives more than half of the votes cast (i.e., whether or not it was selected by a strict majority of the voters).

If $a' \geq 2$, then plurality voting asks each voter to select a" alternatives, for any a" between 1 and a' (inclusive), and to cast one vote for each of the a" alternatives so selected (and to cast no other votes). Plurality voting then ranks the alternatives in order according to the number of votes received and the top a' alternatives are declared to be the winners. Of course, if each voter has ranked all a alternatives from 1 (most preferred) through a (least preferred), then plurality voting can be implemented for any a' by casting a vote for each alternative in one of the first a' places in each voter's rankings.

b. Approval Voting

Approval voting asks each voter to select a subset (of any size) from the set of available alternatives and to cast a vote for each alternative in that subset. That is, each voter votes "yes" (casts a vote) or "no" (does not cast a vote) for each alternative in the set. Approval voting then ranks the alternatives in order according to the number of "yes" votes they receive. If one alternative is to be declared the winner (i.e., $a'=1$), then the alternative that receives the most votes wins, whether or not that alternative was selected by a majority of the voters, and if it was so selected, whether or not some other alternatives were also selected by a majority of the voters. If $a' \geq 2$, then the top a' alternatives (in terms of votes received) are declared to be the winners.

c. Runoff Voting

There are significant theoretical and practical reasons for considering runoff voting when exactly one winner is desired ($a'=1$), and this is the only case considered here. In this case, runoff voting can be implemented in one of two equivalent ways. Either each voter can be asked to select exactly one alternative from the set of available alternatives (as in plurality voting), or each voter can be asked to rank the alternatives from 1 (most

preferred) through a (least preferred), where a is the number of alternatives under consideration.

In the first implementation, if one alternative receives more than half the votes cast, then that alternative wins. If no alternative receives more than half the votes cast, then the voters are asked to participate in a second voting. This second voting matches the top two alternatives as measured by the number of votes received in the first voting (and it includes no other alternatives). In this second voting, each voter is asked to select exactly one of two alternatives being matched, and to cast a vote for that alternative. The alternative that receives the most votes in this second voting is then declared the winner.

In the second implementation, if one alternative is ranked first by more than half the voters, then that alternative wins. If no alternative is ranked first by more than half the voters, then the top two alternatives in terms of the number of first place rankings are compared against each other. Of these two alternatives, the alternative that ranks higher than the other on a majority of the voters' rankings is declared the winner.

d. Borda Voting

As in the second implementation above, Borda voting asks each voter to rank the alternatives from 1 (most preferred) through a (least preferred), where a is the number of alternatives under consideration. Each alternative receives a–1 points for each voter that ranks that alternative first, it receives a–2 points for each second place ranking, and so forth (receiving 1 point for each next-to-the-last place ranking and no points for being ranked last). That is, each alternative receives one point for each competing alternative ranked below it on each voter's preference list. The alternatives are then ranked in order, according to the number of points they receive. If $a'=1$, then the alternative receiving the most points wins. If $a'\geq 2$, then the top a' alternatives (in terms of points received) are declared to be the winners.

e. Hare Voting

Hare voting was designed to directly address elections in which $a'\geq 2$. Of course, it can also be used when $a'=1$ and, for expository purposes, the $a'=1$ case is described first below. Following this description, the Hare voting method for general a' is described. For any value of a', the Hare voting method asks each voter to rank the alternatives from 1 through a, as described above.

Suppose $a'=1$. Then Hare voting proceeds as follows. If one alternative is ranked first by more than half of the voters, then that alternative wins. Otherwise, the alternative that receives the fewest first place votes is deleted from every voter's rankings, so that each voter now has a ranking of a–1 alternatives, and those voters that previously had the deleted alternative in first place now have a new first place alternative. In these new rankings, if an alternative is now ranked first by more than half of the voters, then that alternative wins. If not, then the process is continued by deleting from every ranking the alternative that received the fewest first place votes in the rankings of the a–1 alternatives that survived the first deletion. This second deletion results in each voter having a ranking of a–2 alternatives. The process continues until one alternative receives a strict majority of the first place votes. In the extreme, the process could continue until only two alternatives remain, in which case simple majority rule would determine the winner. Note that, if $a=3$ and $a'=1$, then Hare voting is identical to runoff voting.

A three-part rationale for Hare voting when considering general values of a' (i.e., $1 \le a' < a$) is as follows. First, if $a'=1$ and some particular alternative receives more than one half of the first place votes cast, then no other alternative can receive as many first place votes as that alternative; thus the Hare rationale argues that the alternative in question should be chosen. Similarly, if $a'=2$ and a particular alternative receives more than one third of the first place votes cast, then, at most, one other alternative can receive as many or more first place votes as that alternative and so, at most, one other alternative has a stronger claim to be chosen than that alternative. Since two alternatives are to be chosen, Hare argues that the alternative in question should be one of them. If there were another alternative that received more than one third of the first place votes cast, then that other alternative also would be chosen by Hare. If no other alternative received more than one third of the first place votes, then the alternative in question (the one that did receive more than one third of the first place votes) is chosen and the process is continued to find a second alternative. If $a'=3$, the same logic leads to the selection of any alternative that receives more than 25 percent of the first place votes cast, and so on. In general, if there are v voters, then any alternative that receives $q(v,a')$ or more first place votes is selected, where

$$q(x,y) = \lfloor x/(y+1) \rfloor + 1$$

and $\lfloor x \rfloor$ denotes the largest integer less than or equal to x.

The second part of the Hare rationale is based on the general premise that, once an alternative has been selected, its supporters have had (at least some of) their say and their

voting power should be reduced by some appropriate amount. If q votes are need for selection, the Hare rationale argues that the supporters of the selected alternative have used up q votes worth of voting power. This premise is implemented in the following manner. Each voter starts with an initial voting power of 1.0. If a particular alternative, say alternative i, receives q_i' first place votes where $q_i' \geq q(v,a')$, then: (1) that alternative is selected, (2) that alternative is deleted from the individual rankings of all of the voters, and (3) the voting power of those voters that ranked that alternative first is reduced (i.e., weighted) by a factor of $1-q(v,a')/q_i'$. (the rationale for this weighing factor is discussed in the next paragraph). Let n_1 denote the number of alternatives selected here, i.e.,

$$n_1 = \text{card } \{i: q_i' \geq \lfloor v/(a'+1) \rfloor + 1\}.$$

Note that $0 \leq n_1 \leq a'$. If $n_1 = a'$, then these a' alternatives are the winners by the Hare voting method. If $n_1 = 0$, then the next step in Hare voting involves the third part of the Hare rationale, which will be discussed below. Here, suppose that $0 < n_1 < a'$. Let

$$v_1 = v - n_1 q(v,a')$$

and

$$a_1' = a' - n_1$$

so that v_1 votes-worth of voting power remain to be used and there are a_1' positions left to be filled. If, as a result of deleting the n_1 alternatives just selected, one or more new alternatives now receive at least $q(v_1,a_1')$ first place votes (where the vote of each supporter of a selected alternative is reduced as described above), then those new alternatives are also selected, they are deleted from the individual voter's rankings, and the voting power of each of their supporters is reduced. In particular, if alternative j is selected here, and if alternative j received q_{1j}' (weighted) first place votes in the revised rankings, then the voting power of each supporter of alternative j is reduced by a weighting factor of $1-q(v_1,a_1')/q_{1j}'$. In general, at this stage some voters can have their voting power reduced by $1-q(v,a')/q_i'$ for some i selected in the "first round," some by $1-q(v_1,a_1')/q_{1j}'$ for some j selected in the "second round," some by the product

$$(1-q(v,a')/q_i')\,(1-q(v_1,a_1')/q_{1j}'),$$

and some voters can still have full (1.0) voting power. This process continues until either a′ alternatives have been selected or no alternative receives enough first place votes to be selected. If a′ alternatives have been selected, the selection process is over. If (at this point in the selection process) fewer than a′ alternatives have been selected and no remaining (unselected) alternative has enough first place votes to be selected according to the (revised and weighted) voters rankings, then the third part of the Hare rationale is applied.

Before discussing this third part, the reason for using $1-q(v,a')/q_i'$ as a weighting factor should be noted. As stated above, the basic rationale here is that, if alternative i is selected (in the first round), then the supporters of that alternative are assumed to have used up q(v,a′) votes worth of voting power in selecting this choice. However, they should not lose more than this amount of voting power. That is, these supporters should not have reason to believe that they are wasting their voting power when more than q(v,a′) of them rank alternative i first. Given that q_i' voters have ranked that alternative first and that q(v,a′) of these votes are to be considered as having been expended, these q_i' voters should have a total of $q_i'-q(v,a')$, votes worth of voting power remaining. Distributing this remaining voting power uniformly over the q_i' voters in question means that each such voter should retain a voting power of

$$(q_i' - q(v,a'))/q_i' = 1-q(v,a')/q_i'.$$

If, in tabulating the results of Hare voting, no remaining alternative receives enough first place votes to be selected, yet there are still positions left to be filled, then the third part of the Hare rational is employed. This third part is that the remaining alternative with the weakest claim to be selected is the alternative that is receiving the least number of first place votes in the current rankings. Accordingly, if, in these (revised and weighed) rankings, no alternative is receiving enough first place votes to be selected and one or more positions remain to be filled, then the alternative with the least number of first place votes is deleted from the rankings of all voters, thereby (in general) creating some new first place alternatives. (If not, then the alternative with the next fewest first place votes is removed from all of the rankings, and so on, thereby ensuring that new first place alternatives will eventually be created.) This creation of new first place alternatives allows the process to continue until a′ alternatives have been selected.

The Hare voting method is the logically straightforward (but notationally complex) implementation of this three-part rationale.

f. Copeland Voting

Copeland voting asks each voter to rank the alternatives from 1 through a as described above. For i=1,...,a and j=1,...,a, let

$$w'_{ij} = \begin{cases} 1 & \text{if } i \neq j \text{ and a strict majority of the voters rank alternative i ahead of alternative j} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$l'_{ij} = \begin{cases} 1 & \text{if } i \neq j \text{ and a strict majority of the voters rank alternative j ahead of alternative i} \\ 0 & \text{otherwise.} \end{cases}$$

For i=1, ..., a, let

$$w_i = \sum_{j=1}^{a} w'_{ij} \quad ,$$

$$l_i = \sum_{j=1}^{a} l'_{ij} \quad ,$$

$$t_i = a - 1 - (w_i + l_i) \quad ,$$

and

$$s_i = w_i - l_i \quad .$$

Copeland voting then ranks the alternatives in order according to s_i as just defined. That is, for all relevant i and j, alternative i is ranked ahead of alternative j if and only if $s_i > s_j$, and alternative i is tied with alternative j if $s_i = s_j$.

Note that, if a tie-breaking procedure is being used to break exact ties (where exact ties are as defined at the beginning of Section 3 above), then $t_i=0$ for all relevant i. However, Copeland voting also can result in methodological ties (which occur if $s_i = s_j$ for any $i \neq j$). Under Copeland voting, such methodological ties are not necessarily rare (even in large-scale elections), and there is no intrinsic method for breaking such ties. Of course, there are many extrinsic methods that can be used to break such ties; for example, Borda voting or Hare voting, appropriately modified, could be used to break any methodological ties that occur within Copeland voting.

Note also that, whether or not t_i and t_j are zero,

$$s_i > s_j,$$

if and only if

$$\frac{w_i + t_i/2}{a-1} > \frac{w_j + t_j/2}{a-1} \quad ,$$

and these inequalities hold if and only if

$$2w_i + t_i > 2w_j + t_j$$

In particular, Copeland voting can be viewed as matching alternative i against alternative j for each pair of distinct alternatives i and j. With this viewpoint, each alternative participates in a–1 matches, one each against each of the other a–1 alternatives, with the results being that alternative i wins w_i of these matches, loses l_i of these matches, and ties its opponent in t_i of these matches. If a tie is counted as half of a win and half of a loss, then

$$\frac{w_i + t_i/2}{a-1}$$

gives the winning percentage of alternative i in these matches. Accordingly, Copeland ranking (i.e., ranking by wins minus losses) is identical to ranking by winning percentage (counting ties as a half win and a half loss), which is the ranking method used by most team-sport leagues. (Instead of calculating a winning percentage, ice hockey typically gives each team 2 points for each win and 1 point for each tie, and uses total points to rank the teams. This clearly results in the same ranking as using winning percentage or using wins minus losses to rank the teams when the teams involved have all played the same number of games.)

Finally, it should be noted that if the voters rankings are such that there is one alternative that would win each one-on-one match against each other alternative (i.e., for some particular i, $w_{ij}=1$ for all $j \neq i$), then that candidate is called a Condorcet winner in the voting theory literature. As the example in Section 2 above shows, voters can have individually consistent rankings on a set of alternatives, yet, based on those rankings, no alternative is a Condorcet winner. However, if a Condorcet winner exists, it is unique. A relevant property concerning voting systems is whether or not a given voting system will always rank a Condorcet winner first (i.e., declare it the winner if $a' = 1$) whenever the individual voters' rankings are such that a Condorcet winner exists. Copeland voting

clearly has this property, and it is the only voting system of those discussed thus far in this section that necessarily selects a Condorcet winner if one exists. However, there are many other voting systems that also have this property, one of which is discussed next.[4]

g. Kemeny Voting

Kemeny voting asks each voter to rank the alternatives from 1 through a as described above. Kemeny's approach is to define a metric over the possible rankings (i.e., over permutations of {1, ..., a}), and to find a "consensus" ranking that minimizes the sum of the distances between this consensus ranking and each of the voter's rankings. Specifically, let $p = (p_1, ..., p_a)$ and $q = (q_1, ..., q_a)$ be any two permutations of {1, ..., a}; i.e., for all relevant i and j, $p_i \in \{1, ..., a\}$ and $p_i \neq p_j$ if $i \neq j$, and the same conditions apply to q. Accordingly, p and q can be viewed as being two (not necessarily distinct) voter's rankings of the alternatives involved. For all relevant i and j, let

$$d'_{ij}(p,q) = \begin{cases} 0 & \text{if } i=j \text{ or if } i\neq j \text{ and either alternative i is ranked ahead of alternative j by both p and q or alternative i is ranked behind alternative j by both p and q (i.e., p and q agree when comparing i with j)} \\ 1 & \text{otherwise (i.e., and p and q disagree when comparing i with j),} \end{cases}$$

and define the function d on pairs of rankings (i.e., on the cross product of permutations of {1, ..., a}) by

$$d(p,q) = \sum_{i=1}^{a-1} \sum_{j=i+1}^{a} d'_{ij}(p,q)$$

It is not difficult to show that d is a metric on rankings as defined here. (It also is not difficult to extend this structure to allow any voter to be indifferent over several alternatives, i.e., to allow input ties, which is how Kemeny suggests that this method be used. As above, it is assumed for simplicity here that no such input ties occur.) Assign each voter a distinct numeric label between 1 and v. For k=l,...,v, let

$$p_k = \left(p_1^k, \ldots, p_a^k\right)$$

---

[4] Another voting system that has this property is taking pairwise votes according to a fixed agenda, as described in Section 1 above. That is, if a Condorcet winning alternative exists, then this alternative will be the overall winner of any sequence of pairwise votes in which it participates.

be the preference ranking of voter k. Call q a consensus ranking if q is a permutation of {1,...,a} such that

$$\sum_{k=1}^{v} d(p^k,q) \leq \sum_{k=1}^{v} d(p^k,q')$$

for every possible ranking q' (i.e., for every q' that is a permutation of {1,...,a}). Since a is finite, there is always at least one such consensus ranking. If the voter's preferences are such that there is only one such ranking, then Kemeny voting ranks the alternatives in order, according to that unique consensus ranking. If there are multiple consensus rankings, then Kemeny voting ranks each alternative as being (in general, tied) at the smallest (closest to first) place in which that alternative appears in any of those consensus rankings. Accordingly, as with Copeland voting, Kemeny voting can result in methodological ties. However, also as with Copeland voting, if a Condorcet winning alternative exists, then Kemeny voting will rank that alternative uniquely in first place.

h. Other Voting Methods

As stated above, many other voting systems have been devised. These range from straightforward modifications of the methods described above. E.g., for Kemeny voting, redefine consensus rankings so that q is a consensus ranking if

$$\sum_{k=1}^{v} \left(d(p^k,q)\right)^2 \leq \sum_{k=1}^{v} \left(d(p^k,q')\right)^2$$

for every possible ranking q'-- to qualitatively different systems.

As an example of a qualitatively different type of system, consider the following structure. Each voter is asked to rate each alternative on a scale from 0 to 100 (i.e., on a scale from 0 to 1 to two decimal places), and each alternative receives a number of points equal to its rating by each voter. The alternatives are then ranked in order by the total number of points they receive from all of the voters.

Interested readers should consult the rather extensive voting theory literature for descriptions of other voting methods.

4. Axioms and Arrow's Impossibility Theorem

In addition to defining voting methods, papers on voting theory literature often construct one or more sets of axioms and discuss: (a) whether or not any voting system

can satisfy a given set of axioms, (b) if so, whether or not there is a unique voting system that satisfies that set of axioms, and (c) if not, then given a set of voting systems, which systems satisfy which axioms, and what constitutes counterexamples for those that do not.

Perhaps the best known and most significant such axiomatic structure is that of Arrow's Impossibility Theorem (for which, in part, Kenneth J. Arrow won a Nobel Prize). This theorem can be stated and proved in several different (but essentially equivalent) ways, and precise statements and proofs of it can be found in most texts that address voting theory. A rough description of the relevant axioms and resulting theorem is as follows.

Axiom 1: A voting method satisfies Axiom 1 if that method accepts as input any set of v voter's rankings of a alternatives, where v and a are positive integers, and produces a group ranking of these alternatives. (Ties can, but need not, be allowed in either the group or the voter's rankings. Inconsistencies are not allowed in either set of rankings.)

Axiom 2: A voting method satisfies Axiom 2 if, whenever all of the voters rank a particular alternative ahead of every other alternative, that method also ranks that alternative ahead of every other alternative.

Axiom 3: A voting method satisfies Axiom 3 if it does not allow the existence of a voter such that the group ranking is necessarily identical to that voter's ranking, no matter how the other voters rank the alternatives.

Axiom 4: Let V be any set of voters' rankings of the alternatives, and let x be any one of those alternatives. Construct a revised set of rankings, V^x, by moving x up in one of the voter's rankings, but making no other changes. Then a voting method satisfies Axiom 4 if, for any alternative y, that voting method necessarily ranks x ahead of y according to V^x whenever that method ranks x ahead of y according to V.

Axiom 5: Let V be any set of voters' rankings of the alternatives, and let x and y be any two of those alternatives. Construct a revised set of voters ratings, V^{xy}, by allowing the voters to change their rankings in any manner whatsoever, provided only that those voters that ranked x somewhere ahead of y still do so, those voters that ranked y somewhere ahead of x still do so, and those voters that were indifferent between x and y still are so. Then a voting method satisfies Axiom 5 if that voting method necessarily ranks x ahead of y according to V^{xy} whenever that method ranks x ahead of y according to V and it necessarily is indifferent between x and y according to V^{xy} when it is indifferent between x and y according to V.

Arrow's Impossibility Theorem states that no voting method can satisfy all five of these axioms if there are two or more voters and three or more alternatives. In addition to formally stating and proving this theorem, texts on voting theory often discuss alternative forms and interpretations of these axioms and of this result. A very brief such discussion is as follows.

Axiom 1, or something similar, is needed to define the structure being considered. Occasionally this structure is defined in a preamble to the other axioms, not as a separate axiom. However, it can be useful to present this structure as an axiom both for clarity and because there are several voting methods (such as approval voting) that satisfy Axioms 2, 3, 4, 5, but do not satisfy Axiom 1.

Axiom 2 is needed to prevent an external (non-voting) dictator from determining the result (no matter what the voters' preferences are), and Axiom 3 is needed to prevent an internal (voting) dictator from determining the result (no matter what the other voters' preferences are). Note that, if a voting method bases its results on such a dictator, then that method satisfies Axioms 1, 4, and 5, and either 2 or 3 depending on whether the dictator is internal (thus violating 3) or external (thus violating 2), respectively.

Axiom 4 is redundant in the sense that it is not needed; no voting method can satisfy Axioms 1, 2, 3, and 5. While not needed, it is useful to state Axiom 4 as a separate axiom for (at least) the following reason. Many voting methods satisfy Axioms 1, 2, 3, while none can satisfy all five. Accordingly, it can be useful and important to have objective criteria that discriminate among the methods that satisfy 1, 2, and 3. Axiom 4 provides one such criterion (i.e., does any such voting method also satisfy Axiom 4?). For example, plurality voting, Borda voting, Copeland voting, and Kemeny voting all satisfy Axiom 4, whereas Hare voting does not.

Axiom 5 is, in a sense, the killer -- it sounds somewhat reasonable, but it is just too hard to satisfy. Axiom 5 is frequently referred to using the phrase "independence of irrelevant alternatives." In many voting situations there will be what might be called minor, nuisance, unimportant, and/or insignificant alternatives (e.g., weak third-party candidates). Clearly, one does not want the results concerning the other (i.e., major, primary, important, or significant) alternatives dependent on how (or whether) these minor, etc, alternatives are considered. But this is not (only) what Axiom 5 forbids. Axiom 5 also forbids the results of the vote concerning any two alternatives to depend on how any other alternatives are considered, no matter how meaningful, important, or significant those other

alternatives are. Viewed in this light, it may not be surprising that no voting method can satisfy Axioms 1, 2, 3, and 5.

5. National Security Applications

Voting theory may well have the property that, where it can (reasonably) be applied, it is *obviously* applicable; no ingenuity is needed to determine whether or not voting theory is applicable in a particular situation. How to apply voting theory (where it is applicable) is, in general, very complex and can require considerable knowledge and ingenuity. Accordingly, the interested reader may want to concentrate on developing an understanding of the details and interpretations of various aspects of voting theory, as opposed to searching for unconventional applications.

Two standard applications of voting theory concern (1) determining how citizens elect representatives in a representative democracy, and (2) if some of those representatives are to make group decisions, determining how those decisions are to be reached. Established democracies have established procedures for making these decisions, and voting theory may be best applied in such democracies by suggesting (perhaps a series of) marginal changes to those procedures, where appropriate. Voting theory might be applied more quickly and more extensively in new (emerging) democracies, which do not have well established procedures. Such applicability could have a significant impact on (and so be quite important to) the people in those new democracies. However, this impact is not necessarily important (other than in a humanitarian sense) for people in other countries because those new democracies may be small countries that have little influence or effect on people outside their borders.

If one views the Soviet Union and the Warsaw Pact countries of central and eastern Europe as being emerging democracies in the 1990s, the last comment above does not apply. Indeed, it can be argued that the success of the democratization of these countries is extremely important to the national security of the United States (and all other countries). It also can be argued that democracy in the United States succeeds in spite of, not because of, the particular voting structures currently in use here. Accordingly, it may be quite important for the United States to help the countries of central and eastern Europe establish strong democracies, but it may not be wise for the United States to try to do so by recommending the adoption of the systems, structures, or methods currently in use here. Instead, helping to enhance the understanding and implementation of appropriate aspects of voting theory in those European countries could be quite important to those countries in the

near future, and be quite important to the national security of the United States (and all countries of the world) thereafter.

6. Annotated Bibliography

a. Monographs

An inexpensive, entertaining, and informative monograph on voting is:

P.D. Straffin, "Topics in the Theory of Voting," UMAP Expository Monograph Series; Boston, Basel, and Stuttgart: Barkhäuser, 1980.

The interested reader may want to purchase this monograph directly. A related monograph in the same series is:

S.J. Brams, "Spatial Models of Election Competition," UMAP Expository Monograph Series; Boston, Basel, and Stuttgart: Barkhäuser, 1979.

b. Books

(1) Recent Books on Voting Theory

The books listed below are relatively recent texts that directly concentrate on aspects of voting that are described above (some are available in paperback).

S.J. Brams and P.C. Fishburn, "Approval Voting," Boston, Basel, and Stuttgart: Barkhäuser, 1982.

P.C. Fishburn, "The Theory of Social Choice," Princeton: Princeton University Press, 1973.

J.S. Kelly, "Social Choice Theory, An Introduction," Berlin: Springer-Verlag, 1987.

D.C. Mueller, "Public Choice II," Cambridge: Cambridge University Press, 1989.

T. Schwartz, "The Logic of Collective Choice," New York: Columbia University Press, 1986.

A.K. Sen, "Collective Choice and Social Welfare," New York: Elsevier Science Publishing Company, Inc., 1970.

(2) Books Relating Voting and Game Theory

The books listed below discuss aspects of game theory and voting theory, and the relationship between them. All are relatively recent and all have been published in paperback.

S.J. Brams, "Game Theory and Politics," New York: The Free Press, 1975.

S.J. Brams, "The Presidential Election Game," New Haven: Yale University Press 1978.

P.C. Ordeshook, "Game Theory and Political Theory, An Introduction," Cambridge: Cambridge University Press, 1986.

(3) Books with Chapters on Voting Theory

Several more generally oriented textbooks contain chapters devoted to voting theory. Two examples are listed below.

J. Malkevitch and W. Meyer, "Graphs, Models and Finite Mathematics," Englewood Cliffs, NJ: Prentice-Hall Inc., 1974.

F. Roberts, "Discrete Mathematical Models, with Applications to Social, Biological, and Environmental Problems," Englewood Cliffs, NJ: Prentice-Hall Inc., 1976.

(4) Books of Historical Interest

The books listed below were written a (relatively) long time ago, but are worthwhile to note for their historical contributions.

K.J. Arrow, "Social Choice and Individual Values," Second Edition, New Haven and London: Yale University Press, 1951.

D.Black, "The Theory of Committees and Elections," Cambridge: Cambridge University Press Cambridge, 1958.

J. Buchanan and G. Tullock, "The Calculus of Consent, Logical Foundations of Constitutional Democracy," Rexdale, CA: John Wiley & Sons, 1962.

R. Farquharson, "Theory of Voting," New Haven: Yale University Press, 1969.

R.D. Luce and H. Raiffa, "Games and Decisions, Introduction and Critical Survey," New York: John Wiley & Sons, 1957.

(5) Books of Related Interest

The following book is a fascinating text on a closely related subject.

M. Balinski and H.P. Young, "Fair Representation, Meeting the Ideal of One Man, One Vote," New Haven and London: Yale University Press, 1982.

The following book discusses a potpourri of topics related to voting theory.

S.J. Brams, W.F. Lucas, and P.D. Straffin (eds.), "Political and Related Models," New York: Springer-Verlag, 1983.

The following book discusses the relationship between voting theory and multicriterion decisionmaking.

K.J. Arrow and H. Reynaud, "Social Choice and Multicriterion Decision-Making," Cambridge: The MIT Press, 1986.

c. Papers

A characteristic of papers on voting theory is that they appear in a wide variety of journals. For example, voting theory papers appear in journals on mathematics, applied mathematics, operations research, management science, econometrics, economics, political science and behavioral science. They also appear in specialized journals (such as *Public Choice*) and in popular journals (such as *Scientific American*). A representative but far from exhaustive bibliography of papers on voting theory is given in Section (2), below. The interested reader might want to begin examining this literature by starting (in order of appearance) with the papers listed in Section (1).

(1) Selected Papers

D.H. Blair and R.A. Pollak, "Rational Collective Choice," *Sci. Amer.*, Vol 249, No. 2 (1983), 88-95.

P.C. Fishburn, "Condorcet Social Choice Functions," *SIAM J. Appl. Math.*, Vol 33 (1977), 469-486.

H. Hertzberg, "Let's Get Representative," *The New Republic,* June 29, 1987, 15-18.

P.C. Fishburn and S.J. Brams, "Paradoxes of Preferential Voting," *Math. Magazine*, Vol 56, No. 4 (1983), 207-214.

G. Doron, "Is the Hare Voting Scheme Representative?," *J. of Politics*, Vol 41, (1979), 918-922.

R.G. Niemi, "The Problem of Strategic Behavior Under Approval Voting," *Amer. Pol. Sci. Rev.*," Vol 58 (1984), 952-958.

S.J. Brams and P.C. Fishburn, "Comment on the Problem of Strategic Voting Under Approval Voting," *Amer. Pol. Sci. Rev.*, Vol 79 (1985) 816-818.

R.G. Niemi, "Reply to Brams and Fishburn," *Amer. Pol. Sci. Rev.*, Vol 79 (1985), 818-819.

H.P. Young, "Social Choice Scoring Functions," *SIAM J. Appl. Math.*, Vol 28, No. 4 (1975), 824-838.

D.G. Saari, "Inconsistencies of Weighted Summation Voting Systems," *Math. of Oper. Res.*, Vol 7, No. 4 (1982), 479-490.

P.C. Fishburn, "Inverted Orders for Monotone Scoring Rules," *Discrete Appl. Math.*, Vol 3 (1981), 27-36.

P.C. Fishburn, "Paradoxes of Voting," *Amer. Pol. Sci. Rev.*, Vol 68 (1974), 537-546.

P.C. Fishburn, "On the Sum-of-Ranks Winner When Losers Are Removed," *Discrete Math.*, Vol 8 (1974), 25-30.

P.C. Fishburn, "A Comparative Analysis of Group Decision Methods," *Behav. Sci.*, Vol 16 (1971), 538-544.

W.D. Cook and M. Kress, "Ordinal Ranking with Intensity of Preference," *Management Sci.*, Vol 31, No. 1 (1985), 26-32.

M.A. Satterthwaite, "Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *J. of Econ. Theory*, Vol 10 (1975), 187-217.

H. Nurmi, "Voting Procedures: A Summary Analysis," *B.J. Pol. Sci.* Vol 13 (1983) 181-208.

(2) Representative Bibliography

I. Ali, W.D. Cook, and M. Kress, "Ordinal Ranking and Intensity of Preference: A Linear Programming Approach," *Management Sci.*, Vol 32, No. 12 (1986), 1642-1647.

R.D. Armstrong, W.D. Cook, and L.M. Sieford, "Priority Ranking and Consensus Formation: The Case of Ties," *Management Sci.*, Vol 28, No. 6 (1982), 638-645.

D.M. Barton, "Constitutional Choice and Simple Majority Rule: Comment," *J. of Pol. Economy*, Vol 81, No. 2 (1973), 471-479.

J. Barzilai, W.D. Cook, and M. Kress, "A Generalized Network Formulation of the Pairwise Comparison Consensus Ranking Model," *Management Sci.*, Vol 32, No. 8 (1986), 1007-1014.

J. Benneche and B. Wing, "A Study of Approval Voting," Student Paper for OR 291, George Washington University, Washington DC, December 1984.

J.G. Birnberg, L.R. Pondy, and C.L. Davis, "Effect of Three Voting Rules on Resource Allocation Decisions," *Management Sci.*, Vol 16, No. 6 (1970), B356-B372.

D.H. Blair and R.A. Pollak, "Rational Collective Choice," *Sci. Amer.*, Vol 249, No. 2 (1983), 88-95.

J.M. Blin and A.B. Whinston, "Discriminant Functions and Majority Voting," *Management Sci.*, Vol 21, No. 5 (1975), 557-566.

V.J. Bowman and C.S. Colantoni, "Majority Rule Under Transitivity Constraints," *Management Sci.*, Vol 19, No. 9 (1973), 1029-1041.

V.J. Bowman and C.S. Colantoni, "Transitive Majority Rule and the Theorem of the Alternative," *Oper. Res.*, Vol 22 (1974), 488-496.

S.J. Brams, "Strategic Information and Voting Behavior," *Society*, Vol 19 (1982), 4-11.

S.J. Brams, "Approval Voting in Multicandidate Elections," *Policy Studies J.*, Vol 9, No. 1 (1980), 102-108.

S.J. Brams and P.C. Fishburn, "Approval Voting," *Amer. Pol. Sci. Rev.*, Vol 72 (1978), 831-847.

S.J. Brams and P.C. Fishburn, Reply to Tullock, *Amer. Pol. Sci. Rev.*, Vol 73 (1979), 552.

S.J. Brams and P.C. Fishburn, "Comment on the Problem of Strategic Voting Under Approval Voting," *Amer. Pol. Sci. Rev.*, Vol 79 (1985) 816-818.

D.E. Campbell, "Some Strategic Properties of Plurality and Majority Voting," *Theory and Decision*, Vol 13 (1981), 93-107.

A. Caplin and B. Nalebuff, "On 64% Majority Rule," Dept. of Economics, Harvard University, Cambridge MS (Caplin) and Dept. of Economics, Princeton University, Princeton NJ (Nalebuff), September 1985.

J.A.K. Cave, "A Median Choice Theorem," Rand P-7333, The Rand Corporation, Santa Monica CA, April 1987.

W.D. Cook and M. Kress, "Ordinal Ranking with Intensity of Preference," *Management Sci.*, Vol 31, No. 1 (1985), 26-32.

W.D. Cook and L.M. Seiford, "Priority Ranking and Consensus Formation," *Management Sci.*, Vol 24, No. 16 (1978), 1721-1732.

W.D. Cook and L.M. Seiford, "On the Borda-Kendall Consensus Method for Priority Ranking Problems," *Management Sci.*, Vol 28, No. 6 (1982), 621-637.

O.A. Davis, M.J. Hinich, and P.C. Ordeshook, "An Expository Development of a Mathematical Model of the Electoral Process," *Amer. Pol. Sci. Rev.*, Vol 64 (1970), 426-448.

A. Denzau, W. Riker, and K. Shepsle, "Farquharson and Fenno: Sophisticated Voting and Home Style," *Amer. Pol. Sci. Rev.*, Vol 79 (1985), 1117-1134.

G. Doron, "Is the Hare Voting Scheme Representative?," *J. of Politics*, Vol 41, (1979), 918-922.

J.S. Dyer and R.F. Miles, "An Actual Application of Collective Choice Theory to the Selection of Trajectories for the Mariner Jupiter/Saturn 1977 Project," *Oper. Res.*, Vol 24, No. 2 (1976), 220-244.

P.C. Fishburn, "Arrow's Impossibility Theorem: Concise Proof and Infinite Voters," *J. of Econ. Theory*, Vol 2 (1970), 103-106.

P.C. Fishburn, "The Irrationality of Transitivity in Social Choice," *Behav. Sci.*, Vol 15 (1970), 119-123.

P.C. Fishburn, "Intransitive Indifference with Unequal Indifference Intervals," *J. of Mathematical Psychology*, Vol 7 (1970), 144-149.

P.C. Fishburn, "Comments on Hansson's 'Group Preferences'," *Econometrica*, Vol 38, No. 6 (1970), 933-935.

P.C. Fishburn, "A Comparative Analysis of Group Decision Methods," *Behav. Sci.*, Vol 16 (1971), 538-544.

P.C. Fishburn, "Conditions on Preferences that Guarantee a Simple Majority Winner," *J. of Mathematical Sociology*, Vol 2 (1972), 105-112.

P.C. Fishburn, "On the Sum-of-Ranks Winner When Losers Are Removed," *Discrete Math.*, Vol 8 (1974), 25-30.

P.C. Fishburn, "Social Choice Functions," *SIAM Review*, Vol 16, No. 1 (1974), 63-90.

P.C. Fishburn, "Subset Choice Conditions and the Computation of Social Choice Sets," *Quarterly J. of Economics*, Vol 88 (1974), 320-329.

P.C. Fishburn, "Aspects of One-Stage Voting Rules," *Management Sci.*, Vol 21, No. 4 (1974), 422-427.

P.C. Fishburn, "Paradoxes of Voting," *Amer. Pol. Sci. Rev.*, Vol 68 (1974), 537-546.

P.C. Fishburn, "Condorcet Social Choice Functions," *SIAM J. Appl. Math.*, Vol 33 (1977), 469-486.

P.C. Fishburn, "Multicriteria Choice Functions Based on Binary Relations," *Oper. Res.*, Vol 25, No. 6 (1977), 989-1012.

P.C. Fishburn, "Axioms for Approval Voting: Direct Proof," *J. of Econ. Theory*, Vol 19 (1978), 180-185.

P.C. Fishburn, "A Strategic Analysis of Nonranked Voting Systems," *SIAM J. Appl. Math.*, Vol 35, No. 3 (1978), 488-495.

P.C. Fishburn, "Inverted Orders for Monotone Scoring Rules," *Discrete Appl. Math.*, Vol 3 (1981), 27-36.

P.C. Fishburn, "Discrete Mathematics in Voting and Group Choice," *SIAM J. Alg. Disc. Meth.*, Vol 5, No. 2 (1984), 263-275.

P.C. Fishburn, "Foundations of Decision Analysis: Along the Way," *Management Sci.*, Vol 35, No. 4 (1989), 387-405.

P.C. Fishburn and S.J. Brams, "Approval Voting, Condorcet's Principle, and Runoff Elections," *Public Choice*, Vol 36 (1981), 89-114.

P.C. Fishburn and S.J. Brams, "Paradoxes of Preferential Voting," *Math. Magazine*, Vol 56, No. 4 (1983), 207-214.

P.C. Fishburn and W.V. Gehrlein, "An Analysis of Simple Two-Stage Voting Systems," *Behav. Sci.* Vol 21 (1976), 1-12.

P.C. Fishburn and W.V. Gehrlein, "Borda's Rule, Positional Voting, and Condorcet's Simple Majority Principle," *Public Choice*, Vol 28 (1976), 79-88.

P.C. Fishburn and W.V. Gehrlein, "Towards a Theory of Elections with Probabilistic Preferences," *Econometrica*, Vol 45, No. 8 (1977), 1907-1924.

P.C. Fishburn and J.D.C. Little, "An Experiment in Approval Voting," *Management Sci.*, Vol 34, No. 5 (1988), 555-568.

M.M. Flood, "Implicit Intransitivity Under Majority Rule with Mixed Motions," *Management Sci.*, Vol 26, No. 3 (1980), 312-321.

M. Gardner, "Mathematical Games," *Sci. Amer.* Vol 231, No. 4 (1974), 120-125.

M. Gardner, "Mathematical Games," *Sci. Amer.* Vol 243, No. 4 (1980), 16-26B.

W.V. Gehrlein and P.C. Fishburn, "Condorcet's Paradox and Anonymous Preference Profiles," *Public Choice*, Vol 26 (1976), 1-18.

W.V. Gehrlein and P.C. Fishburn, "The Probability of the Paradox of Voting: A Computable Solution," *J. of Econ. Theory*, Vol 13 (1976), 14-25.

W.V. Gehrlein and P.C. Fishburn, "The Effects of Abstentions on Election Outcomes, *Public Choice*, Vol. 33, No. 2 (1978), 69-82.

A. Gibbard, "Manipulation of Voting Schemes: A General Result," *Econometrica*, Vol 41, No. 4 (1973), 587-601.

M. Gladwell, "When Votes are in, Have the People Spoken?" *The Washington Post*, June 4, 1990, A3

G.J. Glasser, "Game Theory and Cumulative Voting for Corporate Directors," *Management Sci.*, Vol 5 (1958), 151-156.

J. Greenberg and S. Weber, "Multiparty Equilibria Under Proportional Representation," *Amer. Pol. Sci. Rev.*, Vol 79 (1985), 693-703.

B. Hansson, "Group Preferences," *Econometrica*, Vol 37, No. 1 (1969), 50-54.

H. Hertzberg, "Let's Get Representative," *The New Republic*, June 29, 1987, 15-18.

M.J. Hinich, J.O. Ledyard, and P.C. Ordeshook, "Nonvoting and the Existence of Equilibrium Under Majority Rule," *J. of Econ. Theory*, Vol 4 (1972), 144-153.

M.J. Hinich and P.C. Ordeshook, "Abstentions and Equilibrium in the Electoral Process," *Public Choice*, Vol 7 (1969), 81-106.

D.T. Hoffman, "A Model for Strategic Voting," *SIAM J. Appl. Math.*, Vol 42, No. 4 (1982), 751-761.

D.T. Hoffman, "Relative Efficiency of Voting Systems," *SIAM J. Appl. Math.*, Vol 43, No. 5 (1983), 1213-1219.

J. Kellett and K. Mott, "Presidential Primaries: Measuring Popular Choice," *Polity*, Vol 9 (1977) 528-537.

J.G. Kemeny, "Mathematics Without Numbers," *Daedalus*, Vol 88 (1959), 577-591.

D.R. Kiewiet, "Approval Voting: The Case of the 1968 Election," *Polity*, Vol 12 (1979), 170-181.

A. Levenglick, "Fair and Reasonable Election Systems," *Behav. Sci.*, Vol 20 (1975), 34-46.

J.D.C. Little, "An Experiment in Approval Voting," *OR/MS Today*, Vol 12, No. 1 (1985), 18-19.

A. Mas-Colell and H. Sonnenschein, "General Possibility Theorems for Group Decisions," *The Rev. of Econ. Studies*, Vol XXXIX (2), No. 118 (1972), 185-192.

R.M. May, "Some Mathematical Remarks on the Paradox of Voting," *Behav. Sci.*, Vol 16 (1971), 143-151.

R.D. McKelvey, "A Theory of Optimal Agenda Design," *Management Sci.*, Vol 27, No. 3 (1981), 303-321.

R.D. McKelvey, "Constructing Majority Paths Between Arbitrary Points: General Methods of Solution for Quasi-Concave Preferences," *Math. of Oper. Res.*, Vol 8, No. 4 (1983), 549-556.

R.D. McKelvey, P.C. Ordeshook, and P. Ungar, "Conditions for Voting Equilibria in Continuous Voter Distributions," *SIAM J. Appl. Math.*, Vol 39, No. 1 (1980), 161-168.

R.D. McKelvey and R.E. Wendell, "Voting Equilibria in Multidimensional Choice Spaces," *Math. of Oper. Res.*, Vol 1, No. 2 (1976), 144-158.

D.K. Merchant and M.R. Rao, "Majority Decisions and Transitivity: Some Special Cases," *Management Sci.*, Vol 23, No. 2 (1976), 125-130.

S. Merrill, "Approval Voting: A 'Best Buy' Method for Multi-Candidate Elections?," *Math. Magazine*, Vol 52 (1979), 98-102.

S. Merrill, "Strategic Decisions Under One-Stage Multi-Candidate Voting Systems," *Public Choice*, Vol 36 (1981), 115-134.

R.G. Niemi, "The Problem of Strategic Behavior Under Approval Voting," *Amer. Pol. Sci. Rev.*, Vol 58 (1984), 952-958.

R.G. Niemi, "Reply to Brams and Fishburn," *Amer. Pol. Sci. Rev.*, Vol 79 (1985), 818-819.

R.G. Niemi and W.H. Riker, "The Choice of Voting Systems," *Sci. Amer.*, Vol 234, No. 6 (1976) 21-27.

S. Nitzan, "Some Measures of Closeness to Unanimity and Their Implications," *Theory and Decision*, Vol 13 (1981) 129-138.

H. Nurmi, "Voting Procedures: A Summary Analysis," *B.J. Pol. Sci.* Vol 13 (1983) 181-208.

D. Ray, "More Voting Paradoxes," *Math. Magazine*, Vol 57, No. 1 (1984), 57.

D.G. Saari, "Inconsistencies of Weighted Summation Voting Systems," *Math. of Oper. Res.*, Vol 7, No. 4 (1982), 479-490.

M.A. Satterthwaite, "Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *J. of Econ. Theory*, Vol 10 (1975), 187-217.

T. Schwartz, "Rationality and the Myth of the Maximum," *Nous*, Vol 6 (1972) 97-117.

A. Sen, "Social Choice Theory: A Re-examination," *Econometrica*, Vol 45, No. 1 (1977), 53-89.

P.P. Shenoy, "On Committee Decision Making: A Game Theoretical Approach," *Management Sci.*, Vol 26, No. 4 (1980), 387-400.

J. H. Smith, "Aggregation of Preferences with Variable Electorate," *Econometrica*, Vol 41 (1973), 1027-1041.

M. Staring, "Two Paradoxes of Committee Elections," *Math. Magazine*, Vol 59, No. 3 (1986) 158-159.

R. Stearns, "The Voting Problem," *Amer. Math. Monthly*, Vol 66 (1959), 761-763.

G. Tullock, "Constitutional Choice and Simple Majority Rule: Reply," *J. of Pol. Economy*, Vol 81, No. 2 (1973), 480-484.

G. Tullock, Comment on Brams and Fishburn, *Amer. Pol. Sci. Rev.*, Vol 73 (1979). 551-552.

H.P. Young, "A Note on Preference Aggregation," *Econometrica*, Vol 42, No. 6 (1974) 1129-1131.

H.P. Young, "An Axiomatization of Borda's Rule," *J. of Econ. Theory*, Vol 9 (1974), 43-52.

H.P. Young, "Social Choice Scoring Functions," *SIAM J. Appl. Math.*, Vol 28, No. 4 (1975), 824-838.

C. PAIRED COMPARISONS

Consistent with the goal of Section B above, the goal of this section is to provide a general description of the major aspects of paired comparison theory. Precise descriptions

of these aspects are not always given, and not all aspects of paired comparison theory are discussed.

1. Introduction

Paired comparison theory concerns the following situation. A set of objects is to be ranked from first (best) to last (worst), with (in general) ties being allowed in this ranking. This ranking is to be based on a set of comparisons of pairs of these objects where, in the simplest case, the outcome of each such comparison is that one of the two objects involved is strictly preferred over the other, with the degree of preference being irrelevant or immeasurable. Extensions that allow the comparisons to result in a tie (equal preference) are regularly considered, and will be discussed below. Extensions that consider a measure of the degree by which one object is preferred over the other are infrequently addressed, and will not be considered here. In general, not all pairs of objects are compared (i.e., each object is not necessarily compared with each other object), and some pairs of objects might be compared more than once. Further, when two objects are compared more than once, one of those objects does not necessarily win all of those comparisons.

The objects involved can be abstract objects (e.g., stability, deterrence, or social values), medical treatments, sensual stimuli, people or teams (e.g., the training, readiness, or quality of these people or teams), for example. The method of comparison can be based on qualitative judgments, on quantitative measurements or scores, or on a combination of both. A natural setting for a discussion of paired comparisons involves ranking a set of teams that compete against each other in some sport, and the discussion below will use this terminology. The reader should be aware that many other applications exist. Indeed, paired comparison theory seems to have been both developed and applied primarily by people interested in such other applications, and it seems to have been largely ignored by sports enthusiasts. Sports terminology, however, is quite useful for explaining the concepts involved.

Suppose that there are n teams, labeled T_1, T_2, ..., T_n, that play a set of games against each other, where each game involves two teams and results in one of those teams winning and the other losing that game, or (optionally) in a draw (i.e., a tied game). For $i \neq j$, let m_{ij} denote the number of games that T_i plays against T_j (so that $m_{ij} = m_{ji}$). Let r_{ij} denote the number of these m_{ij} games that T_i wins plus, if ties can occur, one half of the number of these m_{ij} games that result in a tie. Thus, for $i \neq j$,

$$r_{ij} + r_{ji} = m_{ij} = m_{ji}.$$

For simplicity, set $r_{ii} = m_{ii} = 0$ for all i. The goal of paired comparison theory is to develop an appropriate ranking of the teams based on the schedule (m_{ij}) and the results (r_{ij}).

The discussions below give general descriptions of various approaches that have been considered in the paired comparison literature--these discussions do not necessarily contain precise definitions or descriptions of the techniques involved. The rather extensive paired comparison literature should be consulted for such details. A very recent and relatively complete treatise on the theory of paired comparisons is given by David (1988). In particular, David describes in some detail most of the approaches discussed below, and he references almost 300 publications concerning paired comparisons. (While these references constitute an extensive bibliography on the subject, a few recent and quite relevant papers are not referenced there--these omissions are listed in Section 6.b, below.)

Two basically different types of approaches to determine rankings have been discussed in the paired comparison literature. One type consists of deterministic combinatorial approaches; the other type consists of approaches that pose some type of probability model of the competition and then determine rankings based on the schedule, results, and model so posed. The deterministic combinatorial methods are described first, followed by a description of the methods based on probability models.

2. Deterministic Combinatorial Methods

a. Winning Percentage

Perhaps the simplest and most straightforward method to rank teams is according to their winning percentage. Let

$$g_i = \sum_{j=1}^{n} m_{ij} \quad \text{and} \quad w_i = \sum_{j=1}^{n} r_{ij},$$

so that g_i is the total number of games played by T_i and w_i is the number of those games won by T_i (counting a tie as a half win and a half loss). Assume theat $g_i > 0$ for all i. Then the winning percentage of T_i, say p_i, is given by

$$p_i = w_i/g_i.$$

Ranking by winning percentage places T_i ahead of T_j if and only if $p_i > p_j$. If $p_i = p_j$ then T_i and T_j are tied in these rankings.

In symmetric round robin competitions (i.e., for all distinct i and j, $m_{ij} = \overline{m}$ for some positive integer $\overline{m}$), this is a quite reasonable and quite commonly used method. Three points should be noted here. First, in some applications, there may be more desirable methods for addressing ties. For example, tied games could simply be ignored (as if they never occurred), which would reduce g_i by the number of tied games involving T_i, and would reduce w_i by one half of this number of tied games. Second, as reasonable and as popular as this approach is for determining rankings in symmetric round robin competitions, other methods have been seriously proposed, and these other methods can produce different rankings than ranking by winning percentage, even if (in the competition being considered) no ties can occur. Two such methods are discussed below. Third, and perhaps most importantly, the reasonableness of ranking by winning percentage is strongly based on the round robin property that each team plays each other team an equal number of times. When some teams do not play some other teams ($m_{ij} = 0$ for some i and j), and/or they play different numbers of games against various other teams, then the "strength-of-the-schedule" becomes an important factor to consider in determining rankings. Indeed, one reason for studying paired comparison theory is to find methods that are appropriate for general schedules (i.e., no restrictions are placed on m_{ij} other than that m_{ij} is a nonnegative integer, $m_{ij} = m_{ji}$ for all distinct i and j, and $m_{ii} = 0$ for all i), yet these methods reduce to ranking by winning percentage in the special case of symmetric round robin competitions.

Many of the papers on paired comparisons mention ranking by winning percentages in some manner. Some do so to demonstrate that their methods differ from this ranking; others do so to show that their methods reduce to this ranking for symmetric round robins. Rubenstein (1980) addresses ranking by winning percentage directly in that he posits three very general and quite plausible axioms, and he proves that ranking by winning percentage is the only ranking method that always satisfies these axioms for competitions in which tied games are not possible and $m_{ij} = 1$ for all distinct i and j. See also David (1971).

b. The Kendall-Wei Method

This section gives a brief description of what is frequently called the Kendall-Wei method for determining rankings based on paired comparisons. Specific details can be found in the references cited below.

As background for the Kendall-Wei method, note that in symmetric round robins g_i is a constant, independent of i, and so ranking by winning percentage is equivalent to ranking by w_i. For simplicity, assume that tied games are not possible (i.e., all games

result in a win for one team and loss for the other), and give each team one "victory point" for each game it wins. Then ranking by winning percentage is equivalent to ranking by wins or by victory points as just defined.

The Kendall-Wei method argues that ranking solely by the number of games won is inappropriate because winning against stronger teams should count more than winning against weaker teams. Further, one measure of the strength of an opposing team is the number of games that the opposing team has won. To implement this measure, suppose that $m_{ij}=1$ for all distinct i and j (i.e., each team plays each others team exactly once and, since ties are not allowed, $r_{ij}=1$ if T_i beats T_j and $r_{ij}=0$ otherwise). Then, instead of giving a team one victory point for each win, this measure of strength could be reflected in the rankings by giving each team one victory point for each win by each opponent that the team in question has beaten.

For example, if T_1 beats T_2 and T_3 but no other teams, and if T_2 wins 4 games and T_3 wins 2 games, then T_1 would be given 6 victory points, not 2. Note that this implementation would treat the second-level teams (e.g., the teams beaten by T_2 and T_3) as counting equally. Instead, a third level could be considered in which, in this example, T_i is given one victory point for each team beaten by each of the 4 teams beaten by T_2 plus one victory point for each team beaten by each of the 2 teams beaten by T_3. Normalizing the resulting vector of victory points allows their process to continue indefinity. Given certain restrictions on the matrix $r=[r_{ij}]$, it has been shown that, in the limit as this process goes to infinity, the normalized vector of victory point approaches the vector v that satisfies the matrix equation

$$rv = \lambda v$$

where λ is the largest positive eigenvalue of r. (The restrictions on r are easily addressed, and the vector v is unique up to the normalization technique being used; the paired comparison literature should be consulted for details.) Teams are then ranked according to v, where T_i is ranked ahead of T_j if and only if $v_i > v_j$. (Note that this ranking is independent of the normalization technique used.)

The rationale above assumes that $m_{ij}=1$ for all i and j, $i \neq j$. However, the equation

$$rv=\lambda v$$

can be solved for v without this restriction and without the restriction that tied games are prohibited (the aforementioned easily-addressed restrictions on r still apply). Again, see the literature on the Kendall-Wei method for details. The important points to note here are

as follows. The Kendall-Wei method can be extended in a consistent manner to allow ties. That is, the rationale for using the Kendall-Wei method (and the relevant numerical techniques) can easily be adapted to allow ties--this is discussed in the relevant literature. However, while the Kendall-Wei method can also be numerically extended to consider schedules other than symmetric round robin tournaments (i.e., to consider cases in which teams play different numbers of games, including possibly no games, against various other teams), there seems to be no particularly good rationale or logical basis for doing so.

Two important characteristics of the Kendall-Wei method are as follows.

First, the Kendall-Wei method can produce a different ranking than that produced by winning percentages, even in the simple case in which there are no ties and $m_{ij}=1$ for all i and j, $i \neq j$. For example, if n=5 and

$$r_{ij} = \begin{cases} 1 & i < j \text{ and } (i,j) \neq (1,5) \\ 1 & (i,j) = (5,1) \\ 0 & \text{otherwise,} \end{cases}$$

then T_3 has 2 wins (over T_4 and T_5) and 3 losses, while T_5 has 1 win (over T_1) and 4 losses, yet the Kendall-Wei method ranks T_5 over T_3.

Second, it is not at all clear that wins against stronger teams should count more heavily than wins against weaker teams, especially if losses to weaker teams are not also counted more heavily than losses to stronger teams. For example, in a symmetric round robin tournament with no ties, if T_i and T_j have the same win-loss record and if T_i has k more wins than T_j over some set of strong teams, then (calling the other teams weak) T_j must necessarily have k fewer losses than T_i against weak teams. In such a case, the Kendall-Wei method would rank T_i over T_j, yet it is plausible here that these two teams should be ranked equally. In a similar vein, it has been shown that, in the Kendall-Wei method, changing all wins to losses and all losses to wins does not necessarily reverse a ranking.

Variations of the Kendall-Wei method have been proposed that more heavily consider both wins over strong teams and losses to weak teams and, not surprisingly, these methods can reduce to simply ranking by winning percentages for symmetric round robin tournaments. See the literature for details.

In addition to David (1988), various properties, limitations, and extensions of the Kendall-Wei method are discussed in David (1971), Goddard (1983), Kendall (1955), Moon (1968), Ramanujacharyulu (1964), and Stob (1985).

c. The Minimum Violations Method

The minimum violations method for paired comparisons is the logical analog of the Kemeny method for voting. Specifically, consider a symmetric round robin tournament with n teams in which $m_{ij}=1$ for all distinct i and j and ties are not allowed. Accordingly, $r_{ij}=1$ if T_i beat T_j in the game they played, and $r_{ij}=0$ otherwise. Let S denote the set of all permutations of $\{T_1,...,T_n\}$, so that S consists of all possible rankings (i.e., standings) of the teams in question. For any given ranking $s \in S$, let $r^s = \left[r_{ij}^s\right]$ be the "perfect prediction" results based on s in that $r_{ij}^s = 1$ if T_i is ranked ahead of T_j according to s, and $r_{ij}^s = 0$ otherwise. Let

$$d(s,r) = (1/2) \sum_{\substack{i,j \\ i \neq j}} (1-r_{ij}^s)(1-r_{ij})$$

so that d(s,r) is the number of pairs (i.e., distinct but unordered i and j) in which the ranking s rates T_i over T_j when, in their game, T_j beat T_i. In other words, given a ranking s and actual game results r, d(s,r) is the number of upsets (according to s) that occurred in those games.

As an example, consider the example given in Section b above, namely n=5 and

$$r_{ij} = \begin{cases} 1 & i < j \text{ and } (i,j) \neq (1,5) \\ 1 & (i,j) = (5,1) \\ 0 & \text{otherwise.} \end{cases}$$

Let

$$s^1 = (T_1, T_2, T_3, T_4, T_5),$$

$$s^2 = (T_1, T_2, T_3, T_5, T_4), \text{ and}$$

$$s^3 = (T_1, T_2, T_5, T_3, T_4).$$

Then

$$d(s^1,r) = 1,$$

$$d(s^2,r) = 2, \text{ and}$$

$$d(s^3,r) = 3.$$

Given r, a ranking is a minimum violations ranking if it minimizes d(s,r) over all $s \in S$.

The minimum violations method has the following characteristics. First, it is relatively difficult to compute. Powerful algorithms (which have recently been developed) and modern (very fast) computers can mitigate problems caused by this characteristic, but it is still worth noting.

Second, like the Kendall-Wei method, the minimum violations method can be extended in a consistent manner to allow tied games. It can also be numerically extended to consider more general schedules then symmetric round robin tournaments; however, as with Kendall-Wei, there seems to be no particularly good rationale or logical basis for doing so.

Third, again like Kendall-Wei, the minimum violations method can produce a different ranking than that produced by winning percentages, even in the simple case in which there are no ties and $m_{ij}=1$ for all i and j, $i \neq j$. An example is as follows. Let n=7,

$$r_{ij} = \begin{cases} 1 & i < j \text{ and } (i,j) \notin \{(1,2), (2,5), (2,6), (3,7)\} \\ 1 & (i,j) \in \{(2,1), (5,2), (6,2), (7,3)\} \\ 0 & \text{otherwise,} \end{cases}$$

$$s^1 = (T_1, T_2, T_3, T_4, T_5, T_6, T_7),$$

and

$$s^2 = (T_2, T_1, T_3, T_4, T_5, T_6, T_7) .$$

Then $d(s^1,r) = 4$ and $d(s^2,r) = 3$. Thus s^2, which ranks T_2 ahead of T_1, has fewer violations than s^1, which ranks T_1 ahead of T_2. However, T_1 has 5 wins and 1 loss in this example while T_2 has 4 wins and 2 losses.

Fourth, the minimum violations method as described here (and as described in the majority of the references) treats every violation in exactly the same way when evaluating a ranking, no matter how close together or how far apart the teams involved are in that ranking. However, it can be argued that teams that are, in fact, close in quality to each other are likely to win some games and lose others to each other, and so violations involving such teams should not be treated in the same manner as violations involving teams that are far apart in quality. According to this argument, a ranking that has six violations, all of which involve teams that are close to their opponents in that ranking, could well be a more reasonable ranking than one that has four violations all involving teams that are ranked far apart from each other.

For further discussions of the minimum violations method, see Ali, Cook, and Kress (1986), David (1971), Goddard (1983), Harary and Moser (1966), Stob (1985), and Thompson (1975).

3. Probabilistic Methods

The probabilistic approaches discussed here all (in some sense) involve calculating probabilities, say p_{ij}, that T_i will win a game played against T_j for all relevant i and j. These probabilities can be interpreted as estimates, based on game results, of true but unknown probabilities. They can also be interpreted as summary output measures that can be used for ranking teams based on (in general, intransitive) game results. These probabilities are determined by making a set of assumptions (different assumptions are made for different methods) and then calculating values for these probabilities that are consistent with these assumptions and with the game results in question. Two points should be noted here. First, the various assumptions made for these probabilistic methods seem to fit within an overall, integrated structure. This structure is presented below. Second, different assumptions do not necessarily lead to different results in that one set of assumptions can be equivalent to another set. Such equivalencies occur here and will also be discussed below. First, some relevant notion is introduced.

a. Notation

As above, suppose that there are n teams under consideration ($n \geq 2$), and that these teams are labeled $T_1, \ldots, T_n$. For $i=1,\ldots,n$, $j=1,\ldots,n$, and $i \neq j$, let m_{ij} denote the number of games that T_i plays against T_j, and let r_{ij} denote the number of these games that T_i wins plus one-half of the number of these games that result in a tie. For $i=1,\ldots,n$, let $m_{ii}=r_{ii}=0$. Note that $m_{ij}=m_{ji}$ and that $r_{ij}+r_{ji} = m_{ij}$ for all $i=1,\ldots,n$ and all $j=1,\ldots,n$. Call the matrix $m=[m_{ij}]$ the schedule matrix and the matrix $r = [r_{ij}]$ the results matrix. For $i=1,\ldots,n$, let

$$g_i = \sum_{j=1}^{n} m_{ij}, \quad \text{and} \quad w_i = \sum_{j=1}^{n} r_{ij} \quad .$$

Some new notation is as follows. Let P denote the set of all n by n matrices $p=[p_{ij}]$ such that, for all i and j from 1 through n,

$$0 \leq p_{ij} \leq 1 \qquad \text{and} \qquad p_{ij} + p_{ji} = 1.$$

Given $p \in P$, let

$$x_{ij} = x_{ij}(p_{ij}) = \begin{cases} p_{ij}/(1-p_{ij}) = p_{ij}/p_{ji} & p_{ij} < 1 \\ \infty & p_{ij} = 1 \end{cases} .$$

Also for $p \in P$, let

$$L(p) = \begin{cases} \displaystyle\prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \binom{m_{ij}}{r_{ij}} p_{ij}^{r_{ij}} p_{ji}^{r_{ji}} & \text{if no game between any two teams ends in a tie} \\ \displaystyle\prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \binom{2m_{ij}}{2r_{ij}} p_{ij}^{2r_{ij}} p_{ji}^{2r_{ji}} & \text{otherwise} \end{cases}$$

and, for $p \in P$ and $\rho \geq 1$, let

$$O_\rho(p) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left| r_{ij} - m_{ij}\, p_{ij} \right|^\rho .$$

Let R_1 denote the set of all n by n matrices such that, given any two different i and j in $\{1,...,n\}$, there exists some $k \geq 2$ such that $r_{i_1 i_2}, r_{i_2 i_3}, ..., r_{i_{k-1} i_k}$ are all strictly positive and either $i_1 = i$ and $i_k = j$ or $i_1 = j$ and $i_k = i$. Let R_2 denote the set of all n by n matrices such that, given any two different i and j in $\{1,...,n\}$, there exists some $k \geq 2$ such that $r_{i_1 i_2}, r_{i_2 i_3}, ..., r_{i_{k-1} i_k}$ are all strictly positive and $i_1 = i$ and $i_k = j$. That is, if $r \in R_1$, then there must be at least one "one-way beats-or-ties" path between any two teams; and if $r \in R_2$, then there must be at least one such path in each direction between any two teams. For example, if $r \in R_1$, then the teams cannot be partitioned into two sets such that both no team in the first set beats or ties any team in the second set and no team in the second set beats or ties any team in the first set. Similarly, if $r \in R_2$, then the teams cannot be partitioned into two sets such that no team in the first set beats or ties any team in the second set. Note that $R_2 \subset R_1$.

b. An Integrative Structure for Assumptions Underlying These Probabilistic Methods

Ignoring ties, each of the various papers that propose probabilistic methods for ranking alternatives seems to make two basic assumptions (as opposed to making just one such assumption or to making three or more such assumptions). The particular assumptions that are made can differ from paper to paper, but each such paper seems to need to make two of them in order to allow the development of a structure sufficiently rich that values for the relevant probabilities can be calculated. Furthermore, ignoring the treatment of ties, each such paper seems to take one of its assumptions from "Group A" and one from "Group B" as listed in Table VI-1.

Some comments on the structure proposed in Table VI-1 are as follows.

First, a formal structure like that of Table VI-1 does not seem to appear in the literature, although the discussions in David (1988) and Stob (1984) come quite close. The interested reader may want to test the hypothesis presented on Table VI-1 by examining various publications on paired comparisons. Table VI-2 presents a start in this direction by relating several such publications to the structure given in Table VI-1. Note that, while all of the publications listed on Table VI-2 consider assumptions in pairs as indicated on Table VI-1, several of these publications consider more than one such pair of assumptions. David (1988) and Stob (1984) consider so many of the possible pairs that it would be pointless to list these publications on Table VI-2. The interested reader here should certainly consider examining these two publications.

Second, note the following distinction between the assumptions in Group A and those in Group B. Each assumption in Group A essentially defines a model in that each requires some particular relationship among the p_{ij}'s to hold. However, none of the assumptions in Group A concern calculating numbers for these p_{ij}'s in that none concern the relationship between these p_{ij}'s and the results matrix r. Conversely, each of the assumptions in Group B determines a relationship that must hold between the probabilities p_{ij} and the outcomes given by the schedule matrix m and the results matrix r. However, none of the assumptions in Group B relate the p_{ij}'s to each other outside of relationships involving m and r, and, in particular, none of the assumptions in Group B place sufficient restrictions on the p_{ij}'s to allow values for these p_{ij}'s to be calculated as functions of m and r. The paired comparison literature shows that values for the p_{ij}'s can be computed as meaningful functions of m and r if (in general) one assumption from each column on Table VI-1 is made and if $r \in R_1$.

Third, the roles played by R_1 and R_2 here are as follows.

R_1 concerns whether or not teams can be compared by any of these methods. If $r \in R_1$, then all of the teams can be compared with each other. If $r \notin R_1$, then there are at least two teams that cannot be compared by any of these methods. For example, one way (but not the only way) that this ($r \notin R_1$) could occur is as follows. Suppose m and r are such that it is possible to partition the teams into four groups, say G_a, G_{b_1}, G_{b_2}, and G_c,

Table VI-1. An Hypothesized Structure for Integrating the Assumptions Underlying Probabilistic Methods for Ranking Alternatives

Select one assumption from Group A and one assumption from Group B.

Then, given m and r, find $p \in P$ satisfying the two selected assumptions.

GROUP A

A1.1: Strong Stochastic Transitivity

For all i, j, k, if $p_{ij} \geq 1/2$ and $p_{jk} \geq 1/2$, then $p_{ik} \geq \max\{p_{ij}, p_{jk}\}$.

A1.2: Moderate Stochastic Transitivity

For all i, j, k, if $p_{ij} \geq 1/2$ and $p_{jk} \geq 1/2$, then $p_{ik} \geq \min\{p_{ij}, p_{jk}\}$.

A1.3: Weak Stochastic Transitivity

For all i, j, k, if $p_{ij} \geq 1/2$ and $p_{jk} \geq 1/2$, then $p_{ik} \geq 1/2$.

A2: Odds Multiply

$x_{ik} = x_{ij}\, x_{jk}$ for all i, j, k such that neither

$$p_{ij} = 1 \text{ and } p_{jk} = 0$$

nor

$$p_{ij} = 0 \text{ and } p_{jk} = 1;$$

equivalently,

$$p_{ij}\, p_{jk}\, p_{ki} = p_{kj}\, p_{ji}\, p_{ik}$$

for all i, j, and k.

A3: Odds are Proportional to Deterministic Strengths

If $r \in R_2$ then n positive numbers, say

$$s_1, \ldots, s_n,$$

exist such that

$$p_{ij} = s_i/(s_i+s_j),$$

or, equivalently,

$$x_{ij} = s_i/s_j,$$

for all i and j.

A4: Wins are Determined by Random Strengths

Let F_μ denote the (cumulative) distribution function of a continuous random variable with mean μ. Given F_μ, if $r \in R_2$ then n real numbers, say

$$\mu_1, \ldots, \mu_n,$$

and n independent random variables, say

$$s_1, \ldots, s_n,$$

exist such that s_i has distribution F_{μ_i} and

$$p_{ij} = \text{Prob}(s_i > s_j)$$

for all i and j, where:

A4.1: Strengths Have Extreme Value Distributions

$F_\mu(x) = \exp(-e^{-(x-(\mu-\gamma))})$
where $\gamma = 0.5772\ldots$ is Euler's constant.

A4.2: Strengths are Normally Distributed With Common Variance

$$F_\mu(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(y-\mu)^2/2} dy.$$

A4.3, A4.4, ... Strengths Have Other Distributions

A4.*t*: Specify $F_\mu(x)$ for $t = 3, 4, \cdots$

GROUP B

B1: Expected Wins Equal Actual Wins

$$\sum_{j=1}^{n} p_{ij} m_{ij} = w_i \quad i = 1, \ldots, n.$$

B2: Maximum Likelihood

$L(p) \geq L(q)$ for all $q \in P$.

B3(ρ): Minimum Deviation

For a selected value of $\rho \geq 1$ (e.g., $\rho = 2$),
$O_\rho(p) \leq O_\rho(q)$ for all $q \in P$.

Table VI-2. The Relationship Between the Structure Proposed in Table VI-1 and Some Selected Papers on Paired Comparison Theory

	Assumptions as Listed on Table VI-1	
Publication	**Group A**	**Group B**
Bradley (1976)	A3	B2
"	A4.1	B2
Bradley & Terry (1952)	A3	B2
David (1971)	A1.1	B2
"	A1.3	B2
Dykstra (1960)	A3	B2
Fienberg & Larntz (1976)	A2	B2
"	A3	B2
Ford (1957)	A3	B2
Jech (1983)	A2	B1
Moon (1968)	A3	B2
Mosteller (1951)	A4.2	B3(2)
Stefani (1977)	A4.2	B3(2)
Stob (_985)	A1.1	B1
"	A3	B2
Thompson (1975)	A4.2	B2
Zermelo (1929)	A3	B2

such that each team in G_a (if any) has won all of its games against teams not in G_a, each team in G_c (if any) has lost all of its games against teams not in G_c, and no team in G_{b1} plays any games against any team in G_{b2}. Then no method described here can compare any team in G_{b1} with any team in G_{b2}. Accordingly, whether or not $r \in R_1$ is important concerning whether or not any of these probabilistic methods can be applied, but it is not relevant for distinguishing among these methods.

If $r \in R_1$ but $r \notin R_2$, then all teams are comparable, but some comparisons are "too easy" in that there are (at least) two teams, say T_i and T_j, such that the only reasonable value for p_{ij} based solely on m and r is $p_{ij}=1$. For example, suppose m and r are such that it is possible to partition the teams as described in the last paragraph above, that G_a, G_{b1}, and G_c are not empty, but G_{b2} is empty, and $r \in R_1$. Let $G_b=G_{b1}$. Then, based on m and r, it is reasonable to conclude that each of the teams in G_a is better then any team in G_b or G_c, each team G_b is better any team in G_c, and that, for any probabilistic model,

$$p_{ij} = 1,\ p_{jk} = 1, \text{ and } p_{ik} = 1$$

for all $T_i \in G_a$, $T_j \in G_b$, and $T_k \in G_c$. In this case, the interesting questions reduce to determining how the teams in G_a rate against each other, how those in G_b rate against each other, and how those in G_c rate against each other. In specific, if $r \in R_1$ but $r \notin R_2$, then the teams can be partitioned into subsets such that the restrictions of r to these subsets have the property that each restricted r belongs to the correspondingly restricted R_2. Accordingly, whether or not $r \in R_2$ is important concerning whether or not any of these probabilistic methods yield values for p_{ij} satisfying

$$0 < p_{ij} < 1$$

for all i and j, but it is not relevant for distinguishing among these methods.

Fourth, if (and, of course, only if) these games (i.e., comparisons) can end in ties, then the manner in which these ties are treated can be quite important.

As stated above, sports leagues tend to consider ties by treating them as half a win and half a loss (or in some essentially equivalent manner). Note that, in terms of Table VI-1, treating ties this way corresponds to treating ties under Group B (by adjusting r), not by changing the model postulated in Group A. For simplicity, Table VI-1 is structured to treat ties this way, which is why L(p) counts each game twice if ties can occur. If ties count as half a win and half a loss then r_{ij} need not be an integer, but $2r_{ij}$ will always be integral. The interested reader should note that this is not a standard approach in the paired comparison literature, but it does allow all of the combinations pairs of assumptions from

Group A and B on Table VI-2 to be implementable when ties can occur. Another way to make the structure on Table VI-1 directly implementable when ties are possible is to treat tied games as if they never occurred (instead of treating ties as half a win and half a loss). That is, if t_{ij} of the m_{ij} games between T_i and T_j result in ties, then replace the schedule and results matrices m and r with $\tilde{m}$ and $\tilde{r}$, respectively, where

$$\tilde{m}_{ij} = m_{ij} - t_{ij} \quad \text{and} \quad \tilde{r}_{ij} = r_{ij} - t_{ij}/2,$$

and rate the teams based on $\tilde{m}$ and $\tilde{r}$. Of course, ignoring ties can yield different rankings than treating ties as half a win and half a loss--a somewhat interesting example is given in Section 4, below.

A different general approach for considering ties in probabilistic models is to change the specifications of the model as determined under Group A by introducing a second matrix of probabilities, say $\tilde{p} = \tilde{p}_{ij}$, where $\tilde{p}_{ij}$ is the probability that a game between T_i and T_j ends in a tie. If this is done, then p and $\tilde{p}$ must have the properties that:

$$0 \le p_{ij} \le 1, \qquad 0 \le \tilde{p}_{ij} \le 1,$$

$$\tilde{p}_{ij} = \tilde{p}_{ji}, \qquad \text{and} \qquad p_{ij} + p_{ji} + \tilde{p}_{ij} = 1$$

for all relevant i and j.

For example, A3 could be changed so that

$$p_{ij} = \frac{s_i}{s_i + \theta\, s_j}$$

for some $\theta \ge 1$, which gives that

$$\tilde{p}_{ij} = \frac{(\theta^2 - 1)\, s_i s_j}{(s_i + \theta\, s_j)\,(s_j + \theta\, s_i)};$$

or it could be changed so that

$$p_{ij} = \frac{s_i}{s_i + s_j + \upsilon\,(s_i s_j)^{1/2}}$$

for some $\upsilon \ge 0$, which gives that

$$\tilde{p}_{ij} = \frac{\upsilon\,(s_i s_j)^{1/2}}{s_i + s_j + \upsilon\,(s_i s_j)^{1/2}}.$$

Alternatively, A4 could be changed by introducing threshold parameters, e.g.,

$$p_{ij} = \text{Prob}\{s_i > s_j + \eta\}$$

for some $\eta \geq 0$, which gives that

$$\tilde{p}_{ij} = \text{Prob}\{\,|s_i - s_j| \leq \eta\}.$$

Instead of threshold parameters, discontinuous distribution functions could be considered under A4, in which case $\tilde{p}$ could be defined by

$$\tilde{p}_{ij} = \text{Prob}\{s_i = s_j\}\ .$$

Introducing $\tilde{p}$ into the model in the specifications under Group A necessitates making the corresponding changes to the restrictions under Group B. In particular, B1 as stated on Table VI-1 would change to

$$\sum_{j=1}^{n} p_{ij}\, m_{ij} = \tilde{w}_i \qquad \text{and} \qquad \sum_{j=1}^{n} \tilde{p}_{ij}\, m_{ij} = \tilde{t}_i$$

where

$$\tilde{t}_i = \sum_{j=1}^{n} t_{ij} \qquad \text{and} \qquad \tilde{w}_i = w_i - \tilde{t}_i/2$$

for i=1,...,n. B2 and B3 remain as stated on Table VI-1; however, the definition of L(p) and $O_\rho(p)$ would change to

$$L(p) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \binom{m_{ij}}{\tilde{r}_{ij}} \binom{m_{ij} - \tilde{r}_{ij}}{t_{ij}} p_{ij}^{\tilde{r}_{ij}}\, \tilde{p}_{ij}^{\,t_{ij}}\, p_{ji}^{\tilde{r}_{ji}}$$

and

$$O_\rho(p) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\left| \tilde{r}_{ij} - m_{ij}\, p_{ij} \right| + \left| t_{ij} - m_{ij}\, \tilde{p}_{ij} \right| \right)^\rho$$

respectively. The interested reader should consult the paired comparison literature for details, references, and further discussions of treating ties by explicitly incorporating the probabilities of ties into probabilistic models.

Finally, and perhaps most importantly, different assumptions can (and here frequently do) lead to the same result. Indeed, much of the paired comparison literature on

probabilistic methods concerns showing that, under some set of conditions, making one particular pair of assumptions from Table VI-1 is completely equivalent to making another pair of such assumptions. Indeed, under fairly robust conditions, making any one of assumptions A2, A3, or A4.1 together with making either one of assumptions B1 or B2 is completely equivalent to making any other pair of assumptions in this subset of assumptions from Table VI-1. See the paired combination literature, especially David (1988), Stob (1984), and Stob (1985), for details.

c. An Archetypical Probabilistic Method

Since several of the various possible pairs of assumptions from Table VI-1 turn out to be practically equivalent to each other, only one probabilistic method will be discussed in greater detail here. In terms of Table VI-1, the assumptions underlying this method are assumptions A2 and B1, and the discussion below is based on Jech (1983). The notation introduced above will be used here (definitions for this notation will not necessarily be repeated below); assumptions A2 and B1 will be repeated and explained below.

(1) Assumptions and Goals

Given values for the schedule matrix m and the results matrix r (with ties counting as half a win and half a loss), one goal here is to find values for a matrix p such that:

1) $p \in P$,

2) $\sum_{i=1}^{n} p_{ij} m_{ij} = w_i \qquad i = 1, \ldots, n$, and

3) for all relevant i, j, and k, if neither

$$p_{ij} = 1 \text{ and } p_{jk} = 0$$

nor

$$p_{ij} = 0 \text{ and } p_{jk} = 1,$$

then

$$x_{ik} = x_{ij}\, x_{jk}$$

where

$$x_{ij} = p_{ij}/(1-p_{ij}) = p_{ij}/p_{ji}$$

(if p_{ij} is the probability that T_i would beat T_j in any given game between them, then x_{ij} is the expected number of games between T_i and T_j that T_i would win per each such game that T_j would win). A second goal here is to determine a ranking (say from best to worst) given that values for the p_{ij}'s have been found satisfying the conditions just stated.

(2) Discussion of Assumptions

The first condition above, that $p \in P$, is just a way of starting the assumption that a probabilistic method is to be used.

The second condition above is, of course, assumption B1 on Table VI-1. With p_{ij} being the probability that T_i wins any given game against T_j, the sum

$$\sum_{j=1}^{n} p_{ij} m_{ij}$$

gives the expected number of games that T_i would win against all other teams if it were to play m_{ij} games against T_j for all other j. But it did play m_{ij} such games, it won w_i of them (counting a tie as a half win), and the p_{ij}'s are assumedly to be based on these results. Accordingly, if games are not to be discounted by some external criteria (such as weighting games played earlier in a season less than games played more recently), then it seems quite reasonable to require that the p_{ij}'s satisfy the property that, for each team, its expected number of wins based on the schedule matrix m equals its actual number of wins according to that schedule, i.e. that assumption B1 holds.

The third condition above is, of course, assumption A2 on Table VI-1. As noted above, x_{ij} can be interpreted as the expected number of games between T_i and T_j that T_i would win for each such game that T_j wins. Thus, the assumption that $x_{ik} = x_{ij}x_{jk}$ can be interpreted as assuming that the expected number of wins by T_i per win by T_k in games between them equals the product of the expected number of wins by T_i per win by T_j (in games between T_i and T_j) times the expected number of wins by T_j per win by T_k (in games between T_j and T_k). Jech (1983) gives the following argument to justify the validity of assumption A2:

> Suppose we can compare the objects T_i and T_k only indirectly by comparing T_i with T_j and T_j with T_k and we do it a large number of times, say M. In $M \bullet P_{ij}$ cases, T_i looks better than T_j, and of these $M \bullet p_{ij}$ cases, T_j looks better than T_k exactly $M \bullet p_{ij} \bullet p_{jk}$ times and worse $M \bullet p_{ij} \bullet (1-p_{jk})$ times. Whenever we find T_i better than T_j and T_j better than T_k we

conclude that T_i is better than T_k, but when T_i is found better than T_j and T_j worse than T_k we reserve our judgment about T_i and T_k.

A similar situation arises in the $M \bullet (1-p_{ij})$ cases when T_j is deemed better than T_i. Thus we have $M \bullet p_{ij} \bullet p_{jk}$ cases when T_i is declared better than T_k and $M \bullet (1-p_{ij}) \bullet (1-p_{jk})$ cases when T_k is considered better. It follows that

$$x_{ik} = \frac{M \bullet p_{ij} \bullet p_{jk}}{M \bullet (1-p_{ij}) \bullet (1-p_{jk})} = x_{ij} \bullet x_{jk} .$$

(3) Two Theorems

Jech (1983) proves several theorems based on the assumptions above. His Theorem 1 is stated as Theorem 1 below, and a slightly specialized version if his Theorem 3 is stated as Theorem 2 below.

Theorem 1. If $r \in R_1$, then there exists one and only one probability matrix p whose entries, p_{ij}, satisfy assumptions B1 and A2.

Theorem 2. If $r \in R_1$, if assumptions B1 and A2 are made, and if the schedules for T_i and T_j are such that $m_{ik} = m_{jk}$ for all k other than $k = i$ or $k = j$, then the probability matrix p which is uniquely determined according to Theorem 1 has the properties that:

$$p_{ij} > 0.5 \text{ (or, equivalently, } x_{ij} > 1) \text{ if and only if } w_i > w_j,$$

and

$$p_{ij} = 0.5 \text{ (or, equivalently, } x_{ij} = 1) \text{ if and only if } w_i = w_j.$$

(4) Discussion and Implications of These Theorems

Throughout this subsection, assume that $r \in R_1$ and let p denote the unique probability matrix that follows from making assumptions B1 and A2 according to Theorem 1 above.

First, note that assumption A2 implies assumption A1.1 and so p determines a unique ranking. To see this, add the teams one-at-a-time to an ordered list in the following manner. Order T_1 and T_2 in the obvious manner (with T_1 and T_2 being tied in this order if $p_{12} = 0.5$). Add T_3 ahead of both T_1 and T_2 if $p_{31} > 0.5$ and $p_{32} > 0.5$, add T_3 behind both T_1 and T_2 if $p_{31} < 0.5$ and $p_{32} < 0.5$, and add T_3 to the list in a tied position with T_i if $p_{3i} = 0.5$ for either i=1 or i=2 (by assumption A2, if both $p_{31} = 0.5$ and $p_{32} = 0.5$ then $p_{12} = 0.5$, so this addition is unambiguous). Finally, if $p_{3i} > 0.5$ but $p_{3j} < 0.5$, where

(i,j) is either (1,2) or (2,1), then add T_3 above T_i but below T_j (if $p_{3j} < 0.5$ then $p_{j3} > 0.5$ and, by A1.1, if $p_{j3} > 0.5$ and $p_{ji} > 0.5$ then $p_{ji} > 0.5$ and $p_{ij} < 0.5$, thus this addition is also unambiguous). Adding the remaining teams to this list in the same manner produces a unique ranking in which T_i is ranked ahead of T_j if and only if $p_{ij} > 0.5$ and T_i is ranked as being tied with T_j if and only if $p_{ij} = 0.5$.

Second, it can be shown that an equivalent ranking can be obtained in the following manner. Let

$$\bar{p}_i = \left(\sum_{\substack{j=1 \\ j\neq i}}^{n} p_{ij}\right)/(n-1) \qquad i = 1,...,n \ .$$

Then $\bar{p}_i$ is the expected winning percentage of T_i in a symmetric round robin tournament (i.e., a tournament is which each team plays each other team the same number of times). Ranking the teams by these winning percentages yields the same ranking as the ordered list approach described just above.

Third, note that the ordered list approach above produces a unique ranking, but it does not produce (meaningful) cardinal ratings for the teams according to that ranking. The expected-round-robin-wins approach produces the same ranking, and also produces cardinal ratings (namely, those winning percentages) associated with that ranking. If $r \in R_2$, then it can be shown that another way to produce cardinal ratings for the teams according to this ranking is as follows. Let j_o be such that $p_{ij_o} \geq 0.5$ for all $i \neq j_o$--that is, j_o is (one of) the team(s) ranked last by the ordered list approach above. Suppose that $r \in R_2$. Then $1 \leq x_{ij_o} < \infty$ for all i. Let $s_{j_o} = 1$ and let

$$s_i = x_{ij}s_{j_o} = x_{ij_o} \qquad i = 1,...,n.$$

Call s_i the strength of T_i, rank T_i ahead of T_j if and only if $s_i > s_j$, and rank T_i as being tied with T_j if $s_i = s_j$. Then ranking by these strengths is identical to ranking by round robin winning percentages and to ranking by the ordered list approach described above. Additionally, it can be shown these strengths have the property that

$$p_{ij} = \frac{s_i}{s_i + s_j}$$

for all i and j, and so assumption A3 on Table VI-1 is also satisfied.

In summary, the method proposed by Jech has the following properties: (1) It turns out to be equivalent to several other methods previously proposed in the paired

comparison literature, see Stob (1984), and so a justification for any of these methods is a justification for all of them. (2) It is not restricted to applying only to symmetric round robin comparisons--it can address any schedule matrix m, and it will produce a unique (ordinal) ranking with meaningful cardinal ratings for any m if $r \in R_1$. (3) It allows the result of any comparison to be a tie, not just a win or a loss, and it can address ties in either one of two distinct ways (either by ignoring all tied comparisons or by treating a tie as half a win and half a loss). This can be beneficial because different applications may have different logical bases for treating ties. (4) Finally, by Theorem 2 above, this method necessarily produces the "standard" ranking according to winning percentages whenever it is applied to a symmetric round robin set of comparisons (i.e., one in which $m_{ij} = \overline{m}$ for some positive constant $\overline{m}$ and all i and j such that $i \neq j$).

4. Examples

The reader not interested in ties should skip directly to Section b, below.

a. Some Hypothetical Examples Involving Ties

(1) Some Alternative Approaches For Treating Ties

Two general approaches for considering tied games were described above; namely, either treat ties as half a win and half a loss, or simply ignore all games that result in ties by treating these games as if they were never played. Each of these general approaches are implementable (in some form) in each of the paired comparison methods discussed above, although some of these methods need slight adjustments to treat ties using these approaches (e.g., counting each game as two separate games for methods that use maximum likelihood functions).

Before presenting examples involving ties, two additional approaches that could be used to consider ties are defined as follows. First, ties could be treated as being almost as good as a win. That is, each tied game could count as $1-\epsilon$ of a win and as ϵ of a loss for both of the teams in that game, where ϵ is a sufficiently small positive real number such that using any smaller such number would not change the resultant rankings. (Given the assumptions that only a finite number of teams are being considered and that they play only a finite number of games, only extremely perverse ranking techniques would be able to rank the teams counting ties as $1-\epsilon$ of a win and ϵ of a loss yet not admit the existence of an ϵ' such that the rankings remain constant for all ϵ such that $0 < \epsilon < \epsilon'$. Of course, ϵ' could depend on the number of teams n, the schedule m, and, perhaps, the ranking method

involved.) Symmetrically, ties could be treated as being almost as bad as a loss. That is, each tied game could count as ε of a win and as $1-\varepsilon$ of a loss for both of the teams in that game, where again ε is some sufficiently small positive real number such that any smaller such number would not change the resultant rankings.

Given ε sufficiently small (in the sense discussed just above), let

$$\hat{w}_i = w_i + (1-\varepsilon)\sum_{j=1}^{n} t_{ij}\,, \qquad \hat{l}_i = g_i - \hat{w}_i\,,$$

$$\ddot{w}_i = w_i + \varepsilon \sum_{j=1}^{n} t_{ij}\,, \quad \text{and} \quad \ddot{l}_i = g_i - \ddot{w}_i$$

for i=1, ..., n. Thus, $\hat{w}_i$ and $\hat{l}_i$ are the resulting number of wins and losses, respectively, if ties count almost as much as a win, and $\ddot{w}_i$ and $\ddot{l}_i$ are these wins and losses if ties count almost as much as a loss. Clearly

$$\hat{w}_i + \hat{l}_i = g_i \quad \text{and} \quad \ddot{w}_i + \ddot{l}_i = g_i$$

for each i, and

$$\sum_{i=1}^{n} (\hat{w}_i + \hat{l}_i) = 2\bar{g} \quad \text{and} \quad \sum_{i=1}^{n} (\ddot{w}_i + \ddot{l}_i) = 2\bar{g}$$

where $\bar{g}$ is the total number of games played, i.e.,

$$\bar{g} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} m_{ij}\,.$$

Of course, corresponding forms of these equations also hold if ties are treated as half a win and half a loss, or if all tied games are completely ignored. However, the following inequalities are strict here if any games end in a tie:

$$\sum_{i=1}^{n} \hat{w}_i > \bar{m}\,, \qquad \sum_{i=1}^{n} \hat{l}_i < \bar{m}\,,$$

$$\sum_{i=1}^{n} \ddot{w}_i < \bar{m}\,, \quad \text{and} \quad \sum_{i=1}^{n} \ddot{l}_i > \bar{m}\,,$$

whereas the corresponding relationships are all strict equalities either if ties are counted as half a win and half a loss, or if all tied games are completely ignored. Due to these

inequalities just above, some of the ranking methods described in Sections 2 and 3 are not directly implementable if ties are counted as almost a win or are counted as almost a loss (although some of these methods--and perhaps all of them--could be implemented if suitably significant adjustments to them were made).

Note, however, that ranking by winning percentage is always implementable. Also, if each team plays the same number of games ($g_i = \bar{g}$ for all i and some $\bar{g}$), then counting ties as almost a win and ranking by winning percentage is the same as simply ranking by fewest (actual) losses, where if two teams have the same number of losses then the team with fewer ties (and hence more wins) is ranked ahead of the other. (Teams are equally ranked if they have same numbers of wins, of ties, and of losses). Likewise, if each team plays the same number of games, then counting ties as almost a loss and ranking by winning percentage is the same as ranking by most (actual) wins, where if two teams have the same number of wins then the team with more ties (and hence fewer losses) is ranked ahead of the other. (Again, teams are equally ranked if they have the same numbers of wins, of ties, and of losses.)

The reason for introducing alternative approaches for considering ties here is certainly not to suggest changes in the way that sports leagues treat tied games. Counting ties as half a win and half a loss has been found to be quite reasonable for sports. Instead, the reason for introducing these alternatives is as follows. Paired comparison theory has largely been developed and applied outside of the sports realm. Many additional applications of this theory (outside of sports) may exist, and it does not necessarily follow that the way that sports treats ties is appropriate for applications outside of sports. Accordingly, to assist in examining the possible merits of future applications, it is useful to have several alternative approaches for treating ties available for consideration.

For example, counting ties as almost a win might turn out to be appropriate in applications where consensus is valuable. Converting the terminology away from sports, suppose that paired comparisons of several alternatives are being made by several judges according to several criteria. If each alternative is involved in the same number of comparisons, then an alternative that wins or ties more of these comparisons than any other (and hence losses fewer comparisons than any other) might be viewed as being preferred on a consensus basis over the other alternatives, even if some of these other alternatives won some more but lost many more (and so tied much fewer) of these comparisons. Similarly, if the judging is being done to select an alternative to face some adversaries in future competitions, then it may be desirable to select a robust alternative--one that has

fewer defects for those adversaries to exploit. Such a "safe and sure" (or satisfying) goal might lead to counting each tie as almost a win. Finally, if the judging is being done to weed out inferior alternatives, then ties might not be considered to be much worse than v ins.

Conversely, ties might be considered as being indicative of possessing a blandness which may be almost as bad as being inferior. That is, it may be that being "just as good as..." is not good enough--the selected alternative might have to face adversaries where the only hope of succeeding is to be strictly better in some ways, even if it is strictly worse in many other ways. For example, the developers of new products may want to be able to carve out sufficiently large shares of the market by being better than the competition according to the criteria important for those market shares, even if those products are worse everywhere else (which could be many more places). Blandness might also be considered as being bad in artistic judging or in places where major breakthroughs are being sought. In these cases, it might be reasonable to count each tie as being almost a loss.

(2) Numerical Examples Involving Ties

As can be inferred from the discussions above, the four ways considered here for *treating ties (almost a win, half win plus half loss, almost a loss,* and ignore the game completely) can produce different rankings from each other. The hypothetical example on Table VI-3 illustrates this outcome. In that example, $n = 6$ and $m_{ij} = 6$ for all distinct i and j from 1 through 6, so that each team plays 6 games against each of the other 5 teams for a total of 30 games.

Some points to note concerning the example on Table VI-3 are as follows. First, it involves a symmetric round robin competition (i.e., it does not yield different rankings because some teams play different teams or different numbers of games than other teams). Second, each team wins at least one game and loses at least one game (i.e., it does not yield different rankings due to zeros in numerators or denominators). Third, each team occupies its own place in each ranking (i.e., none of the rankings result in a tie for any position in any ranking). Fourth, not only are the four rankings different, each ranking produces a different winner.

Two of the techniques to treat ties (half win plus half loss, and ignore tied games) are quite robust in the example on Table VI-3 in that (1) the winner by either comes in second by the other, and (2) these two winners come in second or third in the two cases structured to give different winners for the other two techniques. Thus, in addition to the

general acceptance and general applicability of these two techniques, they may be generally robust. This would give another general argument for using them in applications other than those for which there are specific arguments to the contrary.

Table VI-3. An Example Yielding Four Different Rankings From Four Different Ways To Treat Ties In Calculating Winning Percentages

	OPPONENT (Wins/Losses/Ties)						TOTALS		
TEAM	T_1	T_2	T_3	T_4	T_5	T_6	WINS	LOSSES	TIES
T_1	-	0/0/6	0/0/6	0/0/6	5/0/1	3/3/0	8	3	19
T_2	0/0/6	-	0/0/6	0/0/6	2/0/4	4/2/0	6	2	22
T_3	0/0/6	0/0/6	-	0/0/6	0/0/6	2/1/3	2	1	27
T_4	0/0/6	0/0/6	0/0/6	-	0/0/6	3/3/0	3	3	24
T_5	0/5/1	0/2/4	0/0/6	0/0/6	-	6/0/0	6	7	17
T_6	3/3/0	2/4/0	1/2/3	3/3/0	0/6/0	-	9	18	3

	Rankings By Winning Percentage When Considering Each Tie As				Standings By Winning Percentage When Considering Each Tie As			
TEAM	Almost A Win	1/2 Win + 1/2 Loss	Almost A Loss	If Tied Games Never Occurred	Almost A Win	1/2 Win + 1/2 Loss	Almost A Loss	If Tied Games Never Occurred
T_1	3	1	2	2	T_3	T_1	T_6	T_2
T_2	2	2	3	1	T_2	T_2	T_1	T_1
T_3	1	3	6	3	T_1	T_3	T_2	T_3
T_4	4	4	5	4	T_4	T_4	T_5	T_4
T_5	5	5	4	5	T_5	T_5	T_4	T_5
T_6	6	6	1	6	T_6	T_6	T_3	T_6

One such argument to the contrary can be made in voting theory. By making pairwise comparisons of each alternative with each other alternative, a rank-order-input voting process can be viewed as being a symmetric round robin tournament among the alternatives in which each alternative is matched (i.e., plays a game) against each other alternative exactly once. Indeed, if ties are treated as half a win plus half a loss, this is exactly what Copeland's voting method does (see Section B.3.f above). One way to paraphrase an argument made in voting theory is as follows. If, in this "voters tournament," T_i beats or ties every alternative that T_j beats or ties, and T_i beats or ties T_j, then T_i should be ranked at least as high as T_j.

To help see the implications of the argument, consider the hypothetical example on Table VI-4. In that example, n=6, m_{ij}=1 for all distinct i and j from 1 through 6, and the terms "team" and "alternative" are used synonymously. It is not hard to construct a set of voter's preferences (i.e., individual voter's rankings of the six alternatives T_1 through T_6) that result in the outcome displayed on Table VI-4. In the example on that table, T_1 is ranked first by three of the four tie treating techniques tabulated there. However, T_2 beats or ties every alternative that T_1 beats or ties and, not only does T_2 beat or tie T_1, T_2 beats T_1. Therefore, by this paraphrased voting theory argument, T_2 should be ranked at least as high as T_1. Further, since T_2 beat T_1, it can be argued that a tie in ranking between them should be broken in T_2's favor. Indeed, note that if T_i beats or ties every alternative that T_j beats or ties, and T_i beats T_j, then T_i will necessarily be ranked above T_j when these rankings are determined by winning percentages with ties being counted as almost a win. Accordingly, the fourth tie treating technique tabulated on Table VI-4 ranks T_2 in first place, ahead of T_1, and T_2 can be viewed as being a consensus alternative in the sense described above.

Thus, treating ties by counting them as almost a win may be quite appropriate in "voter's tournaments". However, one should be careful about applying arguments taken from voting theory (which may turn out to not actually have meaningful applications within voting theory). For example, another paraphrased (and slightly modified) argument from voting theory is that if every alternative that beats or ties T_i also beats or ties T_j, and T_i beats T_j, the T_i should be ranked ahead of T_j. This corresponds to treating each tie as being almost a loss. In the example on Table VI-4, every alternative that beats or ties T_6 also beats or ties T_2, and T_6 (strictly) beats T_2. Accordingly, treating a tie as almost a loss would rank T_6 ahead of T_2, which may not be what is actually desired.

Table VI-4. An Example Involving Ties and Arguments From Voting Theory

	OPPONENT (W = Win By Team, L = Loss To Opponent, T = Tie)						TOTALS		
TEAM	T_1	T_2	T_3	T_4	T_5	T_6	WINS	LOSSES	TIES
T_1	-	L	W	W	W	L	3	2	0
T_2	W	-	T	T	T	L	1	1	3
T_3	L	T	-	W	L	W	2	2	1
T_4	L	T	L	-	W	W	2	2	1
T_5	L	T	W	L	-	W	2	2	1
T_6	W	W	L	L	L	-	2	3	0

	Rankings By Winning Percentage When Considering Each Tie As				Standings By Winning Percentage When Considering Each Tie As			
TEAM	Almost A Win	1/2 Win + 1/2 Loss	Almost A Loss	If Tied Games Never Occurred	Almost A Win	1/2 Win + 1/2 Loss	Almost A Loss	If Tied Games Never Occurred
T_1	2	1	1	1	T_2	T_1	T_1	T_1
T_2	1	2	6	2t	T_1	T_2	$T_3T_4T_5$	$T_2T_3T_4T_5$
T_3	3t	3t	2t	2t	$T_3T_4T_5$	$T_3T_4T_5$	$T_3T_4T_5$	$T_2T_3T_4T_5$
T_4	3t	3t	2t	2t	$T_3T_4T_5$	$T_3T_4T_5$	$T_3T_4T_5$	$T_2T_3T_4T_5$
T_5	3t	3t	2t	2t	$T_3T_4T_5$	$T_3T_4T_5$	T_6	$T_2T_3T_4T_5$
T_6	6	6	5	6	T_6	T_6	T_2	T_6

b. A Realistic Example Involving College Football

The examples in Section a are hypothetical, are concerned with the treatment of ties, are symmetric round robins, and form rankings using winning percentages (which is quite reasonable for symmetric round robins). The example here is realistic, only incidentally involves ties, is not a symmetric round robin, and illustrates the use of the method described in Section 3.c above. This example concerns the 1989 college football season, which ended with seven bowl games played on January 1, 1990.

Collecting and entering into a computer data for every game played by every college football team in the 1989 season is beyond the intent of this overview. However, a reasonably interesting and useful example can be constructed by considering individually only the fourteen teams that played on January 1, 1990, and aggregating all of the other teams into one notional opponent team, labeled "All Others" in this example. (This is the same approach used by Jech (1983) to consider National Collegiate Athletic Association (NCAA) Division II and Division III teams.) In addition to this major simplifying assumption, the following two assumptions are also made. First, each game is to be counted equally no matter when it was played--who played against whom can make a difference, but whether two teams played each other on the first day of the season or on New Years Day is to make no difference. Second, only the results of the games in terms of who won, who lost, or whether it was a tie, is to be considered--neither the points scored nor any other measure of how well any team played in any game is to be considered (other than indirectly through wins, losses, and ties). Table VI-5 summarizes the 1989 college football season based on these assumptions. The teams are listed on that table in order according to their relative position in the final Associated Press (AP) sports writers opinion poll, with All Others listed last. Data for the schedule matrix, m, and results matrix, r, can be taken directly from Table VI-5.

Table VI-6 gives the relative rankings of these teams according to the AP poll, the United Press International (UPI) coaches opinion poll, the Kendall-Wei method as described in Section 2.b above, and Jech's method as described in Section 3.c above. Note that Jech's method is equivalent to that proposed by Zermelo (1929), Bradley and Terry (1952), and Ford (1957).

The integers on the left side of the AP and UPI polls give the absolute position of the corresponding teams in those polls (e.g., Clemson came in 12th in the AP poll, between 11 Nebraska and 13 Arkansas, and Clemson came in 11th in the UPI polls between 10 Illinois and 12 Nebraska, but Clemson played on New Years Eve, not New

Table VI-5. A Summary of the Common Games Involving Teams That Played in New Years Day (1990) Bowls

	MIA	NDU	FSU	COL	TEN	AUB	MIC	USC	ALA	ILL	NEB	ARK	UVA	OSU	Other	Totals
Miami	-	W	L	-	-	-	-	-	W	-	-	-	-	-	9-0	11-1
Notre Dame	L	-	-	W	-	-	W	W	-	-	-	-	W	-	8-0	12-1
Florida State	W	-	-	-	-	W	-	-	-	-	W	-	-	-	7-2	10-2
Colorado	-	L	-	-	-	-	-	-	-	W	W	-	-	-	9-0	11-1
Tennessee	-	-	-	-	-	W	-	-	L	-	-	W	-	-	9-0	11-1
Auburn	-	-	L	-	L	-	-	-	W	-	-	-	-	W	8-0	10-2
Michigan	-	L	-	-	-	-	-	L	-	W	-	-	-	W	8-0	10-2
U. So. Cal.	-	L	-	-	-	-	W	-	-	L	-	-	-	W	7-0-1	9-2-1
Alabama	L	-	-	-	W	L	-	-	-	-	-	-	-	-	9-0	10-2
Illinois	-	-	-	L	-	-	L	W	-	-	-	-	W	W	7-0	10-2
Nebraska	-	-	L	L	-	-	-	-	-	-	-	-	-	-	10-0	10-2
Arkansas	-	-	-	-	L	-	-	-	-	-	-	-	-	-	10-1	10-2
Virginia	-	L	-	-	-	-	-	-	-	L	-	-	-	-	10-1	10-3
Ohio State	-	-	-	-	-	L	L	L	-	L	-	-	-	-	8-0	8-4
All Others	0W 9L	0W 8L	2W 7L	0W 9L	0W 9L	0W 8L	0W 8L	0W 7L 1T	0W 9L	0W 7L	0W 10L	1W 10L	1W 10L	0W 8L	-	4-119-1
Totals:																
Wins	11	12	10	11	11	10	10	9	10	10	10	10	10	8	4	146
Losses	1	1	2	1	1	2	2	2	2	2	2	2	3	4	119	146
Ties	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	2
Games	12	13	12	12	12	12	12	12	12	12	12	12	13	12	124	294
	MIA	NDU	FSU	COL	TEN	AUB	MIC	USC	ALA	ILL	NEB	ARK	UVA	OSU	Other	Totals

Table VI-6. Several Alternative Rankings of Teams That Played in New Years Day (1990) Bowls

JECH		KENDALL-WEI		AP		UPI	
1	Notre Dame	1	Notre Dame	1	Miami (39)	1	Miami (36)
2	Miami	2	Miami	2	Notre Dame (19)	2	Florida State (7)
3	Colorado	3	Colorado	3	Florida State (2)	3	Notre Dame (6)
4	Tennessee	4	Tennessee	4	Colorado	4	Colorado
5	Illinois	5	Florida State	5	Tennessee	5	Tennessee
6	Michigan	6	Illinois	6	Auburn	6	Auburn
7	Alabama	7	Michigan	7	Michigan	7	Alabama
8	Auburn	8	Alabama	8	Southern Cal.	8	Michigan
9	Southern Cal.	9	Auburn	9	Alabama	9	Southern Cal.
10	Florida State	10	Southern Cal.	10	Illinois	10	Illinois
11	Nebraska	11	Nebraska	11	Nebraska	12	Nebraska
12	Ohio State	12	Arkansas	13	Arkansas	13	Arkansas
13	Arkansas	13	Virginia	18	Virginia	15	Virginia
14	Virginia	14	Ohio State	24	Ohio State	21t	Ohio State
15	All Others	15	All Others				

Years Day (1990), and is not individually identified in these rankings). The number in parentheses on the right gives the number of first place votes that these teams received in these polls (only the top three teams received any first place votes). The ranking in each of these polls is determined using a truncated Borda voting method. Each AP voter selects 25 teams to rank order, and each such voter's first place choice receives 25 points, each voter's second place choice receives 24 points, and so on through each 25th place choice which receives 1 point. Each UPI voter selects 20 teams to rank order in the same manner (except that the points awarded range from 20 points for each first place through 1 point for each 20th place).

The rankings for Jech's method and the Kendall-Wei method are determined using values for m and r obtained from Table VI-5 (the tie involving USC is counted as half a win and half a loss). A modified version of a computer program originally developed by Miller and Palocsay (1985) was used to perform the calculations. Note that $r \in R_2$ in this example, so these are the unique rankings produced by these methods.

Some points to note concerning this example are as follows.

First, the relative rankings of these fourteen teams by the Jech or Kendall-Wei method could change if all of the college football teams were individually considered. This is not unlikely for lower ranked teams or for closely ranked teams--the relative strengths according to Jech's method (denoted by s_i for i=1,...,n in Section 3.c), renormalized so that Notre Dame has a strength of 1.000, are given on Table VI-7. (That table also repeats the rankings by the Kendall-Wei method and gives the relative AP and UPI rankings of the fourteen individual teams.) As can be seen from Table VI-7, neither Miami nor Colorado are close to Notre Dame in relative strength by Jech's method (similar data applies for the Kendall-Wei method). Also, none of these three teams lost any games to any of the also-rans aggregated into the notional All Others team. Thus, with one exception, it seems quite unlikely that these top three places would change if every college football team were individually considered--the one exception is that Miami and Colorado might switch places (moving Miami to third and Colorado to second) since they are very closely ranked to each other.

Second, it was noted above that the Kendall-Wei method considers the strength-of-schedule as it concerns wins over strong teams versus wins over weak teams, but not as it concerns losses to weak teams versus losses to strong teams. In comparison, Jech's method considers all aspects of the strength of the schedule. This characteristic is evident on Tables VI-5, 6, and 7 in that the differences between their rankings involve Kendall-

Table VI-7. Normalized Strengths By Jech's Method and Relative Rankings From Table VI-6

TEAM	Normalized Strengths	Relative Ranking †			
		Jech	Kendall	AP	UPI
Notre Dame	1.000	1	1	2	3
Miami	.580	2	2	1	1
Colorado	.553	3	3	4	4
Tennessee	.195	4	4	5	5
Illinois	.169	5	6	10	10
Michigan	.168	6	7	7	8
Alabama	.138	7	8	9	7
Auburn	.113	8	9	6	6
Southern Cal.	.108	9	10	8	9
Florida State	.093	10	5	3	2
Nebraska	.052	11	11	11	11
Ohio State	.027	12	14	14	14
Arkansas	.022	13	12	12	12
Virginia	.021	14	13	13	13
All Others	.002	15	15	(12)	(11)

† These rankings are relative to the fourteen individual teams listed here. The highest AP and UPI ranked team not listed (Clemson) had an absolute AP ranking of 12 and UPI ranking of 11, as noted in parentheses.

Wei's higher ranking (compared to Jech) of one team (Florida State) that lost two games to the notional "weak" opponent team, and Jech's higher ranking (compared to Kendall-Wei) of one team (Ohio State) that lost no games to the notional "weak" opponent.

Third, a common argument concerning mathematical techniques is that a (usually more complex) new technique might be theoretically better, but it might yield results that are insignificantly different from those produced by the current approach. This is clearly not the case here. Of course, just because a new technique produces significantly different results in a sports application doesn't necessarily mean that it even has appropriate applications outside of sports, let alone has meaningful applications that would produce significantly different results than currently used approaches. However, the real issue here may be the existence of appropriate applications, not the significance of the impact of using these techniques if such applications were found.

Fourth, a common argument concerning all modeling and abstract representation techniques is that a major purpose of these techniques is to gain insight and understanding, as opposed to obtaining a definitive answer from (say) one computer run using such a technique. This being the case, one might question whether ranking techniques are too sterile to provide any additional insight or understanding. In general, one might be able to gain such insight or understanding by doing sensitivity analyses. For example, what would happen if the outcome(s) of one (or a few) of the paired comparisons were reversed (win-to-loser and vice versa)? In the particular example here, one could try to isolate (and support or discredit) possible reasons why the opinion polls differ so much from, say, Jech's ranking, which takes full account of the strength of the schedule.

Was it because the voters in the polls placed greater weight on more recent games? This could be tested by making relatively simple modifications to Jech's method to place unequal weights on different games and then trying to find a plausible time-discounting scheme that produces results comparable to the polls.

Was it because the voters in the polls considered the scores of the games rather than just who won or who lost? This could be tested by modifying Jech's method to consider game scores and/or by comparing the polls with results produced by other methods that directly consider such scores.

Was it because those voters felt that Notre Dame received unfair rewards in previous years, which they were not going to allow to happen this year? Voters' feelings are hard to test, especially after the fact. However, results and polls from past years could be analyzed to see if any given team, or set of teams, were generally overrated or

underrated in those polls as compared to the more objective rankings produced by Jech's method (or by other methods described above).

Or was it because the polls ranked Colorado first, Miami second, Michigan third, and Notre Dame fourth before the New Year Day bowl games, and the voters did not consider Notre Dame's victory over Colorado in their bowl game to be significant enough to move Notre Dame ahead of Miami? This conjecture has two parts: the first concerns the significance of Notre Dame bowl victory our previously unbeaten and top-ranked Colorado, and the second concerns the appropriateness of the pre-bowl polls, which can be tested by comparing these polls to rankings that Jech's method would produce given the games up to but not including the bowls. Such a comparison is shown on Table VI-8. Since Colorado was undefeated, and since Nebraska's one pre-bowl loss was to Colorado, Jech's method necessarily ranks Colorado first and Nebraska second (note that if r^- is the corresponding pre-bowl results matrix, then $r^- \in R_1$ but $r^- \notin R_2$). Clearly, more analysis could be done concerning this example. However, the point here is not to reach a conclusion as to why the polls ranked Miami ahead of Notre Dame, nor who really should have been No. 1 (according to various criteria) in college football this year, nor even that this is an important issue in the first place. Instead the point is that paired comparison methods are not so sterile as to preclude using them in comparative analyses, and the example here shows (at a minimum) how such an analysis could be begun.

Finally, as an aside, college football was used as an example both here and in Jech (1983). One should not conclude that this means that one of the most appropriate applications of paired comparison theory in sports is to the post bowl determination of college football rankings. Indeed, one could argue that no post (complete) season application could be as important as an application that would affect who would (continue) to play in what games. For example, paired comparison theory could have a quite useful and relatively quite important application in determining which college teams (e.g., in basketball) should be selected to participate in the NCAA post (regular) season tournaments, and to help determine how such teams should be seeded. This theory could also be used to help determine which of, say, several National Football League teams with identical won-lost records should make the playoffs when, as frequently happens, more teams are tied with identical records than there are (wild card) places to be filled by these teams.

Table VI-8. Pre-Bowl and Final 1990 College Football Rankings (Jech and AP)

Team	Pre-Bowl Record	Pre-Bowl		Bowl Result (Opponent)	Final	
		Jech Rank	AP Rank		Jech Rank	AP Rank
Colorado	11-0	1	1	Lost (NDU)	3	4
Nebraska	10-1	2	6	Lost (FSU)	11	11
Notre Dame	11-1	3	4	Won (COL)	1	2
Michigan	10-1	4	3	Lost (USC)	6	7
Illinois	9-2	5	11	Won (UVA)	5	10
Miami	10-1	6	2	Won (ALA)	2	1
Alabama	10-1	7t	7	Lost (MIA)	7	9
Tennessee	10-1	7t	8	Won (ARK)	4	5
Auburn	9-2	9	9	Won (OSU)	8	6
Southern Cal.	9-2-1	10	12	Won (MIC)	9	8
Florida State	9-2	11	5	Won (NEB)	10	3
Ohio State	8-3	12	21	Lost (AUB)	12	24
Arkansas	10-1	13	10	Lost (TEN)	13	13
Virginia	10-2	14	15	Lost (ILL)	14	18

5. Some Characteristics Concerning General Applicability With Emphasis on Combat Analyses

Like voting theory, paired comparison theory does not appear to have been applied in major defense analyses. Unlike voting theory, it may not be immediately obvious whether, in any particular situation, paired comparison theory can be profitably applied. However, increased dissemination of the existence of this theory combined with sufficient ingenuity in structuring analyses may lead to such applications. To assist in finding such applications, the following framework is proposed.

a. A General Framework for Applying Paired Comparison Theory

Paired comparison theory involves making and evaluating comparisons in situations that have the following four characteristics. First, each comparison is made between two alternatives (not among three or more). Second, different pairs of alternatives can (optionally) be compared different numbers of times, including (optionally) no comparisons being made of some pairs. Third, only which alternative won and which lost each comparison (if a tie did not occur) is important, the size of the victory in terms of any magnitude that could be associated with the comparison is irrelevant. Fourth, the comparisons need not be consistent either in that alternative A could win a comparison over B, B over C, yet C could win over A, or in that if A and B are compared several times, A might win some of these comparisons but lose others.

In a reasonable sense, this fourth characteristic either holds or it doesn't, and paired comparison theory has very little to offer concerning the analyses of situations for which it does not hold. However, these four characteristics are clearly not independent, and situations possessing the first and third characteristics above are likely to have the fourth also.

The third characteristic is not necessarily a "definitely holds or definitely does not hold" characteristic. It clearly holds if the comparisons have no magnitudes associated with them. However, many comparisons have (and many of the rest can be easily modified to have) such magnitudes. The issue therefore concerns the relative significance of these magnitudes in the analysis being made. If the relative sizes of these magnitudes across comparisons are quite important, then paired comparisons theory (as described here) may have little to offer. If the sizes of these magnitudes provide insight and explanation, but only within any given comparison, and they are not necessarily meaningful across comparisons, then paired comparison theory might be a useful tool.

The second characteristic clearly either holds or doesn't hold. If it holds, then paired comparison theory is relatively more likely to be applicable (but it is not necessarily so); while if these don't hold, paired comparison theory is relatively less likely to be applicable (but it still might be so).

The first characteristic may be more important than it first appears concerning the applicability of paired comparison theory. For example, it might seem unimportant because any ranking of a alternatives (perhaps with associated magnitudes) can be converted into a(a–1)/2 comparisons, one each for each different pair of alternatives. More generally, v rankings of a alternatives can be converted into va(a–1)/2 comparisons, where each different pair of alternatives is compared v times (this is essentially what Copeland's method does in voting theory). However, if this is done when v=1, then characteristics 2 and 4 do not apply, and paired comparison theory has essentially nothing to offer. (If characteristic 3 applies, then the situation is trivial--the ranking gives the answer. If characteristic 3 does not apply, then analyses that directly address the relevant magnitudes may be useful. However, either way, paired comparison theory contributes nothing here.) If this is done when v>1, then characteristic 2 does not apply, and voting theory (but not paired comparison theory) could apply, depending on characteristic 3.

b. Discussion of Potential Combat Applications in Terms of This Framework

The purpose of this section is to discuss potential applications of paired comparison theory to combat related analyses in terms of the four characteristics described in Section a above. Non-combat defense applications are, of course, possible, and the interested reader could structure such applications in terms of these characteristics if desired. Some of the comments below apply to both combat and non-combat applications.

(1) Making Comparisons in Pairs

At the outset it should be noted that the standard procedure of comparing a set of alternative forces by evaluating each force against a canonical enemy threat does not constitute making paired comparisons as defined here. That is, while each alternative force can be evaluated by pairing it in simulated combat against the canonical enemy force, the alternatives being compared in such an analysis are the different friendly forces, not a friendly alternative and an enemy alternative.

However, it may be that, due to personnel, time, equipment, and/or range constraints, the alternatives must be evaluated in separate groups, that no more than a

certain (maximum) number of alternatives can be evaluated in any one group, and that cross-group comparisons are not valid. (Such cross-group comparisons might not be valid due to changes in personnel, equipment, and/or range conditions between groups.)

For example, if eight alternatives are to be evaluated, but no more than four can be evaluated in any one comparable group, then the following procedure is possible. Alternatives 1, 2, 3, and 4 could be evaluated in one group, alternatives 5, 6, 7, and 8 could be evaluated in a second group, and then two from each, say 3, 4, 5, and 6, could be evaluated in a third group. Such an approach would produce the following paired comparisons:

Pair of Alternatives	Number of Comparisons Made	Pair of Alternatives	Number of Comparisons Made
1,2	1	4,5	1
1,3	1	4,6	1
1,4	1	5,6	2
2,3	1	5,7	1
2,4	1	5,8	1
3,4	2	6,7	1
3,5	1	6,8	1
3,6	1	7,8	1
		All Other Pairs	0

This same point clearly applies if the alternatives are being compared purely by human judgment and the situation is sufficiently complex that no one judge can adequately address all of the alternatives. The example just above would apply if there were three judges, eight alternatives, and each judge could address any four but no more than four of these alternatives. In the extreme, each judge may not be able to address more than two alternatives, which would satisfy the first condition directly.

Finally, while war games and computer models frequently have been used to evaluate alternative forces and tactics against canonical threats, the lack of such threats in the future may preclude use of this technique. Of course, the goal of spending defense dollars wisely remains. However, testing the wisdom of such spending by evaluating alternatives against an arbitrarily specified threat (or a small set of such threats) may no longer be reasonable. Instead, alternatives could be evaluated by matching alternatives in pairs directly against each other using a war game or computer model. This approach is somewhat novel, but it has been suggested before (Grotte and Brooks, 1983), and it may prove to be an appropriate threat-independent way to evaluate alternative future defense expenditures.

(2) Unequal Numbers of Comparisons

Analyzing comparisons in groups as described above would generally result in unequal numbers of comparisons of various pairs of alternatives. In the example above, the pairs (3,4) and (5,6) are compared twice, the pairs (1,2), (1,3), (1,4), (2,3), (2,4), (3,5), (3,6), (4,5), (4,6), (5,7), (5,8), (6,7), (6,8), and (7,8) are compared once, and the remaining 12 pairs are not compared at all.

If comparisons are being made in pairs by matching two alternatives against each other in a fully automated combat model, then there is no inherent reason why all pairs would not be compared an equal number of times. However, if such comparisons are being made in an interactive war game, a field exercise, or any other human-intensive processes, then limitations on personnel, time, equipment, and/or range facilities, or (conversely) special interest by the personnel involved in repeating comparisons of selected pairs, could lead to unequal numbers of comparisons of pairs of alternatives.

(3) Irrelevance of Associated Magnitudes

If comparisons are made in groups as described above, then cross-group comparisons would not be valid precisely when associated magnitudes are not comparable across groups. As indicated above, such magnitudes may not be comparable across groups because key personnel (players, judges/controllers, etc.) change between groups, or because there are sufficient changes in the setting (equipment, range conditions, etc.) that cross-group comparisons are not valid. Note that an important special case here occurs when these groups are all of size two, that is, the alternatives are being compared in pairs and, due to changes in the personnel or conditions involved, magnitudes associated with a comparison of one pair of alternatives are not necessarily commensurable with magnitudes associated with another comparison of that pair or with any comparison of any other pair.

There are several other conditions that would lead to considering only wins, losses, and ties, but not associated magnitudes, in making comparisons. For example, the goal in a war game may be to hold the enemy's advance to a certain amount, or to penetrate the enemy's defenses in at least a certain number of places. That is, the goal is to satisfy a set of criteria, not to maximize any particular war game output. In such a case, the magnitudes of associated outputs would not be directly relevant to any comparisons being made.

A related set of examples concerns Monte Carlo models. One might desire to run sufficiently many trials of a Monte Carlo model in order to achieve a certain level of statistical confidence in the results. One way to do this is to set a goal (e.g., to keep major

bases or ships from being put out of action with probability greater than some specified amount) and to run sufficiently many trials to have high confidence in knowing whether or not a particular defense against a particular attack was capable of meeting that goal. In such a case, the comparisons should depend on whether or not that goal was achieved, not on associated magnitudes (such as how many attacking and defending weapons systems were killed).

Finally, if the associated magnitudes (or measures of the importance of these magnitudes) are determined directly by human judgment, then differences in magnitudes that seem large at the beginning of a set of comparisons may seem small at the end, and (in different cases) vice versa. Such order effects can be addressed by doing many more comparisons; but they are not important (and so need not be addressed) if only the direction (i.e., bigger, smaller, or the same) is to be counted, not the size of the differences in magnitudes.

(4) Inconsistent Comparisons

A comparison between alternative A and alternative B that favors alternative A is clearly inconsistent with another such comparison that favors alternative B, and can be said to be inconsistent with two comparisons, one between B and C and one between C and A, if those two comparisons favor B and C, respectively. As stated above, paired comparison theory is designed to handle such inconsistencies; the question is whether such inconsistencies would occur so that paired comparison theory would be a useful tool.

Such inconsistencies might occur, for example, if the comparisons were made in groups (including groups of size 2) as described above, and if the changes in personnel and/or conditions involved were sufficiently significant to reverse the outcomes of some comparisons.

Such inconsistencies might also occur in the following situation. Suppose a Monte Carlo model is used and insufficiently many trials for statistical significance are run. (For example, using Monte Carlo models in interactive war games frequently involves making only one trial per identical set of conditions.) The inherent randomness involved in such a use of Monte Carlo models could easily lead to the inconsistencies described here.

6. Annotated Bibliography

a. Recommended Reading

The following references, in the following order, are recommended for those who want to read into the literature on paired comparisons.

A good place to start is

T. Jech, "The Ranking of Incomplete Tournaments: A Mathematician's Guide to Popular Sports," *Amer. Math. Monthly*, Vol. 90 (1983), 246-266.

In this paper, Jech presents a clean and clear exposition of the method described in Section 3.c. Theorems are well explained and proofs are nicely set off to assist the reader in understanding the concepts without getting lost in the details or rigor. However, one reason that this paper may be so clean is that it contains no references to any previous work nor does it mention any other non-trivial analytical methods; this is noted by

M. Stob, "A Supplement to 'A Mathematician's Guide to Popular Sports,'" *Amer. Math. Monthly*, Vol. 91 (1984), 277-282,

which should be read second. By relating Jech's paper to previous research, Stob provides a brief but useful and coherent overview of probability models in paired comparison theory.

Third, the interested reader should examine

M. Stob, "Rankings from Round-Robin Tournaments," *Management Sci.*, Vol. 31 (1985), 1191-1195.

In this paper, Stob provides a brief but useful overview of deterministic combinatorial methods in paired comparison theory, and he points out that this is only one of two general approaches--the probability model approach being the other. Curiously, in this paper, Stob references neither his previous paper (cited just above) nor Jech's paper, and he does not reference Ford (1957), which is one of the (at least) three previously published research papers that contain results that Jech rediscovered. (He does cite the other two, but there is a typographical error in the reference to Zermelo--the correct date of that publication is 1929, not 1926.) Since this paper by Stob is oriented toward deterministic combinatorial methods, not probability models, these omissions are quite understandable, but they inhibit exploration of the literature by a new reader. In this paper, Stob reviews a paper by Goddard (1983), but that paper is not a prerequisite for Stob's paper. Another feature of Stob's paper is that it points out some characteristics of deterministic combinatorial methods that can reasonably be viewed as also being major flaws of these methods.

Fourth, the interested reader should examine

H.A. David, "Ranking the Players in a Round Robin Tournament," *Rev. Int. Statist. Inst.*, Vol. 39 (1971), 137-147.

This paper also points out some questionable characteristics (i.e., flaws) of deterministic combinatorial methods, it points out a potentially serious problem with probability models that use maximum likelihood values, and it provides a nice transition from the other recommended references to

H.A. David, "The Method of Paired Comparisons," *Second Edition, Revised*, London: Oxford University Press, 1988,

which should be (at least) skimmed and judiciously examined next. This text seems to be the only recent book on the subject, it is quite good, it covers a broad range of topics concerning paired comparison theory, and, with a few notable exceptions (listed in Section b, below), it contains an outstanding bibliography for paired comparisons.

After examining this book, the reader might continue in any of several different directions. One such direction is to consider probability models based on random strengths with distributions other than the extreme value distribution or the normal distribution. An excellent start is this direction is given in

M.R. Frey, "Predicting the Outcome of Sporting Events: An Application of Probability Theory," Student Paper for OR 291, George Washington University, Washington DC, April 1984.

A reader particularly interested in sports should see

S.P. Ladany and R.E. Machol (eds.), "Optimal Strategies in Sports," Amsterdam, New York, and Oxford: North-Holland Publishing Company, 1977, and

R.E. Machol, S.P. Landany, and D.G. Morrison (eds.), "Management Science in Sports," Amsterdam, New York, and Oxford: North-Holland Publishing Company, 1976,

and should consider the papers listed under Recreation and Sports (classification category 840) in the Subject Index of

K.T. Marshall and F.P. Richards (eds.), "The OR/MS Index 1952-1976," Providence: The Institute of Management Sciences, and Baltimore: Operations Research Society of America, 1978,

J.W. Tolle and R.E. Stone (eds.), "The OR/MS Index 1976-1981," Providence: The Institute of Management Sciences, and Baltimore: Operations Research Society of America, 1983, and

J.W. Toole (ed.), "The OR/MS Index 1982-1987," Providence: The Institute of Management Sciences, and Baltimore: Operations Research Society of America, 1988.

b. Omissions From David's References

In his book, David (1988) cites almost 300 references, one of which (Davidson and Farquhar, 1976) gives an extensive bibliography up to 1976. Accordingly, one might expect every major paper on paired comparison theory published between 1976 and 1988 to be cited there. The following are not.

I. Ali, W.D. Cook, and M. Kress, "On the Minimum Violations Ranking of a Tournament," *Management Sci.*, Vol. 32 (1986), 660-672.

S.T. Goddard, "Ranking in Tournaments and Group Decisionmaking," *Management Sci.*, Vol. 29 (1983), 1384-1392.

T. Jech, "The Ranking of Incomplete Tournaments: A Mathematician's Guide to Popular Sports," *Amer. Math. Monthly*, Vol. 90 (1983), 246-266.

M. Stob, "A Supplement to 'A Mathematician's Guide to Popular Sports,'" *Amer. Math. Monthly*, Vol. 91 (1984), 277-282.

M. Stob, "Rankings from Round-Robin Tournaments," *Management Sci.*, Vol. 31 (1985), 1191-1195.

c. References and Representative Bibliography

The publications listed below include all those cited in this section (on paired comparison theory) above, and includes a representative sample of related publications in this area.

I. Ali, W.D. Cook, and M. Kress, "On the Minimum Violations Ranking of a Tournament," *Management Sci.*, Vol 32 (1986), 660-672.

R.A. Bradley, "Science, Statistics, and Paired Comparisons," *Biometrics*, Vol 32 (1976), 213-232.

R.A. Bradley and M.E. Terry, "The Rank Analysis of Incomplete Block Designs. I. The Method of Paired Comparisons," *Biometrika*, Vol 39 (1952), 324-345.

H.A. David, "Ranking the Players in a Round Robin Tournament," *Rev. Inter. Statist. Inst.*, Vol 39 (1971), 137-147.

H.A. David, "The Method of Paired Comparisons," *Second Edition, Revised*, London: Oxford University Press, 1988.

R.R. Davidson, "On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments," *J. Amer. Statist. Assoc*, Vol 65 (1970), 317-328

R.R. Davidson and P.H. Farquhar, "A Bibliography on the Method of Paired Comparisons," *Biometrics*, Vol 32 (1976) 241-252.

O. Dykstra, "Rank Analysis of Incomplete Block Designs Method of Paired Comparisons Employing Unequal Repetitions on Pairs," *Biometrics*, Vol 16 (1960), 176-188.

S.E. Feinberg and K. Larntz, "Log Linear Representation for Paired and Multiple Comparisons Models," *Biometrika*, Vol 63 (1976), 245-254.

L.R. Ford, "Solution of a Ranking Problem from Binary Comparisons," *Amer. Math. Monthly*, Vol 64 (1957), 28-33.

M.R. Frey, "Predicting the Outcome of Sporting Events: An Application of Probability Theory," Student Paper for OR 291, George Washington University, Washington DC, April 1984.

S.T. Goddard, "Ranking in Tournaments and Group Decisionmaking, *Management Sci.*, Vol 29 (1983), 1384-1392

J.H. Grotte and P.S. Brooks, "Measuring Naval Presence Using Blotto Games," *Int. J. of Game Theory*, Vol 12 (1983), 225-236.

F. Harary and L. Moser, "The Theory of Round Robin Tournaments," *Amer. Math. Monthly*, Vol 73 (1966), 231-246

D.A. Harville, "Football Ratings and Predictions via Linear Models," *Proceedings of the Amer. Statist. Assoc.* 1978, 74-82.

J. Horen and R. Riezman, "Comparing Draws for Single Elimination Tournaments," *Oper. Res.*, Vol 33 (1985), 249-262.

T. Jech "The Ranking of Incomplete Tournaments: A Mathematician's Guide to Popular Sports," *Amer. Math. Monthly*, Vol 90 (1983), 246-266.

M.G. Kendall, "Further Contributions to the Theory of Paired Comparisons," *Biometrics*, Vol 2 (1955), 43-62.

K.J. Koehler and H. Ridpath, "An Application of a Biased Version of the Bradley-Terry-Luce Model to Professional Basketball Results," *J. of Mathematical Psychology*, Vol 25 (1982), 187-205.

S.P. Ladany and R.E. Machol (eds.), "Optimal Strategies in Sports," Amsterdam, New York, and Oxford: North-Holland Publishing Company, 1977.

R.E. Machol, S.P. Ladany, and D.G. Morrison (eds.), "Management Science in Sports," Amsterdam, New York, and Oxford: North-Holland Publishing Company, 1976.

A. Miller and S. Palocsay, "A Comparison of Ranking Methods in Pairwise Competitions," Student Paper for OR 291, George Washington University, Washington DC, December 1985.

F. Mosteller, "Remarks on the Method of Paired Comparisons: I. The Least Squares Solution Assuming Equal Standard Deviations and Equal Correlations," *Psychometrika*, Vol 16 (1951), 3-9.

C. Ramanujacharyulu, "Analysis of Preferential Experiments," *Psychometriku*, Vol 29 (1964), 257-261

A. Rubinstein, "Ranking the Participants in a Tournament," *SIAM J. Appl. Math.*, Vol 38 (1980), 108-111.

J.S. Sagarin and W.L. Winston, "The Use of Exponential Smoothing to Forecast the Outcome of Basketball and Football Games," School of Business, Indiana University, Bloomington IN, October 1983, 1-9.

J.H. Smith, "Adjusting Baseball Standings for Strength of Teams Played," *The Amer. Statistician*, Vol 10 (1956), 23-24.

R.T. Stefani, "Football and Basketball Predictions Using Least Squares," *IEEE Transactions on Systems, Man, and Cybernetics*, February 1977, 117-121.

M. Stob, "A Supplement to 'A Mathematician's Guide to Popular Sports,'" *Amer. Math. Monthly* (1984), 277-282

M. Stob, "Rankings from Round-Robin Tournaments," *Management Sci.*, Vol 31 (1985), 1191-1195.

M. Thompson, "On Any Given Sunday: Fair Competitor Orderings with Maximum Likelihood Methods," *J. of the Amer. Statist. Assoc.*, Vol 70 (1975), 536-541.

P. Tryfos, S. Casey, S. Cook, G. Leger, and B. Pylypiak, "The Profitability of Wagering on NFL Games," *Management Sci.*, Vol 30 (1984), 123-135.

R.C. Vergin and M. Scriabin, "Winning Strategies for Wagering on National Football League Games," *Management Sci.*, Vol 24 (1978), 809-818.

E. Zermelo, "Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung," *Math. Z.*, Vol 29 (1929), 436-460.

VII. OBSERVATIONS

A. DELPHI

The Delphi method is defined by a few loosely specified guidelines for designing a group data collection process. The resulting heterogeneity among nominal "Delphi" studies suggests that, for the purpose of this review and for all practical purposes, there is no "Delphi method" *per se*. Instead, there is a principle that attention can and should be paid to minimizing the effect of deleterious group social processes on group data collection. The guidelines that define the Delphi methodology represent a few steps that can be taken in this regard, although there are conditions under which they probably should not be taken. As we imply, there also are other steps that can be taken to improve data collection process under specified conditions.

However, despite the substantial history of Delphi method applications, there is not a great deal of empirical support for the important statements made about the Delphi method by either its supporters or its detractors. *What steps have actually been shown to improve or worsen the quality of group judgment data? Under what conditions are such steps effective?* More often than not, we do not know the answer to these questions, and, as a result, the skilled analyst is left with not much more than his experience and intuition when it comes to designing a group data collection process for a particular application.

B. THE ANALYTICAL HIERARCHY PROCESS

The AHP has been subjected to a number of criticisms in the literature that cast doubt on the method as it was originally stated in Thomas Saaty's book, *The Analytic Hierarchy Process* (McGraw Hill, 1980). First is the argument that the conventional AHP question, "Which is more important, A or B, and by how much?" is ambiguous. The question does not provide clear referents upon which to base judgments of relative importance, that is, relative importance in terms of what units and what statistics (e.g., maximum, minimum, average). Simple modifications to address this problem have been suggested by the authors of recent papers. Second is the argument that AHP is a manifestation of multiattribute utility theory, but requires restrictive assumptions that

usually are not satisfied in practice. Finally, arguments have been advanced from psychological measurement theory that applications of the AHP rest on untested critical assumptions regarding the form of the human judgment processes. The AHP itself does not provide for evaluating these assumptions. Since most practitioners are unaware that these assumptions are being made, they implicitly assume them to be satisfied. Furthermore, evidence from the literature on human judgment processes suggests that some assumptions about how judges respond to AHP-type relative magnitude questions will not be satisfied. The most immediate consequence of these arguments is that *it is invalid to give a quantitative meaning to AHP results beyond the ordering they indicate for the alternatives. That is, it is not meaningful to interpret the differences or ratios between AHP-derived weights*. Future research may weaken even the limited interpretability allowed by the current argument.

In addition to the assumptions discussed above, we have observed that many AHP practitioners are unaware of the assumptions underlying the AHP stated by Thomas Saaty as part of his original development of the method. Key among these are the assumptions of independence between elements subordinate to a common parent and of criteria from subordinate alternatives. Violation of these assumptions requires specific corrective procedures as described by Saaty in his book, *The Analytic Hierarchy Process* (McGraw Hill, 1980) and in other writings. As demonstrated by Saaty and others, inattention to these assumption can badly compromise the weights produced by an AHP analysis. Most available software does not make provisions for testing these assumptions as part of the normal course of analysis. As a consequence, many practitioners either using this software or operating in "cookbook fashion" from basic AHP texts ignore what Thomas Saaty himself regards as basic diagnostic procedures. Other practitioners may be aware of these assumptions, but assume them to be satisfied without conducting even a cursory empirical evaluation.

C. SUBJECTIVE TRANSFER FUNCTION METHOD

Subjective transfer function is a promising method for collecting and representing expert knowledge of complex systems. However, the STF method has not received critical attention from the analytical community, which is important in identifying strengths and weaknesses, attracting new applications and promoting continued development.

The developers of the STF method have introduced us to valuable concepts from psychological measurement and judgment theory. This is an significant contribution

separate from the STF method because of the issues they raise about collecting and interpreting judgment data. A more thorough understanding of judgment methods in terms of these issues can only improve the use and validity of analyses based on ratings-type data.

D. UTILITY THEORY

Utility theory and its variants have proven to be a uniquely valuable component of the foundation of many of the analytical sciences. Its axiomatic nature has contributed rigor to the powerful theories that incorporate its precepts. Correctly satisfying these axioms, however, when developing actual utility functions, is not an easy matter. Not only must a respondent, in general, evaluate many tradeoff situations, often in the form of non-intuitive lotteries, but the analyst also must be careful to present these situations in a way that minimizes bias. There is a substantial literature that indicates that neither of these obstacles is easy to overcome. As a result, there tend to be more conceptual applications of utility theory in defense analysis than otherwise specific evaluative applications. These factors suggest that utility theory only should be used by analysts literate in the theory and the issues surrounding implementation and only in situations where respondents have the time and the understanding to fully participate in the development of utility functions.

E. VOTING THEORY

Where voting theory applies, it is obviously applicable, and forcing it to fit where it doesn't obviously apply does not appear to be useful. There are many different voting methods, and the choice of which to use can be so important that, given a fixed set of voters' preferences, choosing different methods can result in different winners. Various voting methods have various properties. In a sense, none is perfect. However, based on their properties and on the voting situations involved, some may be deemed better than others. A commonly used method, plurality voting, may be one of the worst for all situations. Plurality voting may be used so frequently because of its extreme simplicity, but this extreme simplicity can result in serious flaws.

F. PAIRED COMPARISONS

Paired comparison theory is not well known outside of a relatively small group of theorists. It has been applied, but not extensively, and apparently not in major defense analyses. Several paired comparison methods exist, and one (which has been regularly

rediscovered) seems to have generally more desirable properties than the others. Specific applications, however, may have specific characteristics that are more suitably addressed by one of the other methods. More widespread dissemination of this theory, together with sufficient ingenuity in adapting it to particular situations, could lead to many additional applications.

G. GENERAL

Common to criticisms of many of the methods we have reviewed above are several recurring themes. Most fall under the mistaken belief that judgment methods are inexpensive and easy to employ while yielding quality results. Two important practices that result are that critical assumptions are untested (practitioners frequently are not schooled well enough in the theory underlying the methods to understand what the critical assumptions are); and judgment-eliciting questions are ambiguous. Question ambiguity gives respondents considerable latitude to impose their own interpretations of what is being asked. As a result, the quality of judgment data is compromised to an unknown degree. Some corrective measures are relatively easy to take. Analysts should pretest questions and questionnaire designs to screen for poorly phrased or ambiguous items and for undesired effects of particular question orders. Other corrective measure require more effort. In particular, analysts should attain a thorough working understanding of the methods they use and should keep their knowledge base current by following developments in the literature.

Developments in the psychology of judgment and expertise suggest that analysts also should give attention to questions of how much weight should be given expert judgments. Expert respondents will not always be able to provide judgments of "expert quality." Nonetheless, they may offer responses that are based on limited or flawed models of the subject of interest. We question whether having these flawed judgment data should be preferred to having no data at all if the flawed data lead to seriously misleading conclusions. There are few if any cleanly diagnostic methods for evaluating the quality of expert judgments. However, careful analysis may go a long way in screening out judgments to which little or reduced weight should be given.