THE ASSESSMENT OF WORTH: A SYSTEMATIC PROCEDURE AND ITS EXPERIMENTAL VALIDATION

BY

JAMES RUMRILL MILLER III

A. B. PRINCETON UNIVERSITY (1959)

M. B. A., HARVARD BUSINESS SCHOOL (1962)

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June, 1966

Signature of Author .....
Sloan School of Management, June 9, 1966

Certified by .....
Thesis Supervisor

Accepted by .....
Chairman, Department Committee on Graduate Students

June 9, 1966

Professor William C. Greene
Secretary of the Faculty
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Dear Professor Greene:

In accordance with the requirements for graduation, I herewith submit a thesis entitled "The Assessment of Worth: A Systematic Procedure and its Experimental Validation."

Sincerely yours,

James R. Miller

ABSTRACT

This thesis addresses itself to the problem of assessing worth. It is assumed that a decision context has been specified and that a fixed set of discrete alternatives has been produced. It then remains to assess the worth of each alternative, to estimate the resource drains required by each, and to combine these considerations, along with considerations of risk/uncertainty, so as to arrive at a final decision. The bulk of this thesis is directed primarily toward worth assessment.

To aid in the assessment process, a detailed procedure has been devised. The purpose of this procedure is set forth, and step-by-step instructions for its actual implementation are presented. A live instance of its complete application is also provided for illustrative purposes.

A dual-purpose experiment was performed. One purpose was to validate the procedure itself (i.e., to demonstrate that it could be carried out successfully by professional decision makers). The other purpose was to use various aspects of the procedure to study individual decision processes in a laboratory setting. Results drawn from the experiment are interpreted and conclusions are drawn from both points of view. Additional implications for decision making are traced out from the normative, from the descriptive, and from the practical points of view. The procedure and the experiment are reviewed critically, and suggestions are made for further research.

Major conclusions are that the procedure can be carried out successfully -- at least by professional decision makers in a laboratory setting -- and that all phases of it exert an important impact upon the decision making process. Critical to its overall impact, and particularly to its success, are its quantitative aspects. Requirements to quantify worth notions induce decision makers to formulate and validate their preferences. These two consequences are received very favorably by the decision makers themselves. An additional consequence is to provide an explicit and logically consistent assessment structure which, if considered acceptable, may be used to guide a final decision.

The procedure is both general and flexible with respect to type of decision, type of alternative, and type of worth criterion. However, its application may be restricted by type of decision making personnel and certain contextual factors.

0.1.0 HISTORY AND SPONSORSHIP OF THE THESIS

Some time ago the writer became disenchanted with profit maximization both as a normative and as a descriptive principle of decision behavior. Many important consequences of a decision either cannot or should not be assessed in dollar terms, and not all of those that can and should be are so assessed in actual practice. This suggested a need for a more general and flexible procedure both to reflect all of the real objectives of decision makers and to convert this information into a single unit of worth (other than dollars), which could then be applied to alternatives.

Additional impetus was given to the development of this idea by a growing enchantment with normative decision theory (e.g., statistical decision theory, operations research, and systems analysis). However, without an explicit and descriptively accurate statement of what is desired from alternatives, such normative techniques have no practical application. How many times has the eager and technically competent consultant been stymied in his efforts to improve practical decision making because the practical decision maker did not know (or at least could not communicate) what he really wanted? How many times has the consultant then made an assumption about what the decision maker does (or should) want, only to find, after performing extensive analyses, that the decision maker did not really want that after all? While working as a systems analyst, the writer ran squarely into both of these problems.

Development of the assessment procedure began about three years ago in an attempt to obviate some of these difficulties. Development occurred in a very practical context, wherein the purpose was to assist

The United States Air Force evaluate and select computers. This effort was supported by the MITRE Corporation under contract to the Air Force.[1]

After three years of collaborative effort involving both MITRE and Air Force personnel, an assessment procedure was produced. Only limited opportunities to validate the procedure were provided by the Air Force - MITRE environment both because of the highly focused nature of this effort (i.e., toward evaluation and selection of computers) and because of certain proprietary constraints.

Skillful guidance and generous support were provided by Massachusetts Institute of Techology to extend substantially the scope of the procedure. Under joint sponsorship of the Sloan School of Management, the National Aeronautics and Space Administration[2], and Project MAC[3], a dual-purpose experiment was designed.

One purpose of the experiment was to validate the procedure. The other purpose was to use various aspects of the procedure to study individual decision processes in a laboratory setting.

---

1. The Electronic Systems Division was the Air Force Sponsoring Agency under contract AF 19(628)-5165.

2. The research described in this thesis was supported (in part) by NsG-235 of the National Aeronautics and Space Administration for research studies in the organization and management of research and development. The findings and the views reported herein are those of the author and do not necessarily reflect those of the sponsoring agency.

3. Work reported herein was supported (in part) by Project MAC, an M.I.T. research program sponsored by the Advanced Research Projects Agency, Department of Defense, under Office of Naval Research Contract Number Nonr - 4102 (01). Reproduction in whole or in part is permitted for any purpose of the United State Government.

The experiment was carried out at and with the cooperation of the Defense Weapons System Management Center at Wright-Patterson Air Force Base, Dayton, Ohio.[4]

The procedure and the experiment constitute the subject matter of this thesis.

4. This management center is sponsored and supported by the Department of Defense. Teaching and research activities are carried out jointly by military personnel from all branches of the Armed Services and by Ohio State University.

0.2.0 PERSONAL ACKNOWLEDGEMENTS

It is obvious from a review of the history and sponsorship of this thesis that many debts are owed to many people. The single most important debt is owed to Professors Donald G. Marquis, H. Martin Weingartner, and Gordon M. Kaufman, who served as my thesis committee. These men both individually and collectively provided substantial assistance in designing and executing the research reported herein. Particular thanks are due to Donald Marquis. For several years he has provided encouragement and direction not only to my thesis but also to my professional development. In addition, he has always been a valued personal friend.

Early motivation for this thesis was provided in large part by my mentors at the Harvard Business School. Professors Howard Raiffa and Robert Schlaifer inspired an interest in normative decision theory. Professor Raymond A. Bauer inspired an equal interest in the descriptive aspects of decision making. Professor John V. Lintner also deserves thanks for nurturing these interests, particularly their integration.

The actual development of this thesis began while I was a systems analyst at the MITRE Corporation. Initial concept formation was aided considerably by frequent and lengthy conversations with Edward I. Friedland, Martin V. Jones, Lee R. Morris, Norman Waks, and Carter F. Wolfe. Concept revision and actual development of the procedure were guided by Jack D. Porter, Eugene D. Lundberg, Bernard H. Rudwick, David F. Votaw, and Robert H. Hamilton. Particular thanks are due to Jack Porter for continual constructive criticism and to Carter Wolfe for substantial collaborative effort in designing and working out detailed procedures. Thanks are also due to Sandi Moss for typing and re-typing numerous early drafts.

A debt of gratitude is owed to Colonel Edward McCloy and Messrs. Homenko, Jones, Lambert, and Joslin of the Electronic Systems Division, Unites States Air Force. These men all provided useful criticism during the early developmental phases -- particularly regarding issues of practical application.

In addition to my thesis committee, useful guidance in the design and execution of the experiment were provided by many Massachusetts Institute of Technology personnel. Early suggestions were offered by Professors William M. Evan, William H. Gruber, and Donald C. Carroll (who was originally a member of my thesis committee). Additional helpful suggestions, particularly regarding the descriptive implications of my work, were provided by Emanuel Kay, David Sirota, and Professor Peer O. Soelberg.

For their generous assistance in pre-testing various research instruments I wish to thank David Sirota, Thomas J. Allen, James R. Brown, Michael Gold, Irwin Rubin, and Lee L. Selwyn. I also wish to thank James H. Morris, David N. Ness and Thomas N. Van Vleck for their advice in programming the analysis.

Implementation of the experiment was aided considerably by the cooperation of the Defense Weapons System Management Center. Thanks are due to Colonel John F. Harris, Captain Kenneth W. Heising, and Lt. Colonel Howard R. Phillips in this regard. Special thanks are due to Colonel James H. Schofield and Bertrand L. Hansen for assuming personal responsibility for the administration of the experiment and for their generous assistance and encouragement. Many detailed tasks were accomplished with the voluntary assistance of Mrs. Jean Vogel.

Useful criticism of the methodology was provided by all members of my committee. Similar assistance was provided by Professors

Howard Raiffa, Robert Schlaifer, James L. McKenney, Richard S. Rosenbloom, and Andrew H. Kahr of the Harvard Business School. Particular thanks are due Professors Donald G. Marquis and Merrill M. Flood and to Alberto Leon for their guidance in the interpretation of results and conclusions.

Typing of the thesis and other essential services were provided by Virginia Stupak, Elizabeth Schneider, Jackie Robinson, Michele Parise, Katherine Blakeslee, Pamela Marsters, Betty Benedetto, Julia Wight, Mary Ellen Van Voast, Marsha Baker, Sandra Lawson, and my wife, Anni. Thanks are also due to John Gilbert and to Katharine DeP. Gilbert for proof-reading the final draft.

Needless to say, the thesis could not have been written without the direct contribution of the sixty officers and civilian personnel who served willingly as subjects of the experiment. Their extraordinary cooperation accounted in large part for its successful outcome. Similarly, I am indebted to my anonymous colleague whose decision among alternative job offers constitutes the subject matter of Chapter III.

I wish also to thank my long-time friend and business partner, John P. Veasy, for providing much useful criticism of the assessment procedure and for originally suggesting that the experiment be performed at the Defense Weapons System Management Center.

Finally, I wish to thank my wife, Anni, not only for her direct contribution in helping to type this thesis, but also for putting up with the great demands placed upon both of us over the last year.

James R. Miller III

TABLE OF CONTENTS

Table of Contents (Cont.)

Table of Contents (Cont.)

Table of Contents (Cont.)

CHAPTER V

AN EXPERIMENTAL TEST OF THE PROCEDURE AND RELATED FACTORS INFLUENCING THE DECISION MAKING PROCESS

Table of Contents (Cont.)

CHAPTER VI

CRITICISMS, OVERALL CONCLUSIONS, AND FINAL INTERPRETATIONS

Table of Contents (Cont.)

Table of Contents (Cont.)

LIST OF ILLUSTRATIONS

CHAPTER I

CONCEPTUAL ORIENTATION

1.1.0 INTRODUCTION

Assessing the worth of "complex" alternatives in an "important" decision situation is generally regarded as difficult. However, this task constitutes but one phase in the still more difficult process of producing such alternatives and making a final decision among them. Some of the factors which make an alternative "complex", which make a decision "important", and which thereby render the overall decision process difficult are stated and illustrated in the paragraphs that follow.

Consider the case of a large business organization which has decided to automate a substantial portion of its day-to-day activities by acquiring and operating an electronic computer. Producing, evaluating, and finally selecting an alternative to implement this decision would very likely be difficult for several reasons.

First, it must be determined which activity or activities are to be automated and to what extent. This requires the decision maker to describe in some detail the job to be performed by whichever alternative is finally selected. In the case of a computer, such a job description would ordinarily be both lengthy and complex.

Second, having formulated an adequate job description, the next task is to explicate the overall purpose in automating whatever activities are described therein. But computers are multi-purpose rather than single-purpose instruments. They satisfy many different objectives simultaneously. For this reason, their overall worth to an organization cannot be reckoned on the basis of a single criterion. This places the following responsibilities squarely upon the shoulders of those entrusted

with making a final decision:

1. The several objectives which are to be satisfied by acquiring and utilizing a computer must be listed, and the list must be fairly complete.
2. From each listed objective must be derived a set of specific worth criteria in terms of which the physical performance of a computer may be assessed.
3. Some means must then be found to organize and integrate these multiple criteria into a consistent and meaningful assessment structure.

Third, both the acquisition and the operation of computing equipment have many important ramifications for an organization. There is no simple or unique consequence on the basis of which an entire decision can be made. There are many performance consequences which must first be ascertained with reasonable accuracy and then assessed meaningfully before a final decision can be reached. This suggests the need for a clear, systematic, and replicable procedure to insure that no important performance consequences are overlooked.

Fourth, since both multiple criteria and multiple performance consequences are present, some means must be found to establish worth connections between the two. But this is not as easy as it may seem. A single performance consequence may be related simultaneously to several worth criteria (e.g., excess core storage capacity over and above minimum requirements might be considered relevant simultaneously to computing capability, to expandability potential, and to providing a hedge against the disastrous consequences of underestimating job requirements). Conversely, many performance consequences may be related

simultaneously to a single worth criterion (e.g., excess core storage capacity, estimated time required to complete a stated job, and various multi-processing features might all be considered relevant measures of total computing capability.)

Fifth, complex patterns of interaction may exist among various physical performance consequences due to the fact that computers are designed not as crude conglomerations of unrelated components, but rather as highly organized and integrated whole systems. This makes it difficult to understand and, therefore, to predict accurately an entire set of specific consequences.

Sixth, even if all performance consequences were known for certain, there may still exist complex patterns of interaction among the various worth criteria imposed by human beings upon this known performance. The structure of human worth notions is itself infested with intricate patterns of interdependence. Even worse, human beings often find it exceedingly difficult to distinguish in their own minds between interaction among performance consequences (a physical phenomenon characteristic of computers) and interdependence among imposed worth criteria (a psychological phenomenon characteristic of human beings). This renders still more difficult any attempt to understand, to assess, and to select computing equipment.

Finally, there is the question of resources expended to acquire and operate a computer. It is not always easy to predict with accuracy the amounts of manpower, materiel, and monetary resources which will of necessity be expended in order to implement any proposed automation alternative. Even if the resource implications of each proposed alternative were predictable, it must still be decided how much of each type

of resource should be expended on the particular job under consideration. In other words, the relative importance of the job must be ascertained in advance, and this must be translated into specific amounts of each type of resource which might appropriately be expended to perform that job.

To summarize, assessing the worth of complex alternatives in an important decision situation is neither simple nor easy. Making a final decision requires still more skill and effort. In the illustrative example of selecting computers, this was because:

1. Describing the activities to be automated entails detailed analysis.
2. Computers are multi-purpose rather than single-purpose, which requires formulating multiple worth criteria and integrating these criteria into a consistent and meaningful assessment structure.
3. Acquisition and operation of a computer have multiple performance consequences rather than a single consequence, which requires some systematic procedure to avoid overlooking important items.
4. Effective assessment requires establishing precise connections between worth criteria and performance consequences, which is difficult due to the many-to-one and one-to-many relationships between the two.
5. Complex patterns of interaction exist among various performance consequences, which makes their accurate prediction difficult.
6. Complex patterns of interdependence exist among various worth criteria, which makes the job of assessment

difficult even when performance consequences are predictable.

7. Expenditures in manpower, materiel, and monetary resources are not always easy to predict and even more difficult tc choose appropriate levels for.

Historically, decision makers have attempted to cope with the above kinds of problems largely on the basis of subjective judgment and intuition. Subjective estimates have been used quite frequently to predict probable resource and performance consequences. Personal judgments have also been used both to assess the worth of different amounts of predicted performance and to effect trade-offs among various worth criteria. The twin problems of physical interaction among performance consequences and conceptual interdependence among worth criteria have been handled similarly - that is, on an intuitive basis. Now if the decision problem under consideration were simple, widely understood, and relatively inconsequential, this might be the best way to proceed. The extra gains realizable from formalizing, systematizing, and making explicit such simple and inconsequential decisions would probably be insufficient to justify the required additional time, effort, cost, and just plain nuisance. However, when the problem becomes even moderately complex (and we have seen that selecting computing equipment is quite complex); or when the items being selected are unfamiliar and poorly understood (and computers are not yet well understood); or when the consequences of making a poor decision are significant (and computers are extremely expensive to acquire and operate as well as massive in their impact upon the daily work routine); then heavy reliance upon intuition and subjective judgment - unaided by systematic logic - becomes

a dangerous gamble indeed. It just does not seem reasonable to expect any human being to make accurate predictions of numerous and intimately interrelated performance and resource consequences, to attach meaningful measures of worth to each of these predictions, and then to effect meaningful trade-offs among worth criteria on the one hand and between total worth received and total resources expended on the other hand - all off the top of his head. Perhaps this explains the reluctance of many to undertake such a Herculean task. But even when an occasional brave soul volunteers for the job, it still seems unreasonable to permit him the privacy of intuitive choice undisciplined either by factual evidence, by logical procedure, or by consensual validation. The potential damage of a poor decision is just too great. For these reasons, therefore, an explicitly stated, logically consistent, and uniformly applicable procedure for assessing alternatives under such circumstances would seem essential.[1] The central purpose of this thesis is to develop just such a procedure and to report an experimental test of its validity.

1. It should be made clear that explicitness, logical consistency, and uniform applicability do <u>not</u> preclude the use of subjective judgment. Quite to the contrary, it is the writer's view that subjective judgment <u>must</u> be used both in assigning measures of worth to various performance consequences and in trading off worth among various criteria. Subsequent sections of this thesis will be specifically devoted to supporting this point of view. Rather, what is being stipulated here is that, when used, subjective judgment should be made explicit, should be thoroughly scrutinized for logical consistency, and should be elicited by a uniformly applicable and replicable procedure. The writer can think of no better way to insure that personal judgments will be free of false assumptions than by stating these assumptions explicitly. Nor can he think of a better way to insure valid reasoning from assumptions to conclusions than by exposing the reasoning process to critical scrutiny. Nor can he think of a better way to elicit a cross-section of opinion and to establish a consensus of preferences than by means of a uniformly applicable and replicable procedure. Most important of all, the writer can think of no better ways than these to obtain feedback on the assessment procedure and, therefore, to provide a constant impetus to its improvement.

1.2.0 STATEMENT OF THE PROBLEM

The overall problem to be treated in this thesis is seven-fold. Assuming that an important decision is to be made among complex alternatives, then the problem is:

1. to describe the job to be performed by whichever alternative is finally selected (i.e., to list the various activities which are to be carried out);

2. to formulate the overall purpose in making the decision (i.e., to abstract a specific set of job objectives from the job description);

3. to produce one or more feasible alternatives (i.e., to design and/or to solicit proposals for at least one alternative whose performance, if selected, would be viewed as satisfying minimum adequacy requirements);

4. to predict the worthwhile performance consequences associated with each alternative (i.e., to predict the types and amounts of worthwhile performance which would be realized from acquiring and utilizing each of the alternatives produced);

5. to assess the worth of these predicted performance consequences (i.e., to assess the extent to which the above-predicted performance would succeed in accomplishing stated job objectives);

6. to predict the resource consequences associated with each alternative (i.e., to predict the types and amounts of

limited resources which would necessarily be expended to acquire and utilize each of the alternatives produced);

7. to reach a final decision (i.e., to match worth of performance received against limited resources expended on each of the alternatives produced so as to determine whether any of them should be selected and, if so, which one).

Actually, this thesis will address itself almost exclusively to the problem of worth assessment (see 5. above). It will henceforward be assumed for discussion purposes that a job has been described, that an overall decision purpose has been formulated, that one or more feasible alternatives have been produced, and that the physical performance of each alternative has been adequately predicted. Nevertheless, despite these simplifications, the residue of the problem is still very difficult. To illustrate the remaining difficulties, let us return to the case of selecting a computer and consider a highly simplified hypothetical example.

Suppose that the job to be automated is a real-time application of inventory control. Suppose, further, that the only performance consequences considered important are the time required to respond to an inquiry and the maximum useable size of the file of items which can be maintained. Suppose, also, that the only resource considered important is the initial dollar investment required to procure hardware. Finally, suppose that three alternative systems have been proposed by competing computer manufacturers and that validated estimates of their performance and cost consequences are as shown in Table 1.

TABLE 1

	Alternative I	Alternative II	Alternative III
Response Time	10 Minutes	20 Minutes	38 Minutes
File Capacity	100,000 Items	175,000 Items	150,000 Items
Investment Cost	$110,000	$125,000	$90,000

The decision maker would now be faced with the task of making trade-offs between different levels of response time and file capacity to arrive at some notion of the overall worth of each alternative, and then he would have to match overall worth against cost on all three. Comparing alternatives I and II, he would have to decide whether the degradation in response time from 10 to 20 minutes were at least offset by the increase in maximum file capacity from 100,000 to 175,000 items. If no, then it certainly would not be worthwhile spending the extra $15,000 to purchase alternative II. If yes, if the increased file capacity more than compensated for the inferior response time, then he would have to decide whether the net gain in worth derivable from selecting alternative II over I warranted spending the additional $15,000. But what about alternative III? It is much cheaper than the other two, its file capacity falls between the other two, and its response time is substantially higher (and, therefore, less desirable) than both of its competitors'. Comparisons similar to those described above would have to be made first between alternatives I and III, and then between alternatives II and III.

However, the above types of comparisons, even if carried out successfully, would not be sufficient to dispose of the problem completely. There still remain the twin dangers of "under-kill" and

"over-kill". One or more of the three alternatives would provide "under-kill" if there existed either a maximum acceptable response time or a minimum required file capacity, and if estimated performance fell beyond either of these limits. This is an obvious kind of danger which can usually be detected with little difficulty - particularly if such mandatory performance requirements have been stipulated in advance on the basis of careful engineering and design considerations.

In contrast, the other kind of danger - the danger of "over-kill" - is far more subtle and much more difficult to detect. The reason is that "over-kill" is an economic rather than an engineering concept. Assessment of "over-kill" requires simultaneous consideration of both performance and resource consequences. The essence of "over-kill" lies <u>not</u> in the mere fact that more performance may be proposed than is necessary, but rather in the fact that whatever additional performance (over and above minimum requirements) is proposed may not warrant expending whatever additional resources are required to receive that additional performance. On economic grounds, it may be preferable to accept lesser performance - or even to accept zero performance (i.e., abandon the project) - and to expend the saved resources on some other project entirely. Returning to our example, it may be that alternative III, even with its relatively long (38 minute) response time, is more than adequate to meet the job requirements. Under such circumstances, it might be economically unwise to spend any more than $90,000 on this job. Alternatively, it may happen that even $90,000 is too much to spend. It may be that a far less efficient manual system costing only $25,000 would also be adequate, and that even the cheapest of the proposed automation alternatives (costing $90,000) would <u>not</u> justify spending the

extra $65,000. That same money might better be spent on some other automation project or, perhaps, on some other project completely unrelated to automation. Before a final decision can be made, all of these issues should be considered, and the decision maker should be prepared to reject any (or even all) of the proposed alternatives if either "under-kill" or "over-kill" become apparent.

The difficulty of making the above types of trade-off decisions - first between different kinds of performance to arrive at an assessment of overall worth, and then between overall worths and their associated costs - is probably quite evident to the reader. And this was a highly simplified example. As the number of worth criteria and related performance consequences increases, the problem quickly reaches unmanageable proportions. If multiple resources are also considered (e.g., manpower and materiel as well as monetary resources), and if complex patterns of both physical interaction and worth interdependence emerge, then effective solution of the problem by sheer intuition becomes just about impossible. This suggests the need for a more formal approach.

The remainder of this thesis will be oriented specifically toward the development of a more formal approach. Chapters I through IV will set forth a systematic procedure to aid in the assessment of worth. Chapters V and VI will report the results of an experiment designed, in part, to validate the assessment procedure and will integrate these results with the preceding discussion.

1.3.0 THE CONCEPT OF WORTH

For purposes of this thesis, the worth of any object, activity, or situation is, roughly speaking, the extent to which such is perceived by a decision maker or group of decision makers as satisfying clearly articulated objectives. Thus, the worth of an alternative in a well-specified choice situation would be defined as the extent to which that alternative satisfies whatever objectives have been articulated regarding the job to be accomplished.

The above notion of worth is intentionally stated in a very general manner. A detailed definition will be presented later. Specifically, step-by-step procedures for assessing worth will be developed in Chapter II, and one purpose in setting forth these procedures is to provide an operational definition of the concept itself. For now, however, it will be useful to outline the intended meaning and scope of the worth concept both to orient future discussion and to preclude the imputation of unintended meanings to the subject matter of this thesis.

1.3.1 The Intended Meaning Of Worth

From the above discussion it is apparent that worth notions constitute an internal property of human decision makers -- not an external property of the physical objects, activities, and situations whose worth is being assessed. Worth is here conceived as inherent within the perceptual apparatus of the decision maker himself. This represents an important departure from traditional approaches to the same problem which focus primarily upon the physical entities being assessed. The detailed procedures developed in Chapter II will clarify this distinction operationally.

Since worth is defined with respect to clearly articulated

objectives, it is necessary that such objectives exist. Operationally, this usually means that a deliberate effort must be made to formulate and articulate clear objectives before worth may be assessed. It also means that worth notions will be multidimensional whenever multiple objectives and/or multiple performance measures are considered relevant (e.g., in the case of "complex" alternatives).

Finally, worth refers to the extent or degree to which some object, activity, or situation satisfies stated objectives. This suggests the need for establishing a definite scale in terms of which various degrees of goal satisfaction (and, therefore, imputed worth) may be expressed. Section 1.4 will address itself to this task.

1.3.2 Implications For The Task Of Assessing Worth

Having discussed briefly the meaning of worth as used in this thesis, we shall now investigate some of the problems involved in performing a practical assessment of worth.

First, the act of formulating and articulating a clear set of objectives in terms of which worth may be assessed is not always easy to accomplish. Decision makers may be either unable or unwilling to formulate and display a complete list of objectives because of:

1. incomplete awareness of the problem at hand;
2. incomplete knowledge of the intricacies of the problem;
3. inability (due to time, money, and/or manpower constraints) to devote sufficient "thinking" effort to formulating a complete and explicit list of objectives.

Alternatively, they may be unwilling to formulate and particularly to display a complete list of objectives because of:

1. fear that some of the "real" objectives will be disapproved if layed bare to public scrutiny;

2. fear that some of the "real" objectives, even if tacitly approved, may not be easily defended in the political arena;

3. realization that some objectives, even if approved and defensible, may not receive complete consensual validation from all interested parties - particularly those who would suffer adverse consequences should the "real" objectives be satisfied.

These latter sources of unwillingness may attain particular motivational importance if decision makers are themselves imbedded in an organizational environment rife with threat, conflict, or a strong tradition of defensive conservatism.

Second, there is the issue of confirming worth judgments. Unlike allegations of fact or scientific predictions, worth judgments cannot be confirmed by empirical test. They are <u>in principle</u> untestable by ordinary scientific means. This is because worth judgments are stated in such a way as to be neither factually true nor factually false. They merely exist in the minds of human beings to be accepted or rejected either in whole or in part by other human beings (or, perhaps, by the same human being at a different point in time). In short, the acceptability of worth judgments is here conceived as a matter of personal taste and based on an act of faith.

A third implication follows from the second. This involves the

identity of decision makers. Different decision makers may very well have different objectives regarding the same situation, which renders the outcome of an assessment highly dependent upon who undertakes to perform the assessment. Stated a bit more simply, the outcome of an assessment depends critically upon whose values are adopted. One way out of this situation is to strive for consensus among potential decision makers, but this is not always possible (and perhaps undesirable) to achieve. In any case, the worth concept is not here defined as requiring consensus.

A fourth implication involves the stability of worth judgments over time. Not only may there exist lack of consensus among separate decision makers at a given point in time, but there may also exist lack of agreement among separate worth judgments made by the same decision maker at different points in time. As additional experience is gained, one would expect (or at least hope) that a given decision maker would alter his worth judgments to account for whatever new insights this additional experience has brought about. Temporal instability is thereby created, but, possibly, in an entirely appropriate manner. In any case, the worth concept is not here defined as requiring temporal stability either.

1.4.0 CONSTRUCTING A MEASURE OF WORTH

Having committed ourselves to creating a formal assessment scheme, we must now tackle the problem of defining a uniform and convenient measure of worth. This suggests (although it does not require) reducing the problem to numbers. Why? Because numbers are familiar, widely used as tools of measurement, and easy to manipulate. However, lest there be any confusion on this score, let it be understood at the outset that the measure of worth to be created, and particularly its numerical scale characteristics, constitute an ad hoc invention specifically designed for the assessment procedure to be developed in Chapter II. No claim is being made that this worth measure or its scale characteristics derive deductively from any set of (normative) axioms. With this proviso firmly in mind, let us endeavor to construct a numerical measure of worth.

1.4.1 The Basic Purpose In Constructing A Measure Of Worth

Perhaps the best way to initiate detailed discussion is with a formal statement of purpose. When a numerical measure of worth is used as a vehicle of assessment, the underlying rationale is that such numbers will be assigned to various objects and activities in such a manner as will reflect their perceived worth. That is, worth numbers will be assigned such that numerical relationships between assigned worth numbers will faithfully reflect perceived worth relationships between the objects and activities to which these numbers have been assigned.

In order to implement the above purpose, it is first necessary to specify the kinds of worth relationships which are to be reflected by means of numerical symbols. It is also necessary to specify the numerical conventions which establish a correspondence between numerical

relationships and the perceived worth relationships which are being depicted thereby. The first of these tasks will be undertaken in Section 1.4.2. The second will be undertaken in Section 1.4.3.

1.4.2 Worth Characteristics To Be Reflected

The most fundamental characteristics of the worth concept to be reflected in our choice of a numerical measure are the three psychological states of preference, aversion, and genuine indifference. A decision maker is said to possess a positive preference for some object or activity if and only if that object or activity elicits a positive affective response from him (e.g., joy, pleasure, interest, excitement, gratification, etc.). Thus, most people possess a positive preference for automobiles because they elicit all of the above positive responses. In addition, most people are willing to part with money (a scarce resource) in order to receive these benefits from an automobile.

A decision maker is said to possess a distinct aversion (negative preference) for some object or activity if and only if that object or activity elicits a negative affective response from him (e.g., distress, anxiety, shame, guilt, disgust, etc.). Most people possess a distinct aversion to death. The very thought of it arouses a great deal of distress and anxiety, and many people are willing to part with substantial amounts of money (i.e., purchase life insurance) in order to ameliorate its unwanted consequences.

A decision maker is said to feel genuinely indifferent toward some object or activity if and only if he possesses neither a preference for nor an aversion toward that object or activity.

Returning to the concept of worth, this is usually thought of as related to positive preferences only. That is, when an object or activity

is said to possess some worth, this usually means that somebody possesses a positive preference for it and/or its consequences. The concept of "negative" worth (referring to objects or activities toward which people feel aversive) is less well defined.

In light of these observations, certain numbers on the worth scale (to be created in Section 1.4.3) will be reserved to indicate positive preferences, and a single number not included in the above range will be reserved to indicate a state of genuine indifference. Negative preferences or aversions will be represented by negative analogs of the positive worth numbers.[2]

Another aspect of the worth concept to be reflected in our choice of a numerical measure involves its boundedness. Is it possible for something to be completely worthwhile? Can a decision maker be completely satisfied (or dissatisfied) wich some object or activity? Although seemingly simple at first glance, this is a very subtle and important question. Let us investigate the issue more closely.

If asked to assess the worth of some object without specifying how or for what purpose that object is to be used, it seems difficult to conceive of any natural, logical outer bounds to the answer. Like the brightness of a color or the loudness of a sound, there exist no apparent natural limits. On the other hand, once a definite job has been specified and once a definite set of objectives has been defined, then the question appears in a somewhat different light. When asked to assess the worth of some object for performing some stated job in accordance with well defined objectives, it seems reasonable to talk in

---

2. More will be said in subsequent sections about negative preferences or aversions and their representation on the worth scale.

terms of the extent to which that object satisfies the stated objectives. Furthermore, since definite objectives have been stated, it seems reasonable to talk about the possibility, at least, of having those objectives completely satisfied. Thus, under these revised circumstances, there appears to be a natural outer bound to the assessment of worth. This will be reflected by placing numerical bounds on the worth scale to be constructed shortly.

Still another aspect of the worth concept to be represented numerically is its continuity or divisibility. It would seem desirable to permit the expression of preferences and preference distinctions to range from infinitesimal magnitudes to large magnitudes. Although decision makers may not always wish to avail themselves of this flexibility, a continuous or everywhere-dense worth scale will be defined to accomodate such notions whenever they are felt.

Finally, there is the question of preference relationships between different objects or activities whose worth is being assessed. There are three kinds of basic relationships which will receive numerical representation by establishing appropriate scale conventions. These are:

1. same-difference relationship (i.e., whether two objects or activities are assessed as possessing the same or different worths);

2. greater than and less than relationships (i.e., whether one object or activity is assessed as possessing more or less worth than another);

3. comparative magnitude relationship (i.e., how many times as much worth one object or activity is assessed as possessing compared to another).

Let us now proceed to construct a numerical worth scale which will reflect all of the above characteristics.

1.4.3 Corresponding Scale Characteristics

A numerical worth scale will be established in accordance with the following ten scaling conventions.

1. Positive numbers will be assigned uniformly to situations assessed as possessing positive worth (i.e., toward which a positive preference is felt).

2. Negative numbers will be assigned uniformly to situations assessed as possessing "negative" worth (i.e., toward which a distinct aversion is felt).

3. The worth scale will be bounded from above by plus one and from below by minus one.

4. Plus one will be assigned only to those situations deemed completely successful in terms of accomplishing positive job objectives. Analogously, minus one will be assigned only to those situations deemed completely "successful" in accomplishing "negative" job objectives (i.e., to situations than which nothing worse is conceivable in the context of the stated job).

5. The number zero will be assigned uniformly to situations assessed as completely worthless (i.e., completely unsatisfactory - but not dissatisfactory; toward which genuine indifference - but not aversion - is felt).

6. All real numbers between plus one and minus one (inclusive) are permissible measures of worth.

7. Two situations will be assigned equal worth numbers if and only if they are assessed as possessing identical worth (i.e., a decision maker feels genuine indifference in choosing between them).

8. One situation will be assigned a higher worth number than another if and only if it is assessed as possessing more worth - that is, if and only if a decision maker prefers the first situation to the second.

9. Numbers between zero and plus one (exclusive) will be assigned to all situations assessed as partially successful in terms of accomplishing positive objectives. Worth numbers will be assigned to such situations according to their proportional or percentage accomplishment of the stated objectives. This defines magnitude comparisons in terms of their ratios.

10. Numbers between zero and minus one (exclusive) will be assigned to all situations assessed as partially "successful" in terms of accomplishing "negative" objectives (i.e., stated avoidance desires). Negative worth numbers will be assigned to such situations according to their proportional

or percentage "accomplishment" of stated "negative" objectives.

1.4.4 Choice Of A Unit Of Worth: Arbitrary Points

The choice of a unit of worth has already been made implicitly by two previous decisions. First, it was decided to bound the worth scale from above and below (i.e., to restrict worth numbers to fall between plus and minus one). This precludes the use of dollars or any other unit whose range lacks intrinsic logical outer bounds. Second, it was decided to assign a worth number to a situation according to the proportional accomplishment of stated objectives achieved by that situation. This means that worth numbers may be viewed as ratios between actually achieved satisfaction and maximum possible satisfaction of stated objectives. As such, no matter what units raw satisfaction might possess, any ratio formed would be dimensionless. That is, such raw units would cancel each other out in forming a ratio, and the result would not possess any explicit units at all. Worth numbers defined in this manner are like index numbers used by various rating schemes (e.g., The Consumer Price Index).

Despite the above considerations, there is an advantage to giving index numbers a definite label so that they may be easily remembered and conveniently discussed. Because such numbers have no physical dimensions and, therefore, no natural unit, any arbitrary label is permissible - so long as it does not suggest anything which possesses either dimensions or a natural unit. Henceforward, our worth numbers will be referred to as "points" or "worth points". This label is being adopted strictly for convenience, and it should always be remembered that worth points possess neither physical dimensions nor natural units. Worth points are completely arbitrary - by definition. Their signifi-

cance is encapsulated within the ten scaling conventions outlined in Section 1.4.3. However, their entire significance may not be immediately obvious upon a single reading of these ten conventions. Consequently, the next section will be devoted to tracing out several important implications which may have escaped a cursory first reading.

1.4.5 The Significance Of Worth Points

We are now in a position to answer several questions concerning the significance of worth point assignments. In particular, we may resolve such practical issues as:

1. the legitimacy of performing basic arithmetic operations on worth points;

2. the legitimacy of assigning worth points to situations toward which no positive preference is felt;

3. the legitimacy of assigning worth points to situations on any basis other than the extent to which such situations satisfactorily accomplish independently pre-established objectives.

First, it would be useful to know what kinds of arithmetic operations may legitimately be performed on assigned worth points. To answer this question, we must investigate the process by which worth points are assigned. Can decision makers give meaningful answers to questions of the following general form?

1. Given a stated objective;

2. Given a well described situation which purports to accomplish the stated objective, at least partially;

3. To what degree does the situation described succeed in accomplishing the stated objective, where degree is assessed in terms of a percentage-like number between zero and one?

If decision makers cannot answer the above type of question at all, then the issue of legitimate arithmetic operations becomes vacuous. There would be no worth points upon which to operate - either legitimately or illegitimately. (Note: The possibility of being unable to answer such questions is considered quite reasonable, and remedial measures will be discussed later in this thesis.)

On the other hand, even if such a question can be answered, there is still the problem of meaningfulness. How meaningful are numerical point scores, when assigned? That is, how confident are decision makers in the interpretability of the numbers they give? If decision makers can muster a reasonable degree of confidence in their own ability to answer the above type of question (and only the decision makers themselves can make such a determination), then all four basic arithmetic operations may be performed on assigned worth points, except for a sign restriction. Worth points may be added, subtracted, and multiplied by a non-negative constant with complete freedom. However, attention must be paid to their sign when multiplying or dividing by one another. Only worth numbers of like sign may undergo these operations, and then only their absolute magnitude is relevant. In the language of scaling theory, each half of the worth scale constitutes a full-fledged ratio scale, with the negative half being treated as a "mirror-image" of the positive half.

The second issue concerns the legitimacy of assigning worth points to situations toward which no positive preference is felt.

Ignoring the case of indifference, which receives a point score of zero, this includes both situations toward which a distinct aversion is felt and situations toward which neither a positive preference nor a distinct aversion are felt directly, but whose indirect consequences are such as to arouse a reduction in positive feelings.

An example of a situation generating direct aversive feelings would be the development of a new drug which proved highly efficacious in one way, but which also produced noxious side effects. Otherwise positive attitudes toward an effective contraceptive device would be substantially mitigated, if not completely overruled by the discovery that it induced permanent sterility. Sterility, here, would constitute a "negative" objective which most people would definitely prefer to avoid and toward which they would feel directly aversive. Negative worth points would be assigned to this situation depending upon the extent to which permanent sterility were induced.

However, there is another type of situation toward which decision makers may feel neither a direct preference nor a direct aversion. This is the situation wherein limited resources must be expended to complete a job. Unless a decision maker possesses miserly feelings, he has no direct aversion to spending money or committing workers or using up capital equipment <u>per se.</u> If the supply of such resources were truly unlimited relative to their demand, the resources themselves would have no worth at all -- either positive or negative. Consequently, spending such unlimited resources could only be regarded with indifference. There would always be enough to go around -- if the supply were truly unlimited. Resources only become valuable when their available supply falls below the total demand for their effective utilization. But even then, their worth is <u>not</u> intrinsic to the resources themselves,

Rather, their worth derives from the fact that they may be diverted to some alternative application which, if carried out, would generate consequences perceived as worthwhile in their own right. Expending limited resources to complete one job precludes using the same resources to complete some alternative jobs, and the worth of completing the alternative job must, therefore, be foregone.

In view of these observations, we may now ask whether it is legitimate to assign worth points directly to the expenditure of resources. The answer is no. Worth points, as defined in this thesis, can only be assigned to situations perceived as worthwhile in their own right because they succeed in accomplishing stated objectives. Although it would be possible to define "conserving resources" as a specific objective, it would be difficult to judge the worth of any given amount of conserved resources unless or until the alternative applications of the same resources had first been ascertained and assessed. Until this is accomplished, no meaningful point scores may be assigned to resource expenditures.

The above conclusions have two important procedural implications. Since worth points are generally not assigned to resource expenditures incurred in acquiring and utilizing a produced alternative, while worth points are assigned to other kinds of consequences related directly to stated job objectives, it is important to define at the outset just which consequences are to be regarded as resource-oriented and to distinguish these clearly from objective-oriented consequences. In addition, some alternative means must be found to reflect the worth implications of expending resources and to incorporate these explicitly into an overall decision making methodology. Section 1.5.2 will discuss briefly various ways of incorporating resource considerations into a final selection methodology.

The third issue concerns the legitimacy of assigning worth points to situations on any basis other than the extent to which such situations satisfactorily accomplish independently pre-established objectives. We have defined the worth of a situation (i.e., an object, activity, or degree of physical performance) as the extent to which that situation satisfies a stated objective in a specified context. We have also defined a worth point assignment as an assessment of the extent to which the stated objective has in fact been satisfied. Therefore, any procedure for assigning worth points - to qualify under these definitions - must measure the worth of a situation against the stated objective (e.g., against user needs and/or requirements). This sounds simple enough - perhaps even undeniable. But let us contrast this basis for making point assignments with an alternative basis whose interpretation and consequences are quite different.

Suppose that a computational job has been specified for which it is required to complete the daily workload within a single 8-hour shift. Suppose, also, that two computer manufacturers propose alternative systems, one of which completes the job in six hours, while the other completes the job in seven hours. Finally, suppose that one of the stated objectives in acquiring and operating the computing system is to receive a capability for expansion over and above the currently stated job and that it has been decided to measure such expansion capability by means of slack time provided between actual job time and the maximum permissible job time of eight hours. Then, the first proposal would provide 8-6 = 2 hours of slack time, while the second proposal would provide 8-7 = 1 hour.

One way of assigning point scores to these two competing proposals would be as follows:

1. Determine before inspecting the actual proposal data what range of possible slack times could conceivably be proposed. In this instance, it would be possible to receive proposals with slack times ranging anywhere from 0 hours (indicating that the entire 8-hour shift is required to complete the current job) up to something near 8 hours (indicating that the current job requires almost no time to complete).

2. Determine a standard or set of standards for judging the worth of every possible proposed slack time before inspecting what is actually proposed and independently of what is actually proposed.

3. Then, upon receipt of actual proposal data, compare the 2-hour and 1-hour slack times against these independently pre-established standards to arrive at an assessment of worth. If one hour of slack time were deemed almost completely satisfactory, then worth point scores of .98 and .95 might be assigned to the two proposals. If, on the other hand, slack time could not be utilized effectively unless at least 1.5 hours were provided, then worth point scores of .50 and .10 might appropriately be assigned. Finally, if at least three hours of slack time were required for effective utilization, then worth point scores of .05 and .01 might be appropriate.

The above procedure is entirely consistent with the definition of worth and the significance of worth points developed in this thesis. Point scores are assigned on the basis of independently pre-established

standards, and they indicate the extent to which proposed slack times accomplish the expandability objective.

Another way of assigning point scores would be as follows:

1. Make no effort to determine either the range of possible slack times which might be proposed or standards for evaluating proposal data before such data are actually received.

2. Upon receipt of proposal data, select some aspect of the data themselves (e.g., maximum proposed slack time, mean proposed slack time, median proposed slack time, etc.), and compare proposals against each other on this basis to arrive at an assessment of worth. Thus, the maximum proposed slack time of two hours might be selected as a basis of comparison, and 1.0 points might be assigned thereto. All lesser proposals might then receive point scores in direct proportion to the maximum proposal such that a 1-hour slack time would receive a point score of $\frac{1}{2} = .50$.

This procedure is not at all consistent with either the definition of worth or the significance of worth points developed in this thesis. Point scores are assigned without any regard for what is needed or desired in the way of slack time. Point scores are assigned solely on the basis of what is proposed. The writer seriously questions the sensibility of any assessment scheme which bases its judgments on what is actually proposed and ignores completely what is desired. Adoption of such a scheme in practical decision situations can easily lead either

to selecting the best alternative proposed - even though it is completely inadequate - or to selecting the least costly alternative proposed - even though it provides an enormous amount of "over-kill". In both of the above instances, it would seem preferable to reject all proposed alternatives and to redirect effort toward producing altogether new alternatives.[3]

1.4.6 Summary

To summarize the discussion in Sections 1.4.1 through 1.4.5, the following things have been concluded.

1. The basic purpose in assigning worth numbers to external situations (i.e., to physical objects, activities, and degrees of performance being assessed) is to reflect explicitly whatever worth characteristics and relationships are perceived by decision makers and imputed to such situations.

2. Numbers are used because they are familiar, well understood, and easy to manipulate.

3. Worth numbers should be assigned in such a manner that numerical relationships existing among assigned numbers will correspond with and, therefore, serve to identify worth characteristics and relationships imputed to the external situations to which worth numbers are assigned.

---

3. The "wrong" procedure illustrated above is no mere hypothetical example. It is an integral part of the assessment scheme used by the United States Air Force in selecting its business-oriented computers. A description of the entire scheme can be found in Rosenthal, S., "Analytical Technique for Automatic Data Processing Equipment Acquisition", 1964 Spring Joint Computer Conference, April, 1964.

4. The most fundamental worth characteristics to be reflected by assigned worth numbers are the psychological states of preference, aversion, and indifference.

5. Although the concept of worth is generally conceived of as referring to situations for which a positive preference is felt, "negative" worth has been defined in terms of "negative" objectives to take care of situations toward which a distinct aversion is felt. The concepts of "negative" objectives, "negative" worth, and their associated negative point scores are defined by "mirror-image" analogy to the corresponding positive concepts and associated point scores.

6. The assessment scheme to be developed later will apply only to situations for which definite objectives can be stated. Consequently, it makes sense to talk of the extent to which a given situation succeeds in accomplishing such objectives. This, in turn, suggests logical outer bounds to the scale of worth numbers corresponding to complete satisfaction of "positive" and "negative" objectives, respectively.

7. The scale of worth numbers is defined as continuous or everywhere-dense within its entire logical range.

8. Two situations are assigned identical worth numbers

4. The most fundamental worth characteristics to be reflected by assigned worth numbers are the psychological states of preference, aversion, and indifference.

5. Although the concept of worth is generally conceived of as referring to situations for which a positive preference is felt, "negative" worth has been defined in terms of "negative" objectives to take care of situations toward which a distinct aversion is felt. The concepts of "negative" objectives, "negative" worth, and their associated negative point scores are defined by "mirror-image" analogy to the corresponding positive concepts and associated point scores.

6. The assessment scheme to be developed later will apply only to situations for which definite objectives can be stated. Consequently, it makes sense to talk of the extent to which a given situation succeeds in accomplishing such objectives. This, in turn, suggests logical outer bounds to the scale of worth numbers corresponding to complete satisfaction of "positive" and "negative" objectives, respectively.

7. The scale of worth numbers is defined as continuous or everywhere-dense within its entire logical range.

8. Two situations are assigned identical worth numbers

if and only if they are perceived as equally worthwhile (i.e., if and only if a decision maker feels genuine indifference in choosing between them).

9. One situation is assigned a higher worth number than another if and only if it is perceived as more worthwhile - that is, if and only if a decision maker prefers the first situation to the second. A reverse statement applies to assigning lower worth numbers.

10. The scale of worth numbers is restricted to contain all real numbers between plus one and minus one (inclusive). The number zero is assigned to any situation which is perceived as worthless (i.e., completely unsuccessful in accomplishing stated objectives) and, therefore, to which genuine indifference - but not aversion - is felt. Plus one is assigned to any situation which is perceived as completely successful in accomplishing positive objectives. Intermediate numbers are assigned to situations perceived as partially successful in accomplishing stated objectives according to the extent or degree (reflected as a proportion or percentage) to which they are successful. Negative worth numbers, when appropriate, are assigned analogously, with minus one being reserved for complete satisfaction of "negative" objectives.

11. Worth numbers are given the label "points". Worth points are arbitrary. They possess no physical

dimension or physical unit. Their meaning and proper interpretation are completely specified by the scaling conventions outlined previously.

12. Addition, subtraction, and multiplication by a non-negative constant may be performed on worth points - provided the process by which they are originally assigned is such as to give them the full meaning specified by the ten scaling conventions. Whether or not they possess this full meaning must be decided by the decision makers themselves. In addition, multiplication and division of the absolute value of worth points with the same sign is also permitted (to compute a geometric mean or a ratio, for example).

13. Worth points are not ordinarily assigned to situations involving the expenditure of resources, unless specific resource-conserving objectives have been stated for the job.

14. Worth points will be assigned to situations strictly on the basis of objectives and standards of worth established prior to and independently of inspecting the performance consequences of produced alternatives. Point scores will <u>never</u> be assigned on any basis of comparison which matches the performance of alternatives against one another. Such bases of comparison are completly dependent upon performance produced and ignore what is needed or desired.

1.5.0 RELATED CONCEPTS

Before moving to the development of a formal assessment procedure, the relationship of worth to the classical concept of utility deserves some brief attention. In addition, the roles of worth, utility, and resources in the overall decision making process deserve a few brief comments. Although neither of these topics will be treated extensively within this thesis, a brief discussion of each will add perspective to our future discussions of worth assessment.

1.5.1 Risk, Uncertainty, And The Classical Concept Of Utility

The worth concept is completely devoid of any risk and/or uncertainty considerations. In assessing the worth of a situation, activity, or performance consequence, it is assumed that such an outcome will occur for certain. Consequently, assigned worth numbers will not reflect the aversion which a decision maker may feel toward either risk or uncertainty regarding the actual occurrence of that outcome. Furthermore, the process of assigning worth numbers provides no mechanism for reflecting perceived trade-offs between the worth of an outcome, conditional upon its actual occurrence, and the variable risk or uncertainty surrounding its occurrence. The worth concept and the related worth measuring and worth assessing procedures are, therefore, incomplete in this sense.

In contrast, the classical concept of utility, as articulated by Von Neumann and Morgenstern (14) and used by statistical decision theorists, does provide an explicit mechanism for reflecting perceived trade-offs between conditional worth, on the one hand, and risk or uncertainty on the other hand. However, the concept of utility ignores the problem of formulating and articulating a measure of conditional

worth. It assumes that the decision maker has already formulated a worth measure and proceeds from there.

A complete assessment procedure should take account of both conditional worth and risk/uncertainty considerations. That is, it should provide a mechanism for assessing worth, conditional upon certainty, and then it should provide an additional mechanism to account for the decision maker's attitudes toward risk and uncertainty. This thesis will address itself exclusively toward the former task (i.e., generating a conditional measure of worth). However, the numerical output of the worth assessment procedure to be developed herein is a perfectly legitimate input to the Von Neumann-Morgenstern utility assessment procedure. Worth point scores can be used as the numeraire to which utility numbers are assigned. More will be said later about this symbiotic bond between the worth and utility concepts.[4]

1.5.2 The Concept Of A Decision Rule

A decision rule might be defined broadly as any uniformly applicable directive which indicates a clear choice among properly specified alternatives in a given selection situation. It is through the mechanism of a decision rule that decision makers specify the trade-offs they are willing to make among worth, risk/uncertainty, and resource considerations.[5] Examples of decision rules which are frequently used in selecting among alternatives include:

1. The Economy Rule, directing decision makers to

---

4. The writer is grateful to Howard Raiffa for pointing out this explicit connection between the worth and utility concepts and for suggesting that worth point scores be used as the numeraire in a utility assignment procedure.

5. The trade-off between conditional worth and risk/uncertainty is assumed throughout this discussion to be encapsulated in a utility index of the variety discussed in Section 1.5.1.

select the least expensive feasible alternative (i.e., the least resource-consuming alternative which satisfies all stipulated mandatory performance requirements and, possibly, physical resource limitations);

2. The Ratio Optimizing Rule, directing decision makers to select whichever feasible alternative maximizes a utility-to-cost (or, equivalently, minimizes a cost-to-utility) ratio;[6]

3. The Weighted Average Utility/Cost Rule, directing decision makers first to assign numerical weights to receiving valuable performance versus expending limited resources, then to assign explicit measures of utility both to received performance and to expended resources, and finally to select whichever feasible alternative maximizes the weighted sum of these separate utility indices.

Obviously, the above list does not exhaust all decision rules that have been or could be used to select alternatives, but it does provide a reasonable basis for discussion. In particular, it provides a reasonable basis for illustrating the primary role of a decision rule in integrating worth, risk/uncertainty, and resource considerations.

In choosing a decision rule, the decision maker must ask himself what he is really trying to accomplish when he finally selects an alternative. He may raise such questions as the following:

---

6. The word "cost" is used throughout this section to indicate the physical process of expending resources including, but not restricted to, monetary resources.

1. Assuming that at least one of the produced alternatives is feasible, must one of them always win the selection; or is it possible to reject all of the alternatives on the grounds that they all provide "over-kill" and that the same resources might better be expended on some other project altogether?

2. Should each successive selection in which the decision maker is required to make a choice be considered separately, without regard to the consequences of that choice on subsequent selection decisions; or should the decision maker assume a broader viewpoint which embraces the whole sequence of decisions he must make?

3. In what sense should valuable performance received be compared with resources expended? Is it worthwhile to expend additional resources in order to receive additional valuable performance over and above minimum requirements? If so, how much more and until what has been achieved?

Answers to these questions should help the decision maker choose a decision rule, or at least narrow substantially the field of candidates. To illustrate why this is so (i.e., how these questions are related to various decision rules), let us consider the implicit answers given to each by the economy rule, the ratio optimizing rule, and the weighted average utility/cost rule, respectively.

First, the economy rule requires that, if at least one feasible

alternative is produced, then one of them must win. It is impossible, under the economy rule, to reject all feasible alternatives - even if the least costly alternative requires a staggering expenditure of resources. No protection against "over-kill" is provided.

Similarly, the ratio optimizing rule (in either of its two equivalent forms) provides no protection against "over-kill". It is quite possible to encounter a set of alternatives - all of which promise performance greatly in excess of what is required (or even desired) and which involve commensurately excessive resource expenditures. Nevertheless, that alternative with optimum ratio would still be defined and, unless the rule were enhanced with a budget constraint or some other protective device, "over-kill" would thereby be suffered.

The weighted average utility/cost rule also fails to provide any protection against "over-kill". As in the case of the ratio optimizing rule, all produced alternatives may promise excessive performance and require excessive resource expenditures. Without an explicit budget constraint or some other protective device, this rule is rendered equally helpless.

Regarding the second question, both the economy rule and the ratio optimizing rule focus attention exclusively on each successive selection considered by itself. No explicit consideration is given to the consequences of one selection decision on other such decisions. In particular, no recognition is given to the fact that, when the total supply of resources is limited (as it almost always is in real-world situations), what must be expended to choose an alternative in one selection cannot be expended on another selection. No limits or any other direct controls are placed on resource expenditures.

This is not the case, however, with the weighted average, utility/cost rule. In principle, a degree of control may be exercised over

the amount of resources expended either by selectively altering the weight on resources or by choosing appropriate utility functions for resource expenditures. Although it may not be clear how to exercise these controls so as to achieve an appropriate allocation of resources over an entire sequence of selection decisions, at least a potential control mechanism exists.

Regarding the third question, the three decision rules give quite different answers. The economy rule rejects completely the notion that additional performance over and above minimum requirements might be worth spending additional resources to obtain. It chooses the cheapest alternative that does the job, even if performance is just barely satisfactory.

In contrast, both the ratio optimizing rule and the weighted average utility/cost rule recognize the potential worth of additional performance over and above minimum requirements, and both rules permit spending additional resources to obtain it. However, the extent to which each rule will spend additional resources and the apparent reasons for spending it are different. Under the ratio optimizing rule, the goal is to get the "best buy" (i.e., the most for whatever resources are expended as evidenced by either a maximum utility-to-cost or a minimum cost-to-utility ratio). Under the weighted average utility/cost rule, additional resources may be spent to obtain additional valuable performance so long as the extra utility received from additional performance equals or exceeds the extra penalty suffered by spending more resources. The balancing of utilities is carried out by means of the utility functions on various performance measures and resources in conjunction with the weights placed on each.

The preceding discussion was intended to indicate that, even

after a satisfactory measure of worth has been defined, and even after risk/uncertainty has been taken into account by means of a satisfactory utility index, there still remains the problem of integrating these considerations with a careful consideration of resource expenditures before a complete decision methodology can be achieved. Choice of a satisfactory decision rule is the means of achieving integration, and, as the preceding discussion illustrated, this is not a simple task.

1.6.0 SOME CONCLUDING REMARKS

In this chapter, we have discussed the overall problem of designing and implementing a rational decision making methodology. An important part of this overall problem is the limited, but by no means simple task of constructing a formal procedure to aid in the formulation, articulation, and measurement of worth notions. Once a worth measure has been achieved, it still remains to combine this with risk/uncertainty considerations (e.g., by defining a utility measure) and then to compare the result with whatever resources must be expended (e.g., by defining a satisfactory decision rule) before a final decision can be reached.

The great bulk of this chapter has been devoted to laying a conceptual foundation for the worth formulating and measuring procedure to be developed in Chapter II. In addition, an attempt was made to delimit the issue of worth determination from the related issues of risk/uncertainty, classical utility, resource expenditures, and an integrating decision rule. Let us now proceed to the development of a systematic procedure for formulating and measuring notions of worth.

CHAPTER II

A SYSTEMATIC PROCEDURE FOR FORMULATING AND MEASURING NOTIONS OF WORTH

2.1.0 GENERAL OUTLINE OF THE PROCEDURE

In Section 1.1.0 of this thesis it was pointed out that worth assessment is an especially difficult task in the case of "complex" alternatives due to:

1. Multiple objectives and assessment criteria to list and arrange in some organized fashion;

2. Multiple performance consequences to predict;

3. Multiple worth connections between listed assessment criteria and predicted performance consequences;

4. Physical interaction among performance consequences;

5. Worth interdependence among assessment criteria.

However, the scope of this task can and will be reduced somewhat by making two simplifying assumptions. First, it will be assumed that validated estimates are freely available for all relevant performance consequences associated with all produced alternatives. Naturally, both obtaining and validating such estimates constitute very real and highly important problems in their own right, but neither of these will be discussed in this thesis. Such omissions are purely for simplification and should not be construed as devaluating either the difficulty or the importance of estimating accurately and validating performance consequences.

Second, it will be assumed that our task is restricted to assessing a fixed set of discrete alternatives. The problem of producing alternatives (i.e., of designing, redesigning, or soliciting proposals for alternatives) will not be considered. This assumption reduces substantially any worries we might otherwise have had concerning physical interaction among performance consequences, since physical interaction is troublesome primarily because it renders prediction of performance difficult under differing design alternatives.[1]

Nevertheless, in spite of these simplifications, we must still worry about listing and arranging multiple objectives and assessment criteria, checking for worth interdependence among them, and establishing worth connections between these criteria and various performance consequences. The remainder of this section will address itself to these tasks.

2.1.1 Listing Major Performance Objectives

The first step in making a formal assessment is to specify what is desired in the way of performance from whatever alternatives may be produced. This means listing objectives. At the outset, objectives may be (and should be) stated in very general terms. After all, the point is to be as all-encompassing as possible initially (to

---

1. The scope of our problem is greatly reduced by this assumption, but not to the point where it no longer possesses practical significance. After all, in any real decision situation, there comes a moment when the process of producing alternatives must be terminated, and an immediate choice must be made among whichever alternatives have already been produced. At that moment of decision it is reasonable to view the choice as among a fixed set of discrete alternatives.

avoid omitting any important objectives which decision makers really possess and are willing to display), and then to work down through a process of successive elaboration to a very specific statement of desired performance. A very specific statement of intentions is required at the end of the process in order to carry out an actual assessment, but this need not concern us too heavily at the beginning. Rather, what should concern us is summarized below.

Any list of major performance objectives should possess the following desirable properties.

1. The list should be complete and exhaustive. That is, all important performance considerations deemed relevant to the final decision should be reflected by the items on the list. This is to guarantee that no important performance considerations are overlooked by the assessment procedure.

2. The list should contain only performance objectives which are mutually exclusive. That is, no listed objective should encompass or be encompassed by (either wholly or in part) any other objective on the list. This is to insure that the assessment procedure will be free of undesirable "double-counting" in the worth sense.

3. The list should be non-redundant. That is, only performance objectives which will be deemed at least somewhat important to the decision should be included in the list. This is to avoid compiling a list of unmanageable size.

4. The list should be restricted to performance objectives of the highest degree of importance. that is, only major objectives should be included. Any objective or criterion which is contained within the meaning of a higher objective does not belong on the list. Similarly, any objective or criterion which is important only because it contributes to satisfying some other objective does not belong on the list. The purpose of these exclusions is to provide a sound basis or starting point from which lower-level criteria may be derived.

5. Finally, the list should contain objectives relatively independent in the worth sense. That is, for any pair of objectives on the list, decision makers should be willing to trade-off or exchange additional satisfaction on one objective for reduced satisfaction on the other at a rate relatively independent of the levels of satisfaction already attained on each of the listed objectives. The rationale underlying this requirement will be made clear in Section 2.1.6. At this point, let it suffice to say that decision makers generally find it difficult to conceive of and, therefore, to assign meaningful worth measures to observed performance on the basis of objectives which are highly interdependent in the worth sense. Consequently, it would be foolish to permit such interdependent objectives to infiltrate the list.

2.1.2. Generating A Hierarchical Structure Of Performance Criteria

Having established a list of major performance objectives, the

second step is to generate a hierarchical structure of successively more specific performance criteria. This involves breaking down or subdividing higher-level criteria into one or more lower-level criteria. The purpose of subdividing is to define more precisely (i.e., in terms of the lower-level criteria) what is intended by or included within the meaning of a higher-level criterion. But what, exactly, is the nature of this problem?

Essentially, our problem is to create a pictorial map of the structure of worth relationships existing within the mind of a decision maker. Just as a cartographer attempts to depict topographical relationships of distance, elevation, contiguity, etc., between masses of land and water in some specified geographical region, we are trying to depict worth relationships between overall performance objectives and successively lower levels of increasingly more specific performance criteria relevant to the selection of a specified alternative for some definite job. Just as the cartographer utilizes certain conventions such as contour lines and special coloring to convey information about the terrain he is describing, we also can use conventions such as pyramidal or treelike arrays to convey information about the worth structure.

Despite these similarities, however, there are a number of important differences between constructing maps of regional topography and maps of human worth structure. First, the cartographer attemps to describe various aspects of our physical environment. We, on the other hand, are attempting to describe various aspects of the inner minds of human decision makers. This suggests that the proper focus of our attention is <u>not</u> the "out-there" physical world of nature, but rather the "in-here" subjective world of human beings. It is to decision makers and

their evaluative responses that we must look in constructing our map.

A second difference follows immediately from the first. Since the cartographer is attempting to map something physical and directly observable, he may utilize direct measuring devices such as compasses and other surveying tools. We, on the other hand, are attempting to map something non-physical and only indirectly measurable. We are therefore forced to utilize indirect measuring devices such as introspection and verbal questioning. From personal thoughts and verbal responses we must <u>infer</u> the underlying structure of human preferences.

A third difference relates to the number and temporal stability of the entities being mapped. Whereas there is only one topographical region to be investigated by the cartographer (the particular region he is interested in mapping), there are frequently more than one decision makers to be investigated in mapping a worth structure (the group of decision makers responsible for making a selection decision). In addition, topographical features of our physical environment are apt to be highly stable over time, while attitudinal features of our assessment structure are apt to change over time with new learning and increased assessment experience.

A fourth difference, and by far the most important one, relates to the perturbing effect of the mapping process itself. The cartographer is concerned with depicting visually a territorial region which has already been formed by the forces of nature. His mapping process does not in any way alter the nature of the physical terrain being mapped. In direct contrast, our mapping process has an enormous impact upon the worth structure being mapped. On the basis of an experiment to be reported in Chapter V, we shall conclude that the single most important

consequence of the entire assessment procedure is to create a worth structure where one did not previously exist - at least not in conscious, well-defined, and easily articulatable form. Participation in this assessment procedure induces the decision maker to formulate a consistent worth structure. At the very least, this entails substantial clarification of what already existed in his mind. Typically, it induces him to alter substantial portions of his prior worth structure. At most, it induces him to create a structure which did not enjoy any prior existence at all in consciousness. Producing a pictorial map of the worth structure, once formulated, constitutes a separate and important consequence of the assessment procedure, but this is not the only consequence, nor is it the most important one.

More will be said later about this joint process of formulating and representing a worth structure. For now, however, we shall concentrate primarily upon the representational or mapping aspects of the process.

From the preceding discussion we may draw the following conclusions.

1. Our current task is to map the structure of worth relationships relevant to selecting an alternative for a specified job.
2. The structure we are trying to map exists only within the mind of a decision maker.
3. Therefore, to draw our map, we must look to the decision maker and his behavioral responses.
4. Since we are mapping an essentially subjective entity, we are restricted to the use of indirect rather than direct measuring devices. We must rely upon intro-

spection, verbal questioning, and observation of behavior.

5. Also, for the same reason, we must expect lack of consensus (at least initially) among different decision makers as to the nature of the worth structure. We must also expect changes over time within the same decision maker's mind as he gains increased knowledge and experience.
6. Finally, we must always remember that what is being mapped is not a fixed entity, but that the formulating and mapping processes are dynamically and intimately interrelated.

2.1.3 Selecting Physical Performance Measures

The third step is to select a single physical performance measure for each lowest-level performance criterion on the hierarchical criterion structure. The purpose of selecting physical performance measures is to give concrete, physical interpretations to their related lowest-level criteria. By this device, a bridge is constructed linking the subjective inner minds (i.e., the worth structures) of decision makers to the objective outer world of physical alternatives. Let us clarify this concept - particularly the distinction between performance criteria and physical performance measures - with further discussion.

A physical performance measure is any tangible reading or concrete observation that can be extracted from the real world. For purposes of assessment, it is any directly measurable attribute of a produced alternative.

Notice, however, that this is not the same thing as a performance criterion. A performance criterion, it will be recalled, is a reflection of how decision makers choose to assess physical performance. Performance criteria are only indirectly measurable and must be inferred via introspection and/or verbal questioning from human responses. Performance criteria are attributes of decision makers, while performance measures are attributes of the physical alternatives they are assessing.

Although this may sound like a mere academic distinction, it will be useful for very practical reasons to keep these two concepts clearly separated. There are three reasons for maintaining the distinction. First, the methods of approach and the people one talks to in formulating performance criteria are different from the methods and people involved in defining physical performance measures. Introspective reflection and discussions with fellow decision makers must be relied upon to formulate, to clarify, and to understand performance criteria. These, after all, reflect what is desired (by decision makers) from an alternative. In contrast, inspection of physical alternatives and discussion with knowledgeable engineers would seem a more useful way to define physical performance measures. These reflect what an alternative will deliver (no matter what is desired).

A second reason for distinguishing between performance criteria and performance measures springs from the very different way in which they will be treated in the process of formal assessment. Once defined, physical performance measures will be used to describe each of the produced alternatives. The description of an alternative in terms of a set of physical performance numbers (and/or other descriptive symbols) will then be converted into equivalent worth point scores by

means of a device called a scoring function (to be discussed in Section 2.1.4).

In direct contrast, worth scores attached by scoring functions to lowest-level performance criteria are not themselves run through scoring functions. Instead, they will be combined with other worth scores already attached to other performance criteria. Such combination will be carried out by means of a device called a weighting function (to be discussed in Section 2.1.5), and the result will be a single, overall index of worth associated with each produced alternative.

A third reason for maintaining the distinction involves the handling of interdependent performance criteria (to be discussed in Section 2.1.6). It will be demonstrated that interdependent performance criteria can sometimes be eliminated by defining a high-level performance measure. But this is getting too far ahead of the story. Let it suffice to say that there are good, practical reasons to keep the distinction between performance criteria and physical performance measures firmly in mind. Some of these reasons will become more clear as the discussion continues.

Returning to the problem of actually selecting physical performance measures, this must be done subjectively. A decision maker must choose a well-defined and easily measurable physical attribute of an alternative which he feels serves to interpret, in phenomenological terms, the meaning of the lowest-level criterion under consideration. Thus, in the case of selecting computers, the performance criterion "printer speed" might be interpreted by means of the physical performance measure "maximum number of lines printed per minute, assuming no jamming or other form of breakdown."

Selecting performance measures raises two questions. First, how does one come up with a likely candidate for selection? Second, if more than one candidate arises, how does one choose from among them? Coming up with candidate measures, like generating subcriteria in filling out the hierarchical structure, requires imagination, ingenuity, and intuition. Both tasks involve creative acts. However, both tasks will be aided considerably if decision makers take the trouble to compile a master list of performance criteria and physical performance measures from which to extract or synthesize particular criteria and measures when needed for a particular decision. This master list might contain all performance criteria and all physical performance measures that have ever been suggested and/or used on past decisions of a similar nature.

As for choosing from among alternative candidate measures to associate with a given performance criterion, this also requires personal judgment. It may happen, for example, that certain printers are known to jam frequently under continued, high-speed operation. Under such circumstances, a better measure of "printer speed" might be "expected number of lines printed per minute" with jamming and other forms of breakdown taken into consideration by means of historical breakdown frequency data relating to each type of printer. However, this kind of choice must be made by decision makers on the basis of historical evidence, their own experience, and personal judgment.

2.1.4 Establishing Specific Worth Relationships Between Lowest-Level Performance Criteria and Their Associated Physical Performance Measures: The Scoring Problem

The fourth step in constructing a formal assessment procedure is to establish specific worth relationships between each lowest-level performance criterion in the hierarchical structure and its associated

physical performance measure. Selecting performance measures (the step just discussed in Section 2.1.3) serves to establish the existence of worth connections, but it does not serve to map out specific worth relationships. Specific relationships are established by means of scoring functions.

A scoring function is a mathematical rule which assigns a unique worth score in points to every possible value of some physical performance measure. It transforms raw performance (measured in terms of whatever physical unit is appropriate to the performance measure under consideration) into worth-of-performance (measured in terms of the worth points discussed in Section 1.4). Just as the selection of a physical performance measure serves to interpret concretely each lowest-level performance criterion and, therefore, to provide a bridge between the subjectively defined worth structure and the objectively defined physical characteristics of an alternative, the specification of a scoring function serves to define precisely the nature, shape, and particular parameters of this bridge.

In formulating a scoring function, it is temporarily assumed that the lowest-level performance criterion in question constitutes the only performance objective in the entire assessment. Then, the worth score assigned by the scoring function to any given amount of performance on the associated physical performance measure is supposed to indicate the extent to which that amount of physical performance actually satisfies the lowest-level criterion. To accomplish this, certain conventions or ground rules must be observed uniformly to insure that all worth scores thereby generated will be comparable with one another and subject to a uniform interpretation. Otherwise, the subsequent procedure by which individual worth scores assigned to separate criteria

are to be combined cannot be meaningfully carried out. A set of scoring conventions designed to insure both consistency and comparability appears below.

1. The outputs of all scoring functions will be in terms of worth points.

2. Worth points will be as defined in the ten scaling conventions presented in Section 1.4.3.

3. All scoring functions will be formulated to cover the entire range of logically possible physical performance - not just the reasonable or expected range. This is to insure that a definite point score will be defined for every conceivable level of produced performance - no matter how unexpected it may be.

4. All scoring functions will be formulated independently of (i.e., without reference to) physical performance actually offered by produced alternatives. This ground rule follows directly from our previous discussion appearing in Section 1.4.5.

5. Hopefully, all scoring functions can be formulated prior to inspecting any performance date. This will help to guarantee that scoring is performed on an independent basis.

6. Most scoring functions will take the form of mathematical formulas and/or graphically depicted mathe-

matical curves. However, some will not be expressed in these terms. Some will take the form of direct judgmental point assignments by decision makers without the aid of either formulas or graphs. In this latter case, scoring functions are thought of as implicit within the minds of decision makers rather than explicitly stated in any precise mathematical form.

7. All scoring functions will be formulated by means of a single, uniform, and replicable procedure. A suggested two-phase procedure (embodying the above six scoring conventions) will be presented in Section 2.2 of this thesis.

2.1.5 Establishing A Procedure For Combining Worth Scores Assigned On The Basis Of Separate Performance Criteria To Arrive At A Single, Overall Index Of Worth: The Weighting Problem

In discussing scoring functions, it was suggested that one temporarily assume each lowest-level performance criterion under consideration to be the sole performance objective in the entire assessment. Obviously, this is an untenable assumption. There are many performance objectives to be satisfied as reflected in the hierarchical structure with its many lowest-level branches. This brings us to the fifth step in formal assessment - combining worth scores assigned on the basis of separate performance criteria to arrive at a single, overall index of worth. This step will be accomplished by defining a weighting function.

A weighting function is a conceptual device by means of which

explicit recognition is given to the existence of multiple objectives and performance criteria. Whereas a scoring function is defined to indicate the extent to which any given level of measured performance succeeds in satisfying its related lowest-level performance criterion, a weighting function is defined to indicate the perceived relative importance of satisfying the criterion itself vis a' vis other performance criteria. In this manner, the temporary assumption of a single criterion made in defining a scoring function is relaxed to reflect reality. Simultaneously, a means of combining worth scores assigned on the basis of separate criteria into a single, overall index is achieved. Let us illustrate these results by means of a very simple example.

Suppose that a computer is to be acquired and operated with two specific objectives in mind. These are:

1. to perform a current job clearly described in terms of workload and applications;

2. to expand at some future date beyond the currently stated workload and/or applications.

Suppose, also, that performance of the current job is to be measured by the time required (in hours) by each produced alternative to process a benchmark program simulating a day's workload and that an appropriate scoring function has been defined to convert all possible benchmark times into equivalent worth scores. Finally, assume that expansion potential is to be measured by immediate-access memory capacity (in number of words) unused by the benchmark exercise and that an appropriate scoring function has also been defined for this performance measure. Then, each alternative computer would be assigned two worth scores - one for performing the current job and another one for expansion potential. How

can these two separate scores now be combined into an overall index of each alternative's total worth? This is the weighting problem.

One way to proceed would be as follows. Decision makers ask themselves which of the two performance criteria - doing the current job or providing expansion capability - should be considered more important. That is, if given the choice between satisfying either of the two criteria to the same extent, which one would they prefer to have satisfied? Alternatively, would decision makers feel genuine indifference in choosing between equal percentage satisfaction of the two criteria? If decision makers would prefer to have the current job criterion satisfied over having the expansion potential criterion satisfied to the same extent, then the former criterion must be considered more important than the latter. If genuine indifference is felt between having the two criteria equally well satisfied, then they must be regarded as equally important.

The next step is to be a bit more precise about the extent or degree of perceived relative importance. Just to say that doing the current job is more important than providing expansion potential is usually not sufficient to distinguish clearly between the overall worths of competing alternatives. The magnitude of this perceived relative importance must also be indicated. How much more important is it to satisfy the current job criterion than to satisfy the expansion potential criterion? Twice as important? Ten times as important? Representation of relative magnitudes once again suggests resorting to numbers.

Suppose that performing the current job were considered three times as important as providing expansion potential. Then, any pair of numbers standing in the ratio of 3:1 could be used to convey this inform-

ation about perceived relative importance. In particular, the numbers 3/4 and 1/4 could be used. Then, whatever scores are attached by scoring functions to these two criteria could be combined by:

1. multiplying the score assigned to performing the current job by 3/4;

2. multiplying the score assigned to expansion potential by 1/4;

3. adding the two products to arrive at a weighted average score, using the importance ratios as constant weights.

The resulting sum of weighted scores might then be interpreted as an overall index of each alternative's total worth.

The above procedure has a definite appeal in its simplicity and directness. It seems to solve the problem of combining scores on separate criteria, and it seems to arrive at a single, overall index of worth. What's more, by requiring the set of constant weights to add internally to one (as was done in the example above), the resulting overall worth score (computed as the sum of weighted individual criterion scores) also lies between zero and one and may be subjected to the same interpretation as worth point scores assigned to individual performance criteria. This renders far more manageable the task of checking assigned weights for intuitive reasonableness and consensual validation. The same questions may be asked of weighted sums as are asked of individual criterion scores. Since worth scores cannot be validated by any other means (recall that they are <u>in principle</u> untestable by ordinary scientific techniques), uniform interpretability becomes an extremely important and valuable asset.

However, in spite of its simplicity and immediate appeal, the

above procedure should be subjected to critical scrutiny before accepting it and incorporating it into a formal assessment scheme. It would be wise to inquire a bit more carefully into what this weighting procedure is really assuming about how decision makers view multiple assessment criteria, how they trade off worth among multiple criteria, and what procedural implications these assumptions have for the practical task of assessment. It will be shown that the key to understanding these issues lies in the concept of worth interdependence among separate performance criteria. This concept and its procedural implications will be discussed in the next section.

2.1.6 Identifying And Eliminating Worth Interdependence Among Separate Performance Criteria

The preceding example of combining worth scores by means of weighting and summing to arrive at a single index of overall worth assumes implicitly the following things.

1. The relative importance of satisfying separate performance criteria does not depend upon the various degrees to which each separate criterion has itself been satisfied by some amount of measured performance. Rather, their relative importance is conceived as being constant in this respect.

2. The rate at which a given increase in the degree to which any criterion is satisfied contributes to overall worth is independent of the levels of satisfaction already achieved on that and other criteria. Rather, such rates are viewed as constant in this respect.

3. The rate at which decision makers would be willing to

trade off decreased satisfaction on one criterion for increased satisfaction on other criteria so as to preserve the same overall worth is <u>independent</u> of the levels of satisfaction already achieved by any and all of the criteria. Such trade-off rates or rates of substitution are viewed as <u>constant</u> in this respect.

These three logically interrelated statements, taken together, define the concept of worth independence among separate performance criteria. This concept is represented mathematically in the weighting function by defining <u>constant</u> relative importance or trade-off rates and using these as weights by which to first multiply individual criterion scores and then sum weighted scores to arrive at an overall index of worth.

To clarify further this concept of worth independence and, more particularly, to distinguish it from its opposite, worth interdependence, let us consider two contrasting examples. First, we shall return to the example given in the preceding section and argue that performing the current job and providing expansion potential constitute worth-independent criteria for assessing performance. Then, we shall concoct a counter-example to illustrate worth-interdependent criteria.

In our previous example, the two performance criteria were assumed to be:

1. system performance potential with respect to the currently stated workload and applications;

2. expandability of the system beyond the currently stated workload and applications.

Now it is claimed that these two criteria could easily be viewed as independent of one another in the worth sense. That is, within the bounds

of probable proposed performance, many decision makers would be willing to exchange additional performance potential with respect to current needs for reduced expandability to accommodate future needs (or vice versa) at the same rate, no matter how much current performance potential or how much future expandability a proposed system already promises. Naturally, this statement is further qualified by restricting trade-offs or exchanges to those situations in which all relevant mandatory requirements are met. Otherwise, the system would not work at all, in which case it would be superfluous even to discuss trade-offs.

The basis for claiming that these major performance criteria are relatively independent is strictly intuitive. Current performance potential and future expandability are generally perceived as relatively independent objectives. Decision makers would like both objectives to be satisfied, but the desire to have either one satisfied is relatively uninfluenced by the degree to which the other has already been satisfied. No matter how well an alternative computer performs the currently stated job, decision makers would still prefer that it be expandable; and no matter how expandable it is, they would still prefer improved performance on the current job. Satisfaction of either objective is no substitute for satisfaction of the other. Both are independent objectives, desired in their own right. Consequently, they are represented by performance criteria claimed to be relatively independent in the worth sense.

Now let us concoct an example of substantial worth interdependence. Suppose that, in terms of expanding the job beyond the currently stated workload and applications, decision makers perceive as valuable the presence of additional memory capacity. This might be provided either by additional core storage capacity, by additional disk capacity, by additional drum capacity, or even by additional tape capacity.

Suppose, further, that decision makers attempt to reflect this by subdividing the higher-level criterion, additional memory capacity, into four sub-criteria:

1. additional core storage capacity;

2. additional disk storage capacity;

3. additional drum storage capacity;

4. additional tape storage capacity.

Now it is claimed that these four sub-criteria are not independent in the worth sense. They are highly interdependent. None of them are generally perceived as independent objectives, desirable in their own right. They constitute four alternative means of satisfying the same single objective--providing additional memory capacity. They are worthwhile only insofar as they accomplish this end. Because of this, satisfying any one of these four sub-criteria is a more or less acceptable substitute for satisfying any of the others, and the rate at which decision makers would be willing to exchange additional satisfaction on any one of the four for reduced satisfaction on any of the other three depends critically upon the extent to which any or all of them have already been satisfied.

Having illustrated the twin concepts of worth independence and worth interdependence and having shown that an additive scheme using constant trade-off weights implicitly assumes worth-independent criteria, the time has come to ask two more fundamental questions. First, what happens if constant trade-off weights are applied to worth-interdependent criteria? The answer is that easily computable, but uninterpretable and meaningless numerical results would be obtained. This is very dangerous.

Numerical sums of weighted scores would be defined mathematically, but it would not be at all clear what these results signified (if anything) in terms of overall worth. This suggests that one of two remedial steps be taken under such circumstances. Either a different weighting function involving variable trade-off weights which depend upon the levels of criterion scores should be defined in place of the simpler function involving constant weights; or the simpler weighting function should be retained, and the hierarchy of performance criteria should be purged of worth-interdependent members. Alternatively, some combination of the above two remedies might be adopted.

In choosing a remedy, we are forced to ask a second fundamental question. How do decision makers really think about making worth trade-offs among separate criteria? How flexible are they in conceptualizing the issue? Judging from the frequent occurrence of additive combinatorial schemes using constant trade-off weights both in government and industry, it would appear that decision makers characteristically prefer to represent their trade-off notions in this very simple manner. Previous attempts on the writer's part to induce numerous decision makers to articulate their trade-off notions in a more complex manner all ended in confusion and failure. Now this does <u>not</u> mean that <u>all</u> notions of worth trade-offs are or should be viewed in terms of an independent and additive scheme. A counter-example of substantial worth interdependence has just been presented. It just means that this seems to be a <u>convenient</u> way to think and that more complex ways of thinking may be too difficult to articulate explicitly and, therefore, to incorporate systematically within a formal assessment scheme.

In light of these considerations, the writer has chosen to adopt a modified version of the second remedial strategy. That is, the

formal assessment procedure to be developed herein will be based on an independent and additive weighting function. To make the procedure meaningful in terms of interpreting its numerical results, instances of substantial worth interdependence will be detected and purged from the hierarchical criterion structure.

This decision is an essentially pragmatic one. It is based on the assumption that decision makers would find it too difficult in practical situations to articulate trade-off relationships in a more complex manner. If subsequent research should indicate this to be a false assumption, the decision might appropriately be reversed.

On the other hand, complete elimination of all instances showing even the slightest trace of worth interdependence might reduce the hierarchical structure to almost nothing. This would serve only to eliminate from explicit consideration many important aspects of the assessment problem. Therefore, only instances of <u>substantial</u> worth interdependence will be purged from the hierarchy. A specific procedure for identifying and eliminating instances of substantial worth interdependence among performance criteria will be presented in Section 2.2 of this thesis.

2.1.7 <u>The Meaning And Interpretation Of Weights</u>

Just as it was useful to establish by means of explicit scale conventions the meaning and proper interpretation of worth point scores, so also is it useful to establish a similar logical basis for numerical weights. This will be accomplished by stating and discussing briefly ten weighting conventions.

1. A set of numerical weights will be defined for every set of sub-criteria into which a higher-level criterion in the hierarchical criterion structure is subdivided. In the case of the highest-level or major performance

criteria, these are construed as "sub-criteria" of "overall worth" and, therefore, each of these will also receive a numerical weight. In all cases, a single weight will be defined for each such sub-criterion.

2. The numerical weight attached to each sub-criterion will be interpreted as an indication of the perceived relative importance of satisfying that sub-criterion in the context of the higher-level criterion within whose meaning it is alleged to be included. Relative importance means "relative to the other sub-criteria in the set."

3. Relative importance will be reflected in the ratios of any two weights assigned, respectively, to two separate sub-criteria in a given set. It is in such ratios that trade-off rates or rates of substitution will be embodied.

4. Weights will be assigned only to sub-criteria perceived as devoid of substantial worth interdependence. A definite procedure will be presented to identify and eliminate sub-criteria displaying substantial worth interdependence.

5. Weights will be restricted to fall within the range of non-negative numbers. This is to indicate that the concept of relative importance possesses only "positive" connotations. Restricting weights to fall within the range of non-negative numbers guarantees that all

trade-off rates (i.e., all ratios between pairs of weights) will be non-negative.

6. Theoretically, a weight of zero would be assigned to any sub-criterion in a given set of sub-criteria if and only if satisfying that sub-criterion were perceived as completely unimportant. In practice, however, a sub-criterion to which a zero weight might appropriately be attached will be ignored (i.e., such a sub-criterion will <u>not</u> be included in the hierarchical criterion structure), since, by the above definition, its satisfaction is viewed as totally unimportant. This definition is included only to provide a logical lower bound to the range of permissible weight numbers and to give the lower bound a definite interpretation.

7. All of the weights in any given weight set (corresponding to a given set of sub-criteria) will add to a finite positive constant, and the same positive constant will apply to all weight sets. This serves to normalize assigned weights so that a given weight number will always have the same significance (i.e., indicate the same relative importance) in all weight sets. Consequently, the task of validating weight assignments by visual inspection becomes easier.

8. The finite positive constant to which all weights in any given weight set add will be one. Any such constant would be permissible, but setting this number

equal to one has a certain conceptual appeal. Since all weights are non-negative and add to one, each weight must lie between zero and one. Hence, relative importance may be viewed as if it were a percentage or proportion, which decision makers may find to be a convenient and familiar conceptual aid.

9. Assigned weight numbers cannot exceed one, and a weight of exactly one will only be assigned in cases where a set of sub-criteria contains a single member. Then, that single sub-criterion must receive full-weight. As such, it must be interpreted as completely synonymous in the worth sense with its related higher-level criterion.

10. Any positive real number equal to or less than one will be a permissible weight. This is to permit the formation of any desirable trade-off ratio between pairs of weights.

2.1.8 Adjusting The Weights To Reflect The Relative Interpretability Of Each Physical Performance Measure

Another issue, which has not yet been discussed, concerns the relative extent to which each physical performance measure previously selected to interpret (in physical terms) its associated lowest-level performance criterion does in fact succeed in providing an adequate interpretation thereof.[2] Decision makers might view "expected number of lines

2. The writer is indebted to H. Martin Weingartner for originally raising and noting the importance of this issue. The writer is also indebted to Howard Raiffa for criticising constructively the particular manner in which this issue is treated in the assessment procedure.

printed per minute" as an excellent measure of the lowest-level criterion "printer speed". This is becuase it reflects very well the intended meaning of "printer speed" within the context of the particular job under consideration. In contrast, "total number of discrete promises" found in a formal proposal submitted by a computer manufacturer to perform that job might be considered a poor measure of "manufacturer's good faith". This is because "manufacturer's good faith" refers to an attitude on the part of corporate executives, and this attitude may not be clearly reflected in the text of their formal proposal. Discussions with executives and review of their historical behavior in similar contractual situations should provide vastly superior measures of their good faith.

To the extent that wide differences emerge in the relative interpretive quality of various performance measures, this could have a seriously distorting impact upon the outcome of a decision. It is quite conceivable that a relatively important criterion (deserving a large numerical weight) cannot be interpreted with any measures of good quality because the decision maker is unable to articulate in explicit physical terms what he means by this criterion. The decision, therefore, should not be unduly influenced by such criteria, especially if other criteria - even though considered relatively less important - are much easier to interpret in terms of high-quality measures. In short, there should be some explicit mechanism for reflecting the relative quality of each criterion's interpretive measure as well as the relative importance of satisfying that criterion. A procedure will be presented in Section 2.2 to achieve this result.

2.1.9 Summary

The first step in formal assessment is to define explicitly what is desired in the way of performance from produced alternatives to complete a stated job. This means listing overall objectives or major performance criteria and insuring that the list is:

1. complete (i.e., contains all criteria which decision makers are able and willing to formulate and display);

2. mutually exclusive (i.e., contains criteria which neither encompass nor are encompassed by other criteria on the list);

3. non-redundant (i.e., contains only criteria which are deemed at least somewhat important);

4. of major significance (i.e., contains only highest-level criteria);

5. free of worth interdependence (i.e., contains only worth-independent criteria).

Having established a list of major performance objectives, the second step is to generate a hierarchical structure of successively more specific performance criteria. This involves breaking down or subdividing higher-level criteria into one or more lower-level criteria alleged to be included within the meaning thereof.

The third step is to select a single physical performance measure for each lowest-level performance criterion on the hierarchical structure. The purpose of selecting physical performance measures is to establish concrete connections between the hierarchical criterion structure (existing in the subjective minds of decision makers) and the outer world of

physical alternatives.

However, merely establishing connections is not sufficient in itself to permit formal assessment. Specific worth relationships must be mapped out between each lowest-level performance criterion and its related physical performance measure. This constitutes the fourth step. It is implemented by defining scoring functions which assign a unique worth score in points to every possible value of a physical performance measure. Scoring functions must be defined, either explicitly or implicitly, for every lowest-level criterion.

The fifth and final step is to combine worth scores assigned on the basis of separate performance criteria to arrive at a single, overall index of worth. This is accomplished by defining a weighting function. An additive weighting function with constant weights is being adopted for this purpose. However, this requires that all sub-criteria intended by or contained within the meaning of a higher-level criterion be relatively independent of one another in the worth sense. Consequently, special procedures are required to identify and eliminate instances of substantial worth interdependence from the hierarchical criterion structure. In addition, a special procedure is required to adjust relative importance weights to reflect the relative interpretive quality of each physical performance measure. This procedure is implemented after defining, but prior to applying the weighting function.

* * * * * * * *

Having moved through an outline of the assessment procedure, the reader now faces two choices. If detailed familiarity with the procedure is desired, continue to the next page and complete Chapter II. If general familiarity is considered sufficient, proceed immediately to Chapter III.

2.2.0 A SPECIFIC STEP-BY-STEP PROCEDURE FOR ASSESSING WORTH

The procedure to be presented in this section is first of all intended to generate an assessment algorithm. This algorithm is supposed to encapsulate the worth notions of a particular decision maker (or group of decision makers) at a particular point in time with respect to a particular and clearly specified job. Once generated, the algorithm may then be applied to any feasible alternative produced to accomplish that job. Application of the algorithm to any one of the alternatives converts a description of that alternative, in terms of physical performance measures, into a single, overall index of that alternative's worth. It will be well to keep in mind the two-stage nature of the assessment procedure (i.e., first generate an assessment algorithm, and then apply the algorithm to generate a worth measure for each produced alternative). Otherwise, a confused interpretation may very likely result.

It is assumed that the following preliminary steps have been successfully completed prior to embarking upon the assessment procedures.

1. The job for which produced alternatives are being assessed has been adequately described.

2. From the job description a set of mandatory performance (and possibly resource) requirements has been extracted and recorded in physical terms.

3. At least one alternative has been produced.

4. The performance and resource estimates associated with each produced alternative have been validated (i.e., investigated for accuracy and truthfulness).

5. These validated estimates have been checked against stipulated mandatory requirements, and at least one alternative has been shown to be feasible.

6. Any alternatives which failed to satisfy one or more stipulated mandatory requirements have been deleted from further consideration.

7. The residue of feasible alternatives is to be assessed formally.

8. The first step in the formal assessment procedure has been completed, at least tentatively. A list of major performance objectives has been formulated with reference to the job description and in accordance with the ground rules set forth in Section 2.1.1.

Let us now proceed to the task of formal assessment.

2.2.1 A Procedure For Generating Sub-criteria: Filling Out The Rest Of The Hierarchical Criterion Structure

As mentioned previously, it will be most helpful for implementing this and subsequent procedures if a master list of candidate performance criteria and performance measures has been compiled in advance. Although not necessary, experience has shown that reference to such a master list facilitates considerably the essentially creative process of filling out a criterion hierarchy and selecting performance measures. For purposes of discussion, it will be assumed that such a master list exists.

Beginning with one of the major or highest-level performance criteria, we ask what this means in the context of the stated job. To

render the discussion concrete, let us return to the example of assessing a computer and select the major criterion, "system performance potential with respect to the currently stated workload and applications". With reference first to the job description, we might decide that the following sub-criteria are all intended by or subsumed under this major criterion:

1. capability to perform certain functions stated in the job description (e.g., process 1000 payroll records) within a stated time limit (e.g., weekly);

2. additional equipment capabilities related to performing the stated job.

Next, with reference to the master list, we might discover a third sub-criterion omitted from the formal job description, but which we also consider to be a part of system performance potential in the stated job context. This third item might be:

3. reasonable reliability in performing the stipulated functions.

If we feel that this more or less exhausts the intended meaning of system performance potential in the stated job context, we can put this aside temporarily and proceed to repeat the process on the three sub-criteria just defined.

Beginning with the first sub-criterion, we again ask ourselves the same question. What does capability to perform certain stated functions mean in the context of the job? We might decide to define the following sub-sub-criteria in response to this question:

1. capability to solve a prepared benchmark problem or set of problems within a certain time limit -

preferably in less time;

2. capability to perform certain other standard data processing tasks not reflected within the benchmark exercise, but anticipated within the job environment.

At this point, we might decide that further conceptual subdivision is unnecessary. The time has come to select physical performance measures for each of these two sub-sub-criteria (e.g., observed number of hours required to perform the benchmark problem or problems and observed or estimated number of minutes required to perform a particular mix of standard data processing tasks).

Returning to the second sub-criterion (additonal equipment capabilities), we might decide to subdivide this into:

1. additional speed capabilities;

2. additional capacity capabilities;

Then, additional speed capabilities might be further subdivided into:

1. instruction speed;

2. peripheral equipment speed.

For instruction speed, we might select a physical performance measure (e.g., average time in microseconds required to process a single instruction) and then subdivide peripheral equipment speed even further into:

1. card reader speed;

2. card punch speed;

3. tape speed;

4. printer speed.

For these four criteria we might select appropriate physical performance

measures (e.g., maximum number of cards per minute, maximum number of characters per second, maximum number of lines per minute), which would close out these branches of the hierarchical tree.

The hypothetical example partially developed in the preceding paragraphs could be carried further, but the general idea should by now be clear. One starts at the highest level of the hierarchy with one of the major performance criteria, asks himself what this means, defines one or more sub-criteria in response to this question, and then repeats the procedure with each of the defined sub-criteria. This process continues until it is decided that further subdivision is unwarranted. A physical performance measure is chosen, and that branch of the tree is considered filled out. A retreat is then made back up the tree to the first level containing incomplete branches. The process of successive subdivision is initiated at that point and carried out until another physical performance measure is defined. By so moving up and down the tree, an entire hierarchical structure may be generated. The final signal to stop occurs when no more incomplete branches exist (i.e., when physical performance measures have been attached to every branch of the tree).

Because the process just illustrated is recursive (i.e., because it involves successive reapplication of the same sequence of steps to move up and down the hierarchical tree), only the reiterated sequence of steps need be specified in any great detail to describe completely the entire process. A formal presentation of this reiterated sequence of steps follows immediately.

Step 1. Locate an incompleted branch on the hierarchial tree (i.e., any major criterion or sub-criterion without an attached physical performance measure). At the outset, incompleted branches will occur only at the top level of major performance criteria.

Step 2. With reference to the job description and to the master list, decide whether the criterion under scrutiny is to be further subdivided or interpreted directly by means of a physical performance measure. If it is to be further subdivided, proceed to Step 3. If a physical performance measure is to be selected for it, proceed to Step 5.

Step 3. Again, with reference to the job description and to the master list, subdivide the criterion under scrutiny into one or more sub-criteria. That is, decide what sub-criteria are intended by or logically subsumed beneath the criterion under scrutiny. All ~~Each~~ of these now constitute new incompleted branches of the hierarchy.

Step 4. Choose any one of the sub-criteria defined in Step 3 as a starting point and return to Step 1.

Step 5. With reference to the job description and to the master list, select a physical performance measure judged relevant to the criterion under scrutiny.

Step 6. Move backwards up that particular branch of the hierarchy until the first level containing at least one incompleted branch is encountered. If this occurs at other than the top level of major criteria, choose the incompleted branch (any one of the incompleted branches if more than one exists), and return to Step 1 with this as a new starting point. If no incompleted branches are encountered until reaching the top level, proceed to Step 7.

Step 7. Inspect the top level of the hierarchical tree. If all major performance criteria have been completely "filled out" (i.e., if all branches starting at the top level have been completed), the process is over. A complete hierarchical structure has been constructed. However,

if one or more incomplete branches remain, choose any one of those remaining as a starting point, and return to Step 1.

This completes the procedure. A tentative criterion structure has been created and given concrete interpretation by means of the various physical performance measures attached to each of the lowest-level criteria. Subsequent procedures will test this tentative structure for worth interdependence and clarify the process of selecting physical performance measures (see Step 5).

2.2.2 A Procedure For Identifying Substantial Worth Interdependence Among Performance Criteria In The Hierarchical Structure

The preceding section outlined a procedure for generating lower-level performance criteria intended by or included within the meaning of a higher-level criterion. This procedure was presented in step-by-step form. Step 3 in the procedure is the exact point at which a higher-level criterion is to be so subdivided. The question now is, what guidelines can be provided to aid in this process of subdivision?

Perhaps the best way to answer the question is to look at the final use to which subdivided criteria will be put. After an entire hierarchical worth structure has been formulated and mapped, decision makers will investigate first the set of major performance criteria and then each set of sub-criteria. For every such set, they will determine the relative importance of each sub-criterion as a component of its related higher-level criterion. The determined relative importance of each sub-criterion will then be reflected by a numerical weight assigned thereto. Finally, these numerical weights will be used to transform intermediate point scores assigned to the sub-criteria (one score to each sub-criterion) into a single point score to be assigned to their related higher-level criterion.

Now it was pointed out in Section 2.1.6 that use of an additive weighting function with constant weights is only legitimate when applied to performance criteria judged independent of one another in the worth sense. Therefore, whatever guidelines are developed to aid in the process of subdividing higher-level criteria should certainly include a means of identifying instances of substantial worth interdependence. Two specific questions are presented below to help distinguish worth-independent sub-criteria from those displaying substantial worth interdependence.

1. In comparing a candidate sub-criterion with its related higher-level criterion, which of the following statements better describes the apparent relationship between the two?

   (a) The sub-criterion is intended by, included within the meaning of, or an integral part of the higher-level criterion.

   (b) The sub-criterion is one alternative means of satisfying the higher-level criterion and important only insofar as it contributes thereto.

2. In comparing one candidate sub-criterion with another sub-criterion already judged as appropriately included within the same set, which of the following statements better describes the apparent relationship between the two?

   (a) Willingness to accept reduced satisfaction on either sub-criterion in return for increased

satisfaction on the other would not be influenced by the degree of satisfaction already obtained on each.

(b) Willingness to accept reduced satisfaction on either sub-criterion in return for increased satisfaction on the other would depend markedly on the degree of satisfaction already obtained on each.

In order to qualify for final inclusion in the hierarchical structure, every candidate sub-criterion must receive an "a" answer to both of the above questions.

A specific, step-by-step procedure incorporating the above pair of questions appears below. It is intended that a first pass be made at creating a tentative criterion structure by means of the procedure presented in Section 2.2.1. Then, this procedure may be applied to the candidate sub-criteria generated thereby. An alternative approach would be to perform this testing procedure every time a higher-level criterion is subdivided into a set of sub-criteria (i.e., after Step 3 in the procedure presented in Section 2.2.1). Either approach would work; however, the step-by-step procedures have been written under the assumption that they will be performed sequentially rather than concurrently.

Step 1. Begin with any set of candidate sub-criteria previously generated in filling out the hierarchical structure (see Section 2.2.1, Step 3, for the exact point at which a set of candidate sub-criteria is generated).

Step 2. Arrange them in a sequence. It makes no difference how they are arranged - any arbitrary sequence will suffice.

Step 3. Compare the first sub-criterion in the sequence with the higher-

level criterion of which all of the sub-criteria are alleged to be component parts. Ask which of the following two statements better describes the relationship between the candidate sub-criterion under scrutiny and its related higher-level criterion.

A. The sub-criterion is intended by, included within the meaning of, or an integral part of the higher-level criterion.

B. The sub-criterion is one alternative means of satisfying the higher-level criterion and important only insofar as it contributes thereto.

If statement A is selected as more descriptive than statement B, move to the next sub-criterion in the sequence, and repeat the same question regarding its relationship to the higher-level criterion. Continue in this manner until the entire sequence has been exhausted; then proceed to Step 4. On the other hand, if statement B is selected as more descriptive than statement A, the sub-criterion under scrutiny does not properly belong in the set. Delete this sub-criterion from the set, lay it aside temporarily, and reconsider it later. (Note: Suggested procedures for handling deleted sub-criteria are discussed later in this thesis). Move to the next candidate sub-criterion in the sequence and repeat the same question, continuing in this manner until the entire sequence has been exhausted.

Step 4. Select another set of candidate sub-criteria as yet unchecked for worth interdependence, and return to Step 2. If all sets of sub-criteria have been checked, proceed to Step 5.

Step 5. At this point, the entire hierarchical worth structure has been tested (at least partially) for worth interdependence. Quite possibly,

some candidate sub-criteria have been deleted and set aside pending subsequent reconsideration. However, it will be useful to check the remaining structure to insure that all sub-criteria are really worth-independent. This can be accomplished by repeating Steps 1 through 4 on the entire hierarchy, but with a new question substituted for the old question in Step 3. A revised form of Step 3 is presented below to facilitate this "second pass" at testing the hierarchy.

Step 3 (revised). Compare every possible pair of sub-criteria in the sequence. (Note: If there are N sub-criteria in the sequence, there are $\frac{1}{2}N^2 - \frac{1}{2}N$ such pairwise comparisons to be made.) Ask which of the following two statements better describes each pair-wise relationship.

A. Willingness to accept reduced satisfaction on either sub-criterion in return for increased satisfaction on the other would not be influenced by the degree of satisfaction already obtained on each.

B. Willingness to accept reduced satisfaction on either sub-criterion in return for increased satisfaction on the other would depend markedly on the degree of satisfaction already obtained on each.

If statement A is selected as more descriptive than statement B, move to the next pair of sub-criteria, and repeat the same question. Continue in this manner until all of the $\frac{1}{2}N^2 - \frac{1}{2}N$ pair-wise comparisons have been made; then proceed to Step 4. On the other hand, if statement B is selected as more descriptive than statement A, at least one of the sub-criteria in the pair-wise comparison does not properly belong in the set. Move to the next pair of sub-criteria, and repeat the same question. Continue in this manner until all pair-wise comparisons have been made. Then, by

inspecting pairs which contain at least one improper member, delete and set aside those sub-criteria which do not belong in the set pending subsequent reconsideration.

This completes the identification procedure. Suggestions for ways of handling candidate sub-criteria which are identified as displaying substantial worth-~~independence~~ interdependence will be presented subsequently.

2.2.3 A Procedure For Selecting Physical Performance Measures To Interpret Lowest-Level Performance Criteria

Let us now investigate the task of selecting physical performance measures. After accomplishing sufficient conceptual refinement through successive subdivisions of higher-level criteria into sets of lower-level criteria, a single performance measure must be chosen to interpret concretely each of the lowest-level criteria in the generated hierarchical structure. In essence, our problem is to select for each lowest-level criterion some physically measurable attribute which is perceived as embodying or providing a concrete interpretation of that criterion. Thus, if one lowest-level criterion were job time, a suitable performance measure would be the time in hours required to complete a benchmark exercise. If more than one benchmark exercise were undertaken, the average benchmark time (averaged over the several benchmark exercises) would be even more appropriate. Alternatively, if it were known that a benchmark time understated the true on-site time to perform a stated job, and if the amount by which this true time were understated could itself be estimated, then an estimate of this downward bias could be added to each alternative's recorded benchmark time, and this would constitute the appropriate performance measure.

From the preceding illustrative discussion, the reader may be somewhat disturbed to see that more than one physical performance measure

may be applicable to any given lowest-level performance criterion. Furthermore, where more than one performance measure appears applicable, it may not always be obvious which one to choose. In short, judgment on the part of the decision maker must again be exercised to select an appropriate measure just as it was in generating sub-criteria.

Another factor to consider in selecting performance measures is the question of their order (i.e., their degree of generality). An example of an extremely high-order measure would be the observed time (in hours) to complete a benchmark exercise. This would reflect numerous lower-order performance measures such as raw processing times (e.g., add time, multiply time, cycle time, etc.), and other more specific attributes of computer performance.

An example of a moderately high-order performance measure would be overall storage capacity provided either by core, by disk, by drum, or by tape. This high-order measure might be computed by adding the contributions (in bits or words) made by each of the above sources.

Contrast each of these two examples of high-order measures with the case of raw add time (in microseconds). Raw add time is a very basic, low-order measure. It is not easily decomposable into more elementary component measures. More importantly, it is not clear that raw add time would ever be a useful measure in and of itself.

Now the order of a physical performance measure is important for two reasons. First, as illustrated above, high-order measures are generally much more relevant to attempts at assessment than are low-order measures. Consequently, an effort should be made to select and/or concoct high-order measures whenever possible. Second, as illustrated by the overall storage capacity measure, creation of high-order measures out of lower-order measures can sometimes be used as a means of retrieving

sub-criteria which have been temporarily deleted from the criterion structure and set aside due to worth ~~interpendence~~ INTERdependence. Such deleted sub-criteria can be replaced by a single higher-level criterion, and a single high-order performance measure can be selected to go with it.

In summary, what guidelines can now be provided for the selection of appropriate physical performance measures? Five guidelines are suggested.

1. Consult the master list to obtain a set of candidate measures.

2. Augment this set by inventing any additional measures not contained in the master list, but which seem appropriate in the context of the lowest-level criterion under consideration and the stated job.

3. Check candidate measures for practical feasibility (i.e., to insure that all data included in the measure can be conveniently and promptly gathered).

4. Attempt to combine candidate measures into higher-order measures, where possible and appropriate.

5. On an intuitive basis, select the seemingly most appropriate and highest-order of the practically feasible candidate measures.

A specific step-by-step procedure incorporating the above five guidelines appears below. It is intended that this procedure be implemented concurrently with the generating procedure outlined in Section 2.2.1.

Step 1. Begin with any one of the lowest-level performance criteria occurring at the base of the previously generated hierarchical structure.

Step 2. Consult the master list of performance criteria and physical performance measures. Looking only at the physical measures contained in the master list, identify those which are perceived as significantly related to the lowest-level criterion under consideration. This may be done by asking the following question about the relationship perceived to exist between the criterion and every physical performance measure on the master list.

> Would changes in the state or numerical value of the performance measure be capable of bringing about either significant increases or significant decreases in the extent to which the lowest-level criterion under consideration is satisfied?

If the answer to the above question is yes, then a significant relationship is said to exist between the lowest-level criterion and the physical performance measure. If the answer is no, then no such relationship is perceived.

Step 3. Add to the set of physical measures drawn from the master list and perceived as significantly related to the lowest-level criterion any additional measures which can be thought of and which also seem related. In this manner, decision makers can supplement the master list with their own imagination and experience.

Step 4. Looking now at all candidate measures generated by Steps 2 and 3, check to see whether each is practically feasible. That is, insure that all data necessary to form each measure can be conveniently and promptly gathered. Delete any candidate measures which are discovered to be practically infeasible.

Step 5. Inspect the residue of feasible candidate measures remaining after Step 4. Either choose one straightaway (by intuitive judgment) as the most appropriate single measure by which to interpret the lowest-

level criterion or, if none of the feasible candidates seem really appropriate, attempt to construct a higher-order physical measure out of two or more individual measures.

Step 6. Proceed to another lowest-level criterion in the hierarchical structure, and repeat Steps 2 through 5. Continue in this manner until all lowest-level criteria have been assigned a corresponding physical performance measure.

This completes the procedure.

2.2.4 A Procedure For Attaching Numerical Weights To Hierarchically Arranged, Worth-Independent Performance Criteria

At this point, it is assumed that a complete hierarchy of worth-independent performance criteria has been constructed. It is also assumed that every lowest-level criterion in this hierarchy has been given a concrete interpretation by assigning to it a physical performance measure. It now remains to attach weights to all criteria in the hierarchy and to specify (by defining scoring functions) the precise worth relationships perceived as linking each lowest-level criterion to its assigned performance measure. The first of these two tasks will be performed in this section. The second will be performed in the next section.

The weight-setting procedure to be developed herein is divided into two sequential phases. In the first phase, an individual decision maker attempts to produce his own numerical weights corresponding to each of the sub-criteria contained in some specified set of sub-criteria appearing in the hierarchical structure. In the second phase, individual weight sets assigned by separate decision makers are compared, and lack of consensus among decision makers (if there are more than one) is resolved by an averaging technique.

The first phase of the procedure involves two major operations.

1. All sub-criteria subsumed under a given higher-level criterion are ranked in order of ascending perceived importance.

2. Then, starting with the most important pair of sub-criteria appearing at the head of the list, successive pair-wise comparisons are made between contiguous sub-criteria, and decision makers are asked to indicate in terms of a ratio the degree of perceived relative importance of the two. Stated alternatively, decision makers are asked to indicate the rate at which they would be willing to accept reduced satisfaction of one sub-criterion in return for increased satisfaction of the other.

A step-by-step procedure to implement this first phase follows immediately. The resulting individual weights generated by this procedure are all positive, they sum to one, and they are interpretable in accordance with the weighting conventions stipulated in Section 2.1.7. However, one word of warning seems appropriate. Although this procedure (if it can be carried out at all) guarantees that the resultant weights will possess certain desirable logical properties (i.e., consistency, transitivity, and preservation of the preselected importance ratios), the validity of the weights themselves still remains the responsibility of informed judgment on the part of decision makers. Neither this procedure nor any other procedure based solely on logical considerations can guarantee their validity. Only clearly articulated judgment can ever provide that.

Step 1. Begin with any set of sub-criteria subsumed under a higher-level performance criterion.

Step 2. List these sub-criteria in approximate order of relative importance, starting with the most important sub-criterion at the top of the list and the least important sub-criterion at the bottom. It is not necessary to have the sub-criteria perfectly ranked or ordered on this first pass, since subsequent operations will be performed to guarantee complete ordering.

Step 3. Compare the first two sub-criteria on the list.

a. If the first sub-criterion is deemed relatively more important than the second, proceed directly to Step 4.

b. If both sub-criteria are deemed roughly equal in importance, proceed directly to Step 4.

c. If the second sub-criterion is deemed relatively more important than the first, invert their positions on the list (i.e., place the first sub-criterion where the second used to be on the list, and vice-versa), and then proceed to Step 4.

Step 4. Compare the lower-ranked sub-criterion from Step 3 with the next sub-criterion on the list. Repeat the comparisons and stipulated operations in Step 3 on this new pair of sub-criteria. Continue in this manner all the way down to the end of the list until pair-wise comparisons have been made between all contiguous criteria.

Step 5. After the list has been completely exhausted, go back and determine whether any inversions (position changes) occurred.

a. If none occurred, proceed directly to Step 6.

b. If one or more occurred, return to the head of the list, and repeat the entire procedure described in Steps 3 and 4.

Step 6. Eventually, the list will become so arranged that successive pair-wise comparisons will generate no inversions. It may require several passes to achieve this result, but it will occur in the end (assuming that the decision maker's notions of relative importance among sub-criteria are both consistent and transitive). When the list has achieved an arrangement wherein no inversions occur, it will then reflect the decision maker's judgments of relative importance in terms of direction, but not yet in terms of magnitude. Relative magnitudes are determined by subsequent steps.

Step 7. Take the first sub-criterion on the rearranged list, and assign to it the number 1.0 or one hundred percent.

Step 8. Compare the second sub-criterion with the first, and assess their relative importance in terms of a ratio or fraction. That is, if satisfying the second sub-criterion seems only one-half as important as satisfying the first, assign the fraction $\frac{1}{2}$ or its decimal equivalent .5 to the second sub-criterion. In like manner, fractions such as $\frac{3}{4}$, $\frac{9}{10}$, etc. or their decimal equivalents might equally well have been assigned. (Note: It may be difficult to set weights when the question is phrased in the above manner. An alternative form of the same question would be, "At what rate would reduced satisfaction of the first sub-criterion be acceptable in return for increased satisfaction of the second so as to maintain the same overall worth considering satisfaction of both sub-criteria jointly?" The answer to this question, expressed in the form of a ratio, may then be assigned as before to the second sub-criterion.)

Step 9. Compare the second and third sub-criteria, assess their relative importance or trade-off rate in terms of either a fraction or a ratio, multiply the number assigned to the second sub-criterion by this fraction or ratio, and assign the resultant product to the third sub-criterion. For example, assuming that the second sub-criterion were assessed as being $\frac{1}{2}$ as important as the first, while the third were assessed as being $\frac{9}{10}$ as important as the second, the appropriate computation would be $\frac{1}{2} \times \frac{9}{10}$, and the number $\frac{9}{20}$ would be assigned to the third sub-criterion.

Step 10. Repeat the above procedure for all successive pair-wise comparisons until the list of sub-criteria has been completely exhausted. Then, each sub-criterion will have been assigned a number equal to the product of its importance relative to the next higher sub-criterion times the number previously assigned to the next higher sub-criterion.

Step 11. Add the numbers assigned to all sub-criteria on the list, and then divide each one by the computed sum. This will serve to convert relative importance ratios into normalized weights. Each weight will be positive, and the whole set will add to one. In addition, the relative importance ratios will be preserved in the ratios of any pair of weights.

This completes the procedure.

Now there may not always be complete agreement among separate decision makers concerning the proper collection of weights to be attached to any set of sub-criteria. In fact, numerical differences, and perhaps even rank-order differences, are to be expected among separate decision makers - particularly if they set weights without first consulting one another. This lack of consensus would seem quite healthy, in the writer's opinion, and should be encouraged rather than discouraged. Unless any single decision maker is willing to claim that his weights are precisely

correct and, therefore, that anybody who disagrees with him is necessarily wrong, then some method for combining group opinion would seem appropriate.

One way of combining group opinion would be to subject differences of opinion to open discussion in hopes of achieving greater consensus. This would be a particularly effective remedy for those situations where some decision makers possess greater knowledge and experience than others. By open discussion, the less knowledgable and less experienced decision makers could benefit from their better endowed compatriots and thereby gain a sounder basis for assessment.

However, open discussion would <u>not</u> be effective against genuine differences of opinion held by equally knowledgable and equally experienced decision makers. <u>Nor</u> would it be effective against whatever differences remain after open discussion has enlightened those decision makers who did not possess initially the same knowledge and experience as others, but who altered their opinions somewhat in the face of ensuing discussions. Some sort of compromise procedure would seem appropriate in these two instances.

One way of achieving a compromise would be by averaging individual weights across separate decision makers. That is, to each sub-criterion in a particular set, separate decision makers would assign their own individual weights. Then, an average weight would be computed for each sub-criterion by adding the weights assigned by separate decision makers and dividing the total by the number of decision makers. It can be shown that, if this averaging procedure is applied to each sub-criterion in a set, then the computed average weights assigned to each of the sub-criteria will sum to one. In addition, the resultant average weights would reflect group opinion instead of one single individual's opinion.

In actual practice, both of the above procedures would seem appropriate if carried out in sequence. First, a group of decision makers would meet to discuss the relative importance of sub-criteria in some designated set. By open discussion, all decision makers would be accorded a similar basis for formulating their own individual opinions. Then, each decision maker would reflect his individual opinion in a set of numerical weights (generated via the step-by-step procedure just presented). Finally, remaining differences of opinion would be handled by averaging over individual weight assignments to arrive at a final set of group weights for each of the sub-criteria. In this manner, spurious differences of opinion arising from differences in knowledge and experience would be minimized, while genuine differences of opinion arising from genuinely different views of the situation would be adequately reflected in the final average weights.

A step-by-step procedure to implement the second phase of the weight-setting process, designed to average out remaining differences of opinion, is presented below.

Step 1. Collect whatever individual weights have been assigned by separate decision makers to a set of sub-criteria in the hierarchical structure.

Step 2. Suppose that there are N separate decision makers and M sub-criteria in the set to which individual weights have been attached. (Note: Both N and M are assumed to be greater than one. If N = 1, there would be no problem of lack of consensus. If M = 1, there would be only one sub-criterion in the set, and it would therefore have to receive a full weight of 1.0.)

Step 3. Lay out the individual weights assigned by separate decision

makers in N parallel columns of M weights each. The resulting rectangular array may be thought of as a matrix with M rows and N columns.

Step 4. Compute and record the sum of the weights appearing in each of the M rows of the above matrix. (Note: If it is considered desirable to weight some opinions more heavily than others, compute an appropriately weighted sum.)

Step 5. Divide each computed row sum by N. This gives an average weight, averaged across the N separate decision makers, for each of the M subcriteria. (Note: The M average weights must add to one - except, perhaps, for small rounding errors. If they do not add to one, check the computations for algebraic errors.)

This completes the procedure.

2.2.5 A Procedure For Establishing Independent Scoring Functions To Link Lowest-Level Performance Criteria To Their Assigned Physical Performance Measures

The next task is to formulate scoring functions by which each lowest-level performance criterion may be linked to its assigned measure of physical performance. Once formulated, these scoring functions may be used to convert measured physical performance into equivalent worth scores, and these worth scores may then be combined via the pre-established weights into a single numerical index indicating the overall worth of a proposed alternative.

The scoring procedure itself will be broken down into two major phases. The first phase will contain an ordered sequence of questions designed to determine the general nature and shape of whatever scoring function is to be formulated. The nature and shape of each scoring function will be inferred from answers to the following questions:

1. Is the physical performance measure to be scored discrete or continuous?

2. If discrete, how many measurement categories are contained in the physical performance scale; is there any inherrent order or sequence built into this scale; and are there any qualitative distinctions to be made concerning observations within each measurement category?

3. If continuous, is the physical performance scale bounded from above and/or from below?

4. If bounded, where do the boundaries of the physical performance scale fall?

5. With which points on the physical performance scale are zero and one hundred percent satisfaction of the related lowest-level performance criterion associated, respectively?

6. Does satisfaction increase or decrease with increases in measured performance?

7. Does the rate at which satisfaction increases or decreases with increases in measured performance ever change, or does it remain constant over the entire range of the physical performance scale?

8. If the above rate changes, does it always increase, or does it always decrease, or does it both increase and decrease over selected intervals within the range of the physical performance scale?

The second phase will contain a step-by-step procedure designed to select a specific scoring function of the general nature and shape indicated in the first phase. Actually, two alternative procedures will be presented to implement this second phase - one involving visual and graphic methods, the other involving numerical methods. The choice between these two alternative procedures will be left up to the discretion of decision makers.

2.2.5.1 Phase I Of The Scoring Procedure: Determining The General Nature And Shape Of A Scoring Function

The general nature and shape of a scoring function will be determined by answers to an ordered sequence of questions. In all, there are twelve possible questions which might be raised, but not every one of the twelve will be relevant to formulating any given scoring function. Consequently, all twelve questions will first be stated and clarified by example. Then, a step-by-step questioning procedure will be presented to indicate which of the twelve questions are relevant to any given scoring function.

Of the twelve questions, six refer to the scale of the physical performance measure itself for which a scoring function is to be formulated. These first six questions will appear, along with illustrative examples, in Section 2.2.5.1.1. The remaining six questions refer to relationships presumed to exist between the worth scale (calibrated in points ranging from zero to one) and the scale of the physical performance measure (calibrated in whatever physical units are appropriate). These six questions will appear, along with illustrative examples, in Section 2.2.5.1.2. Finally, Section 2.2.5.1.3 will present a step-by-step questioning procedure designed to carry decision makers in an orderly fashion to a final conclusion concerning the general nature and

shape of the scoring function to be formulated.

2.2.5.1.1 Questions About The Scale Of The Physical Performance Measure

Question 1. Consider the scale of the physical performance measure. Is it continuous or is it discrete? If all conceivable numbers within the range of the physical performance scale are possible measurements, then the scale is continuous. An example of a continuous performance scale would be "benchmark time." This scale is bounded from below by zero (i.e., it is logically impossible to record a negative benchmark time), but it does not possess any definite upper bound. Therefore, the logical range of this scale falls between zero and positive infinity. Any conceivable number of hours and portions thereof within this range constitutes a possible benchmark observation. For this reason, "benchmark time" possesses a continuous or everywhere-dense scale.

On the other hand, if the physical performance measure can assume only certain specified values within its logical range, then the scale of that measure is discrete. An example of a performance measure with a discrete scale would be "number of on-call maintenance personnel provided by a computer manufacturer." This measure can only assume positive integral values or zero. Fractional values cannot occur, since human beings exist only in integral numbers. (Note: One might argue that a maintenance man can spend a portion of his time being on-call, but then the performance measure would be "number of on-call maintenance man-hours provided by a computer manufacturer." This measure possesses a continuous scale, but it is a different measure.)

Question 2. For a physical performance measure possessing a discrete scale, how many discrete categories or levels are included within its

range? There must be at least two such categories; otherwise, the performance measure could never discriminate among alternatives (i.e., every alternative would fall in the same category, if only one category were defined, and in no category, if no categories were defined). However, it is possible for a discrete scale to possess two, three, or any higher integral number of categories or levels.

An example of a performance measure possessing a two-level discrete scale would be "provision of some specified analytical software routine" such as a canned regression program. Such a routine is either present or absent from any alternative.

An example of a performance measure possessing a five-level discrete scale would be "type of memory device" whose scale categories might include:

1. core
2. disk
3. drum
4. some combination of the above
5. other

Our previous example of "number of on-call maintenance personnel provided by a computer manufacturer" illustrates a many-level discrete scale. Here, any positive integer or zero constitutes a possible measurement value.

Question 3. Does the physical performance measure possess a scale which is neither purely discrete nor purely continuous, but is some hybrid or combination of the two? In other words, does the performance scale display discrete and continuous properties simultaneously? An example of such a hybrid measure would be "expected number of on-call maintenance personnel provided by a computer manufacturer." Here, a manufacturer

might promise to provide no on-call maintenance men, or he might promise to provide at least one, but with the stipulation that such personnel provided would not always answer a call. If the manufacturer promises to provide at least one maintenance man on this conditional basis, and if the relative frequency with which each conditional man provided will in fact answer a call can be estimated, then the expected number of on-call maintenance personnel provided by that manufacturer may be computed by multiplying his conditional number provided by their estimated relative frequency of answering a call. The discrete aspects of this scale derive from the fact that a positive integral number or zero maintenance men may be promised, but perhaps only on a conditional basis. The continuous aspects of the scale derive from the fact that, where uncertainty exists concerning whether or not a call will be answered, this uncertainty is reflected by concocting a frequency-weighted average or expected value as a measure.

Question 4. For a physical performance measure possessing a discrete scale, do the various categories or levels contained within the range of that scale fall into some natural order or sequence, or are they strictly nominal in character? Our previous example of the "number of on-call maintenance personnel provided by a computer manufacturer" also illustrates this concept of natural ordering. Since zero men is less than one man is less than two men, and so forth, the discrete levels contained within this scale do fall into a natural sequence.

In contrast, our previous example of "type of memory device" does not possess a scale whose categories fall into any natural order. These categories are strictly nominal in character. The only thing we can say about core, disk, drum, some combination, and other is that they are different categories. There is no meaningful physical sense in which core

is either less than or greater than disk. They are just different. The same conclusion may be drawn from all other pair-wise comparisons between scale categories associated with this performance measure.

Based on these contrasting examples, a more formal definition may be given for the conceptual distinction between strictly nominal discrete scales and naturally ordered discrete scales. If, by inspecting the discrete categories contained within the range of the scale, "greater than" and "less than" relationships are naturally defined in the physical sense, then these categories fall into a natural order, and the scale is said to be an ordered scale. If, on the other hand, inspection of the component scale categories permits only "same as" and "different from" distinctions, then these categories are strictly nominal in character, and the scale is said to be a nominal scale.

One word of caution seems appropriate before leaving the distinction between nominal and ordered scales. This distinction does not refer to the worth imputed by decision makers to various scale categories. Thus, core storage may be preferred to disk storage as a memory device due to the former's faster access time. However, the ordering here is a preference ordering imputed by decision makers, and it derives not from the physical identity of the two kinds of memory devices per se, but rather from an associated characteristic - access time. The only physical thing we can say about the sheer identity of core storage devices versus disk storage devices is that they are different. If we were concerned with memory devices because of their access time, then we should define "access time" directly as our performance measure, and we should ignore the nominal identity of alternative types of memory devices which provide this access time. Under these circumstances, "access time" would constitute a performance measure whose scale values fall into a natural order, but this is not the same measure as "type of memory device." On

the other hand, where a performance measure with a discrete, but only nominal scale is appropriate (e.g., "presence or absence of some specified analytical software routine"), the existence of a preference ordering imputed by decision makers (presence is preferred to absence) be clearly distinguished from the existence of a natural ordering inherent in the physical measure itself (presence is just different from absence in the sheer physical sense).

Question 5. For a physical performance measure possessing a continuous scale, can either a logical lower bound or a logical upper bound or both be identified with that scale? Let us first consider the issue of a logical lower bound. Most continuous physical performance scales are defined in such a way as to exclude the possibility of negative measurements. It is impossible, for example, to conceive of performing a benchmark exercise in negative time or of acquiring a computer with one or more negative processing speeds. This means that most continuous performance scales do possess definite lower bounds and that such bounds fall either at zero or at some positive number. In light of this, and because negative performance measures introduce certain conceptual and computational complications, it will be desirable to insure that no performance scales can ever include negative measurement numbers. A procedure for accomplishing this objective will be presented later.

Having established by fiat that all continuous performance scales will be restricted in range to exclude negative measurement numbers, we have, in effect, established the existence of a lower bound. All such scales must have lower bounds falling either at zero or at some positive number. The next question is to determine whether zero or some positive number constitutes that lower bound.

In the vast majority of cases, logical lower bounds will fall

naturally at zero. "Benchmark time," "slack time between benchmark time and some mandatory maximum daily, weekly, or monthly operating time," and various types of "processing speeds" all possess continuous performance scales with logical lower bounds falling exactly at zero. It is logically impossible to receive proposal data containing negative observations on any of these scales, but it definitely is possible to receive zero and various positive observations on all of them.

In some cases, however, logical lower bounds will occur at strictly positive numbers. Nevertheless, even in these (rare) instances, it will be convenient to re-scale the performance measure such that its logical lower bound falls exactly at zero. Re-scaling procedures will be presented later to accomplish this end.

Question 6. For a physical performance measure possessing a continuous scale, can a logical upper bound be identified with that scale, and, if so, at what measurement number does this upper bound fall?

An example of a physical performance measure possessing a continuous scale lacking any logical upper bound is "benchmark time." It is conceivable that any positive number of minutes or hours could be recorded from a benchmark exercise. This means, in effect, that the scale of "benchmark time" extends from zero to positive infinity.

Now what happens if some mandatory upper limit on "benchmark time" has been stipulated? Suppose, for example, that daily work-loads must be processed within a single eight-hour shift in order that an alternative may be considered feasible. Does eight hours constitute a logical upper bound to "benchmark time"? No. It does not. Mandatory performance requirements are set by decision makers to reflect maximum or minimum levels of acceptable performance. They do not necessarily reflect the bounds of logically possible performance.

Alternately, it might be argued that benchmark times will very likely fall in a range from one or two hours to ten or twelve hours. However, twelve hours would not constitute a logical upper bound in this case either. It is conceivable, although unlikely, that some observed benchmark time would exceed twelve hours. In fact, no matter where we attempt to set an upper bound, so long as we set it at some finite positive number, it is still conceivable that at least one observed benchmark time might exceed it. The higher we set the upper bound, the more confident we can be that all observed benchmark times will fall below it; but we can never be absolutely certain. It is for this reason that the scale of "benchmark time" must be regarded as unbounded. Any positive number, no matter how unlikely, is still logically possible.

2.2.5.1.2 Questions About Relationships Presumed To Exist Between The Worth Scale And The Scale Of The Physical Performance Measure

Question 1. With which levels on the scale of the physical performance measure are zero worth points and one worth point to be associated, respectively? Stated alternatively, what level of physical performance is to be regarded as completely unsatisfactory, and what level is to be regarded as completely satisfactory?

This question is easy to answer in the case of performance measures with discrete, two-level scales. One level must be completely unsatisfactory, and the other must be completely satisfactory. Thus, absence of an analytical software routine would be completely unsatisfactory (receiving zero worth points), while its presence would be completely satisfactory (receiving one worth point).

In the case of performance measures with discrete, many-level scales, it depends upon whether the discrete scales are strictly nominal or ordered. If strictly nominal, then the question generally cannot be

answered at all. A procedure to handle this type of situation will be presented later. However, if the discrete, many-level scale is ordered, then the setting of zero and one hundred percent satisfaction levels (i.e., zero and one worth point) depends upon the range (i.e., placement of logical lower and upper bounds) of the physical scale. The same is true of performance measures with continuous scales.

It has been established by fiat that all continuous performance scales will possess logical lower bounds falling exactly at zero. In most cases, this occurs naturally, but where exceptions arise, remedial procedures will be employed to make this occur. Similarly, discrete, ordered, many-level performance scales will be made to have logical lower bounds falling exactly at zero - if this does not occur naturally. The only remaining question, then, is whether or not such scales have logical upper bounds. Three examples will be drawn to illustrate the fitting of zero and one hundred percent satisfaction levels to these types of performance scales. Two examples will involve scales without logical upper bounds, and one will involve a scale with a definite logical upper bound.

Consider, first, the case of "benchmark time." The scale of this performance measure is bounded from below by zero, but is unbounded from above. Clearly, an observed benchmark time of zero (minutes, hours, days, etc.) would be one hundred percent satisfactory. Equally clearly, an infinitely large observed benchmark time would yield zero satisfaction. Intermediate levels of benchmark time would yield intermediate degrees of satisfaction. Consequently, we could associate a worth point score of one with zero benchmark time and a point score of zero with infinite benchmark time. But is this not a rather extreme - even peculiar - way to set zero and one hundred percent satisfaction levels? No <u>real</u> benchmark observations will fall exactly at zero time or anywhere near infinite time. In

addition, we would be "almost" one hundred percent satisfied with benchmark times substantially above zero, and we would regard as "almost" completely unsatisfactory large benchmark times falling far below positive infinity. So why set such extreme limits? The answer to this question is that we should set as extreme limits as possible both to avoid the requirement of predicting the range of actual benchmark observations (which predictions are subject to error) and to recognize the fact that variations in benchmark time at both extremes of the time scale do yield variations in satisfaction - even though very slight. The scoring procedures to be presented in Appendices at the end of this thesis will permit making variations in worth point scores as small as desired at the extremes of the performance scale; but we might as well construct scoring functions which are flexible enough to encompass all logically possible situations. This costs next to nothing in effort and may very well provide substantial savings in terms of future headaches.

Consider, next, the case of "number of on-call maintenance personnel provided by a computer manufacturer". This performance measure possesses a discrete, ordered, many-level scale bounded from below by zero, but unbounded from above. Zero satisfaction would be derived from zero maintenance personnel, and one hundred percent satisfaction would be derived from an infinite number of such personnel. Once again, we would set these end-points of the worth scale at the most extreme possible levels of the performance scale to insure logical completeness. Zero worth points would be assigned to zero maintenance personnel, and one worth point to infinite personnel.

As a third example, consider the case of "slack time between benchmark time and some mandatory maximum daily, weekly, or monthly operating time." The scale of this performance measure is continuous and

doubly bounded between zero and the mandatory maximum time. Zero worth points would be assigned to zero slack time, and one worth point would be assigned to the mandatory maximum level of slack time.

In summary, fitting the end-points of the worth scale to a physical performance scale is usually a simple job of matching. Either zero worth points or one worth point is assigned to zero performance (depending upon the performance scale), and the other end of the worth scale is assigned either to the logical upper bound of the performance scale (if such an upper bound exists) or to infinite performance.

Question 2. For physical performance measures possessing either continuous scales or discrete, ordered, many-level scales, what is the direction of the preference relationship? Does a higher level of physical performance yield greater satisfaction, or does it yield less satisfaction? Additionally, is the direction of the preference relationship uniform over the entire logical range of the physical performance scale?

The first part of this question is very easy to answer. A higher observation on "benchmark time" is obviously less satisfactory and, therefore, deserving of a lower worth score than is a lower observation on "benchmark time." This same type of reverse preference relationship characterizes most physical performance measures expressed in terms of time required to do something. It is generally preferable to have things accomplished in a shorter rather than in a longer time.

In contrast, a direct preference relationship exists in the case of "slack time," "number of on-call maintenance personnel provided by a computer manufacturer," "efficiency of compilers," "number of machine reliability checks," and many other physical performance measures. In all of these cases, more physical performance is considered better than less performance. This type of direct preference relationship character-

izes most capacity measures as well as many other operating measures.

The second part of the question is also easy to answer in most cases. Almost all physical performance measures have associated preference relationships which remain uniform in direction over the entire range of their physical scales. Thus, if a direct preference relationship exists in one region of a physical performance scale, it will generally hold in all other regions - although, perhaps, to a greater or lesser extent. If one reliability check is preferred to zero reliability checks, then one hundred such checks will be preferred to ninety-nine, and so on up the line. Similarly, reverse preference relationships usually hold uniformly over the entire range of performance scales. If a benchmark time of four hours is preferred to a time of five hours, then four minutes would be preferred to five minutes, and four days would be preferred to five days.

Occasionally, a physical performance measure may appear with a non-uniform preference relationship. That is, the preference relationship may change direction at some point or points within the range of the physical performance scale. For a while, more performance may be preferred to less; but after some point, less performance may be preferred to more. An example of this sort of situation occurs when one scratches an itching portion of his skin. For a while, continued scratching is preferable to discontinued scratching; but if the scratching process is continued too long or too intensively, it is preferable to scratch more lightly or to discontinue the process altogether. Special provisions will have to be taken in such instances, since the scoring procedures to be presented herein are not intended to cover them.

Question 3. For physical performance measures with uniformly directed preference relationships, how does the rate of increase (or decrease) of worth behave with increases in physical performance? Does the rate of change of worth remain constant over the entire range of the performance scale, or does this rate vary over different regions of the scale?

To clarify this question, let us define more explicitly what is intended by the notion of a rate of change of worth. First, we know that worth should change over various regions of a physical performance scale; otherwise, there would be no point in scoring the physical performance measure. Every possible level of performance would receive the same worth score under these circumstances, and such a measure would fail to discriminate among competing alternatives. Second, given that worth does change with performance, it is important to know the direction of change (i.e., the direction of the preference relationship discussed in Question 2). But knowing only the direction of change is not sufficient. We must also know the rate of change. Does worth change rapidly with changes in performance, or does it change only slowly. Finally, we must also know the acceleration of change. That is, does worth change with performance at a constant rate or at a variable rate (i.e., either at an accelerating rate or at a decelerating rate)?

The question posed here is about the acceleration of changes in worth with increases in performance. If worth changes at a constant rate with increases in performance, then there is no acceleration or deceleration inherent in the preference relationship. A linear or straight line relationship with a constant slope (indicating a constant rate of change) would be selected to depict this sort of situation. Equal changes in performance always receive the same increase or decrease in worth score throughout the entire range of the performance scale.

However, if worth changes at a variable rate with increases in performance, there is some degree of acceleration or deceleration inherent in the preference relationship, and a non-linear scoring function would be selected to depict this type of relationship.

An example of a performance measure whose preference relationship displays a constant rate of change would be "slack time between benchmark time and some mandatory maximum daily, weekly, or monthly operating time". Assuming that slack time would be used to increase the workload by merely processing additional records (e.g., additional personnel records), and assuming that the worth of processing each additional record were the same for all records, then a simple, straight line, direct relationship would exist between slack time and the worth of slack time. Under the above assumptions, the rate of increase of worth with increases in slack time would be constant, since each additional unit of slack time would be utilized to process additional records of equal worth.

An example of a performance measure whose preference relationship displays accelerating changes would be "frequency of machine breakdowns." Here, the direction of the preference relationship would be reverse (i.e., lower frequencies are preferred to higher frequencies), and the rate of this reverse preference increases as the frequency of breakdowns increases. Stated alternatively, the rate of decrease of worth with increasingly frequent breakdowns itself increases. Each breakdown generally requires backing up operations to some point in time prior to the actual moment of breakdown, which often ruins already accomplished work. This is particularly true of multi-processing systems.

An example of a performance measure whose preference relationship displays decelerating changes would be "number of on-call maintenance personnel provided by a computer manufacturer." The direction of

this preference relationship would be direct (i.e., the more the better), but the rate of this direct preference decreases as more and more maintenance personnel are provided. The additional worth of each additional man decreases steadily, since there is just so much maintenance work to be done, and, after a point, providing additional personnel does not serve to get the work done either much faster or much better.

Question 4. In the case of either accelerating or decelerating rates of change of worth with increasing performance, does the preference relationship always accelerate or always decelerate, or does it do first one and then the other over different regions of the performance scale?

Our previous example of "number of on-call maintenance personnel provided by a computer manufacturer" illustrates an always decelerating preference relationship. Each additional man is worth somewhat less than his predecessor, and this is true throughout the entire logical range of the performance scale.

On the other hand, "frequency of machine breakdowns" illustrates a preference relationship which, although uniformly reverse, first accelerates and then decelerates. As discussed in Question 3, increases in the frequency of breakdowns serves to reduce the worth of the machine at an accelerating rate. But this does not continue forever. After a while, when the frequency of breakdowns gets high enough, the worth of the machine has already been reduced to such a low level that an even higher breakdown frequency will result in only a little additional deterioration. The situation is already so bad that it cannot get much worse - even if the machine were to shut down completely (an infinite breakdown frequency). This sort of situation would be depicted by a scoring function in the shape of a mirror-image "S".

Question 5. Does a preference relationship with uniformly changing rate always accelerate or always decelerate? The meaning of this question has already been illustrated in discussing Question 4.

Question 6. Does a preference relationship with non-uniformly changing rate first accelerate and then decelerate, or does it first decelerate and then accelerate? The meaning of this question has also been illustrated in discussing Question 4.

2.2.5.1.3 A Step-by-Step Questioning Procedure

In this section, the twelve questions presented in the two previous sections will be arranged in an ordered sequence. The purpose of providing such a step-by-step questioning procedure is to direct decision makers systematically toward a scoring function of particular nature and shape.

Step 1. Consider the scale of the physical performance measure. Is it continuous or is it discrete? If discrete, proceed to Step 2. If continuous, proceed to Step 7.

Step 2. Is the discrete scale purely discrete or is it a hybrid, containing continuous aspects as well as discrete aspects? If purely discrete, proceed to Step 3. If hybrid, treat it as if it were continuous, and proceed to Step 7.

Step 3. How many categories or levels are contained within the discrete scale identified in Step 2? If two, proceed to Step 4. If three, four, or five, proceed to Step 5. If more than five, proceed to Step 6.

Step 4. If the discrete, two-level scale identified in Step 3 is merely a case of presence or absence of some desirable attribute, proceed to

scoring procedure 1 in Appendix I. If presence of the desirable attribute is to be qualified by an additional measure of relative worth, proceed to scoring procedure 2 in Appendix II.

Step 5. Is the discrete scale identified in Step 3 strictly nominal or is it ordered? If strictly nominal, proceed to scoring procedure 3 in Appendix III. If ordered, proceed to scoring procedure 4 in Appendix IV.

Step 6. Is the discrete scale identified in Step 3 strictly nominal or is it ordered? If strictly nominal, proceed to scoring procedure 5 in Appendix V. If ordered, treat the scale as if it were continuous, and proceed to Step 7.

Step 7. Does the continuous scale identified in Step 1, Step 2, or Step 6 possess a logical lower bound? If yes, proceed to Step 10. If no, proceed to Step 8.

Step 8. It is very unlikely that a performance measure will have been selected whose scale is unbounded from below (i.e., where negative observations are possible and may range all the way to negative infinity). Therefore, ask once again whether the scale under scrutiny possesses a logical lower bound. If the answer is now yes, proceed to Step 10. If the answer is still no, look for a logical upper bound. If the scale possesses no logical upper bound either, the performance measure must be rejected. The scoring procedures presented herein are not equipped to handle doubly unbounded performance scales. Choose a new performance measure, and return to Step 1. However, if the scale does possess a logical upper bound, proceed to Step 9.

Step 9. Transform the scale identified in Step 8 by multiplying every number contained therein by minus one. This transformed scale will now

possess a logical lower bound, but no logical upper bound. Proceed to Step 10, but keep in mind that the new transformed scale is just the reverse of the original scale. Consequently all subsequent questions about the transformed scale must be answered with this reversed aspect in mind.

Step 10. Does the logical lower bound fall exactly at zero? If yes, proceed to Step 12. If no, proceed to Step 11.

Step 11. Identify the numerical value of the logical lower bound. Transform the scale by subtracting this number from every number contained in the scale. Keep this transformation in mind, and remember that all subsequent questions will refer to the transformed scale. Proceed to Step 12.

Step 12. Does the scale possess a logical upper bound? If yes, proceed to Step 13. If no, proceed to Step 24.

Step 13. It has been determined that the performance scale is bounded from below by zero and from above by some finite positive number. What is the direction of the preference relationship? If direct, proceed to Step 14. If reverse, proceed to Step 19.

Step 14. Now fit the end-points of the worth scale to the logical lower and upper bounds of the performance scale. Assign zero worth points to zero performance and one worth point to the logical upper bound of the performance scale. Proceed to Step 15.

Step 15. Is the direct preference relationship identified in Step 13 uniform over the entire logical range of the performance scale? If yes, proceed to Step 16. If no, the scoring procedures presented herein are inadequate to handle such a performance measure. Define a new performance measure, and return to Step 1.

Step 16. Does the direct preference relationship identified in Step 15 maintain a constant rate of change of worth, or does it display a variable rate of change (i.e., either accelerating, decelerating, or both in sequence)? If constant, proceed to scoring procedure 6 in Appendix VI. If variable, proceed to Step 17.

Step 17. Does the variable rate of change of worth identified in Step 16 display uniform acceleration, uniform deceleration, or first one and then the other? If uniform acceleration, proceed to scoring procedure 7 in Appendix VII. If uniform deceleration, proceed to scoring procedure 8 in Appendix VIII. If first one and then the other, proceed to Step 18.

Step 18. Does the variable rate of change identified in Step 17 start by accelerating and then end by decelerating, or does it start by decelerating and then end by accelerating? If the former, proceed to scoring procedure 9 in Appendix IX. If the latter, proceed to scoring procedure 10 in Appendix X.

Step 19. Now fit the end-points of the worth scale to the logical lower and upper bounds of the performance scale. Assign zero worth points to the logical upper bound of the performance scale, and assign one worth point to zero performance. Proceed to Step 20.

Step 20. Is the reverse preference relationship identified in Step 13 uniform over the entire logical range of the performance scale? If yes, proceed to Step 21. If no, the scoring procedures presented herein are inadequate to handle such a performance measure. Define a new performance measure, and return to Step 1.

Step 21. Does the reverse preference relationship identified in Step 20 maintain a constant rate of change of worth, or does it display a

variable rate of change (i.e., either accelerating, decelerating, or both in sequence)? If constant, proceed to scoring procedure 11 in Appendix XI. If variable, proceed to Step 22.

Step 22. Does the variable rate of change of worth identified in Step 21 display uniform acceleration, uniform deceleration, or first one and then the other? If uniform acceleration, proceed to scoring procedure 12 in Appendix XII. If uniform deceleration, proceed to scoring procedure 13 in Appendix XIII. If first one and then the other, proceed to Step 23.

Step 23. Does the variable rate of change identified in Step 22 start by accelerating and then end by decelerating, or does it start by decelerating and then end by accelerating? If the former, proceed to scoring procedure 14 in Appendix XIV. If the latter, proceed to scoring procedure 15 in Appendix XV.

Step 24. It has been determined that the performance scale is bounded from below by zero, but that the scale possesses no logical upper bound. What is the direction of the preference relationship? If direct, proceed to Step 25. If reverse, proceed to Step 28.

Step 25. Now fit the end-points of the worth scale to the performance scale. Assign zero worth points to zero performance and one worth point to infinite performance. Proceed to Step 26.

Step 26. Is the direct preference relationship identified in Step 24 uniform over the entire logical range of the performance scale? If yes, proceed to Step 27. If no, the scoring procedures presented herein are inadequate to handle such a performance measure. Define a new performance measure, and return to Step 1.

Step 27. The following facts have been ascertained concerning the nature and shape of the scoring function for this performance measure.

1. The worth scale is bounded between zero and one (by convention).
2. The physical performance scale is bounded from below by zero, but it possesses no logical upper bound.
3. The preference relationship is uniformly direct over the entire range of the performance scale.

From these three facts, we must conclude that both a constant rate of change of worth and a uniformly accelerating rate of change of worth are logically impossible. The only remaining possibilities are uniform deceleration or initial acceleration followed by deceleration. If uniform deceleration, proceed to scoring procedure 16 in Appendix XVI. If initial acceleration followed by deceleration, proceed to scoring procedure 17 in Appendix XVII.

Step 28. Now fit the end-points of the worth scale to the performance scale. Assign one worth point to zero performance and zero worth points to infinite performance. Proceed to Step 29.

Step 29. Is the reverse preference relationship identified in Step 24 uniform over the entire logical range of the performance scale? If yes, proceed to Step 30. If no, the scoring procedures presented herein are inadequate to handle such a performance measure. Define a new performance measure, and return to Step 1.

Step 30. The following facts have been ascertained concerning the nature and shape of the scoring function for this performance measure.

1. The worth scale is bounded between zero and one (by convention).

2. The physical performance scale is bounded from below by zero, but it possesses no logical upper bound.
3. The preference relationship is uniformly reverse over the entire range of the performance scale.

From these facts, we must conclude that both a constant rate of change of worth and a uniformly accelerating rate of change of worth are logically impossible. The only remaining possibilities are uniform deceleration or initial acceleration followed by deceleration. If uniform deceleration, proceed to scoring procedure 18 in Appendix XVIII. If initial acceleration followed by deceleration, proceed to scoring procedure 19 in Appendix XIX.

This completes the ordered sequence of questions designed to determine the general nature and shape of the scoring function.

2.2.5.2 Phase II Of The Scoring Procedure: Determining A Specific Scoring Function Of The General Nature And Shape Determined In Phase I

Twenty individual scoring procedures have been developed to define a unique scoring function to reflect a decision maker's worth judgments. In Phase I of the overall procedure, the general nature and shape of this function was determined by proceeding through an ordered sequence of questions. On the basis of answers given to these questions, one of the first nineteen procedures in Phase II is to be selected. The steps contained in the twentieth procedure are common to many of the other nineteen. Therefore, it has been segregated to economize on space.

The reader is hereby referred to Appendices I through XX for these twenty individual scoring procedures.

2.2.6 A Procedure For Adjusting The Weights To Reflect Differential Interpretive Quality Among The Physical Performance Measures

The last step in formulating an assessment algorithm is to adjust the weights to reflect differential interpretive quality among the physical performance measures. A step-by-step procedure to accomplish this is presented below.

Step 1. Compute the "effective" weight associated with each lowest-level performance criterion. That is, identify the chain of weights linking each lowest-level criterion to the apex of the hierarchy, and compute the product of all weights in this chain. Then, each of the "effective" weights associated, respectively, with one of the lowest-level criteria will be positive, and they will sum to one.

Step 2. Now consider the relationship between each lowest-level criterion and its associated physical performance measure. Recalling the scoring function which has been defined for each of these linked pairs, assess the extent to which the performance measure serves to interpret, through its scoring function, the intended meaning of the lowest-level criterion. Assess its interpretive quality on a percentage scale, where zero means that the performance measure bears no relation at all to the performance criterion, and one hundred percent means that the performance measure interprets perfectly the intended meaning of that criterion.

Step 3. Assign percentage numbers to each linked pair at the base of the criterion structure.

Step 4. Multiply each "effective" weight by the corresponding percentage number assigned in Step 3.

Step 5. Add the products computed in Step 4.

Step 6. Divide each product computed in Step 4 by the sum computed in Step 5. The result is a set of "adjusted effective" weights.

This completes the procedure.

2.2.7 A Procedure For Computing Each Alternative's Total Worth Score

The process of formulating an assessment algorithm was completed in Section 2.2.6. This Section describes the second stage in the complete assessment procedure wherein that algorithm is applied to an alternative to generate an index of its overall worth. A step-by-step procedure to accomplish application of the algorithm is presented below.

Step 1. Select one of the feasible alternatives.

Step 2. Select one of the performance measures in terms of which that alternative has been described.

Step 3. Referring to the scoring function associated with that performance measure, convert measured performance into an equivalent worth score.

Step 4. Multiply the equivalent worth score computed in Step 3 by the associated "adjusted effective" weight computed in Section 2.2.6.

Step 5. Repeat Steps 2 through 5 for all performance measures.

Step 6. Add the products computed in Step 5. The resulting sum constitutes an index of the selected alternative's overall worth.

This completes the procedure.

2.3.0 SOME CONCLUDING REMARKS ON THE ASSESSMENT PROCEDURE

A general procedure for assessing the worth of well-defined alternatives has been presented. Before moving on to Chapter III, wherein the complete procedure will be illustrated with a live example, it might be well to summarize our conclusions to date and close with some brief critical comments.

The first step in assessment is to define explicitly what is desired in the way of performance from produced alternatives to complete a stated job. If there were only one performance objective, then the task would be quite simple. A single performance criterion would be formulated, a single physical performance measure would be chosen, a single scoring function would be defined to convert promised performance into an equivalent worth score, and that alternative with the highest score would be defined as most worthwhile.

Unfortunately, however, the real task of assessment is much more complicated. Decision makers have more than one performance objective. Even worse, most performance objectives encompass more than one meaning. Consequently, a more complicated map of performance objectives must be constructed to reflect the more complex goal structure existing in the minds of decision makers. One convenient way to reflect this complex structure is by generating a hierarchy of worth-independent performance criteria.

At the base of the hierarchical criterion structure lie numerous lowest-level performance criteria. With each one of these is associated a physical performance measure designed to give it a concrete interpretation. Then, a scoring function is defined to link each physical performance measure with its lowest-level performance criterion. Frequently, these scoring functions take the form of mathematical formulas or curves (hopefully established prior to and always established indepen-

dently of receiving actual data describing alternatives); but sometimes they take the form of direct judgmental point assessments by decision makers without the aid of mathematical formulas or curves. In the latter case, scoring functions are implicit within the minds of decision makers rather than explicitly stated. Now what can be said in the way of interpreting scores, weights, and weighted scores generated by the above procedure?

One interpretation which is conceptually clear and logically sound appears below in outline form. (Note: This may not be the only clear and sound interpretation, but the writer cannot think of any others.)

1. The hierarchical structure of performance criteria reflects what is desired in the way of performance to complete a stated job.

2. Since more than one kind of performance is desired, more than one performance criterion appears in the hierarchical structure.

3. Since most of the performance criteria in the structure encompass several sub-criteria, this fact is reflected by making the structure hierarchical and tree-like.

4. To bridge the gap between desired performance and offered performance, physical performance measures are chosen for each lowest-level performance criterion in the structure, and scoring functions are defined (either explicitly or implicitly) to convert physical performance into equivalent worth scores.

5. A numerical scale (calibrated in points) is defined to reflect the perceived worth of whatever physical performance each feasible alternative promises. The output of each scoring function is some number of points in this worth scale.

6. If there were only one performance objective and, therefore, only one performance criterion in the hierarchy, that alternative which satisfied this single criterion to the greatest extent would be judged the most effective, worthy, etc.

7. However, there are many performance objectives reflected by many hierarchically arranged performance criteria. In addition, it is considered relatively more important to satisfy some of the numerous performance criteria than it is to satisfy others. Consequently, numerical weights are attached to each of the criteria themselves to reflect their perceived relative importance, and these weights are adjusted to reflect differential interpretive quality among the performance measures.

8. With this conceptual framework in mind, it is now possible to interpret scores, weights, and weighted scores.

   (a) A point score attached by a scoring function to a physical performance measure indicates the extent to which the related lowest-level performance criterion has been satisfied.

(b) An "effective" weight attached to a lowest-level criterion indicates how important it is to satisfy that criterion relative to satisfying other lowest-level criteria.

(c) Each "effective" weight is adjusted to reflect differential interpretive quality among the performance measures.

(d) A weighted point score (i.e., the product of "adjusted effective" weight and score assigned to a lowest-level performance criterion) indicates the individual contribution to total worth made by that particular piece of performance.

(e) The sum of the weighted point scores (summed over all performance criteria) indicates the total contribution to worth and, therefore, the total worth of each alternative.

Next, let us turn our attention to some particular aspects of the overall assessment methodology. First of all, it should be kept clearly in mind that the specific intent of the foregoing procedure has been to provide an orderly path by means of which decision makers may arrive at <u>their own</u> criterion hierarchy, <u>their own</u> weights, and <u>their own</u> independent scoring functions. No attempt has been made by the writer to substitute <u>his</u> judgments or <u>anybody else's</u> judgments for the judgments of those actually responsible for assessment. It is the writer's firm belief that this responsibility should <u>not</u> be usurped by a consultant (who might suggest his own hierarchy, weights, and/or scoring functions), <u>nor</u> should this responsibility be abdicated by decision makers and left up to the collective judgment of those who produce alternatives. Once it has been

agreed that responsibility for defining a hierarchy, assigning weights, and formulating scoring functions must be assumed by decision makers, and once it has been recognized that this responsibility must be assumed anew on each new decision, then there arises the need for a general assessment procedure. The procedure presented herein attempts to satisfy that need. It suggests an orderly and replicable sequence of steps - a general accounting framework - by means of which notions of worth may be formulated and mapped out explicitly in the form of an assessment algorithm.

Second, an attempt has been made to create an assessment procedure which will apply to almost all situations. Naturally, it would be foolhardy to allege that no exceptions will ever occur, but it is claimed that this procedure is sufficiently flexible to cover most situations.

Third, an attempt has been made to organize both the scoring and the weighting procedures in such a way as to exploit the collective experience of several decision makers while still preserving their separate and independent judgments. Thus, questions designed to determine the general nature and shape of a scoring function should be raised and answered in collective, group discussion. This will serve to bring everyone up to a common, agreed-upon basis of understanding. Then, individual scoring functions may be formulated separately and independently (cf., scoring procedure 20, Steps 3 and 4), and the results may be averaged (cf., scoring procedure 20, Steps 6 through 9). This will serve to reflect a more stable, less individually biased, and more representative cross-section of group opinion. This type of approach is also embodied in the weighting procedures.

Fourth, although a first reading of the overall procedure presented herein may be somewhat frightening, subsequent readings and - more particularly - a little actual practice using it should dissipate this

initial reaction. With practice, Phase I of both the scoring and the weighting procedures may be dispensed with in all but difficult situations. In the case of scoring, a Phase II procedure may frequently be selected immediately after only a cursory examination of the performance measure in question and a little informal thought about its relationship to the lowest-level performance criterion. In the case of weighting, sub-criteria may often be ranked immediately on the basis of mere inspection, particularly if there are a small number to be ranked. However, before either of these Phase I procedures may be circumvented, it is important that their content be thoroughly understood and that whatever short-cut methods are substituted for them generate the same results.

CHAPTER III

A LIVE EXAMPLE

3.1.0 BACKGROUND

One of the writer's colleagues, a graduate student at Massachusetts Institute of Technology, became interested in the assessment procedure when he was faced with securing employment directly following graduation.[1] He had already solicited several job offers and, on the basis of preliminary analysis, he had reduced these to a set of four feasible and non-dominated alternatives. It was at this point that he undertook the task of formal assessment.

After reading completely an earlier draft of this thesis and obtaining clarification on various procedural points from the writer, he set out to generate a criterion hierarchy, to establish weights, to define scoring functions, to adjust the weights, to assess the four alternative job offers, and, finally, to make a terminal decision. His progress through these sequential steps will be reported in the following sections of this chapter.

3.1.1 The Criterion Hierarchy

It would require too much space to present a complete historical record of this individual's progress through the various procedures involved in generating a criterion hierarchy, purging it of worth-interdependent members, and selecting physical performance measures. He made at least four separate passes at creating and revising a hierarchy over a period of several weeks time. What will be presented instead is the end state of this process. The hierarchy of worth-independent criteria

1. The individual involved has chosen to remain anonymous.

and associated performance measures which he finally selected as providing a satisfactory description of his job objectives is described below.

Four major objectives or highest-level performance criteria were defined:

1. monetary compensation;
2. geographical location;
3. travel requirements;
4. nature of work.

Monetary compensation was broken down to include:

1. immediate compensation;
2. future compensation.

Immediate compensation was further subdivided to include:

1. starting salary;
2. fringe benefits, which included -
   (a) insurance benefits;
   (b) retirement benefits.

Future compensation was subdivided to include:

1. anticipated salary in three years;
2. anticipated salary in five years.

His second major objective, geographical location, was broken down to include:

1. proximity to relatives;
2. degree of urbanity associated with the location;
3. climate.

His third major objective, travel requirements, was broken down to include:

1. daily commuting requirements to and from the place of work;
2. extended trips.

Extended trips was further subdivided to reflect:

1. proportion of time away from home;
2. duration of extended trips.

His fourth major objective, nature of work, was broken down to include:

1. immediate training requirements;
2. continuing aspects.

Continuing aspects of the work were further subdivided to include:

1. personal interest in the technical content of the job;
2. degree of variety implicit in the job;
3. amount of training in management skills realizable from the job.

The above hierarchy contained fifteen lowest-level criteria, each one of which was interpreted by defining a single performance measure. These fifteen lowest-level criteria and their associated performance measures were as follows:

1. starting salary - locally adjusted after-tax annual dollars;[2]
2. insurance benefits - locally adjusted after-tax annual dollars;[2]

2. All dollar figures were adjusted to account for differences in average living costs associated with different geographical locations in the United States.

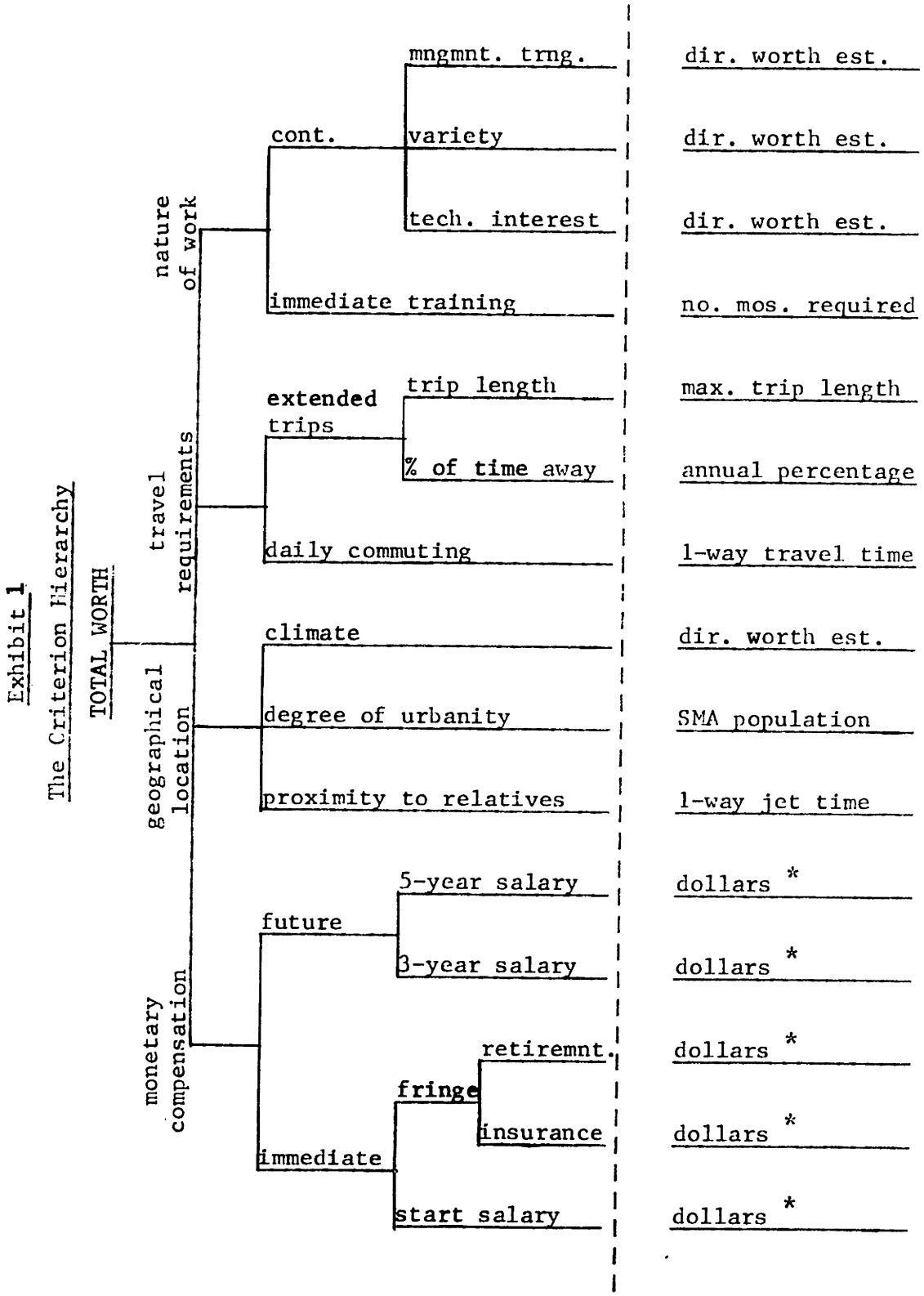
3. retirement benefits - locally adjusted after-tax annual dollars;[2]
4. anticipated three-year salary - locally adjusted after-tax annual dollars;[2]
5. anticipated five-year salary - locally adjusted after-tax annual dollars;[2]
6. proximity to relatives - one way jet flight time in hours;
7. degree of urbanity - standard metropolitan area population;
8. climate - direct worth estimate;[3]
9. daily commuting requirements - one-way travel time in hours;
10. proportion of time away from home - annual percentage;
11. duration of extended trips - maximum trip length in days;
12. immediate training requirements - required training time in months;
13. personal interest in the technical content of the job - direct worth estimate;[3]
14. degree of variety implicit in the job - direct worth estimate;[3]
15. amount of training in management skills realizable from the job - direct worth estimate;[3]

A pictorial display of this criterion hierarchy, complete with performance measures, is shown in Exhibit 1. The dotted horizontal line indicates the region of demarkation between performance criteria and performance measures. The reader will notice that abbreviations are sometimes used in Exhibit 1 to conserve space. However, review of the text

3. A direct worth point score was assigned subjectively to each alternative in this instance.

Exhibit 1

The Criterion Hierarchy



* All dollar figures locally adjusted after taxes.

should clear up any doubts about the meaning of these abbreviations.

3.1.2 The Weights

Numerical weights were then assigned to sub-criteria at every branching point in the hierarchy. For the major criteria, this process yielded the following weights:

1.	monetary compensation	.33
2.	geographical location	.17
3.	travel requirements	.17
4.	nature of work	.33
	Total	1.00

Within monetary compensation, weights were assigned as follows:

1.	immediate compensation			.70
	(a) starting salary		.90	
	(b) fringe benefits		.10	
	(1) insurance benefits	.60		
	(2) retirement benefits	.40		
	Total	1.00		
	Total		1.00	
2.	future compensation			.30
	(a) anticipated three-year salary		.65	
	(b) anticipated five-year salary		.35	
	Total		1.00	
	Total			1.00

Within geographical location, weights were assigned as follows:

1.	proximity to relatives	.40
2.	degree of urbanity	.40
3.	climate	.20
	Total	1.00

Within travel requirements, weights were assigned as follows:

1.	daily commuting requirements		.20
2.	extended trips		.80
	(a) proportion of time away from home	.40	
	(b) duration of extended trips	.60	
	Total	1.00	
	Total		1.00

Finally, within nature of work, weights were assigned as follows:

1.	immediate training requirements		.40
2.	continuing aspects		.60
	(a) personal interest in the technical content of the job	.50	
	(b) degree of variety implicit in the job	.30	
	(c) amount of training in management skills realizable from the job	.20	
	Total	1.00	
	Total		1.00

The above assignment of weights lead to the following distribution of "effective" weights on each of the fifteen lowest-level performance criteria:

1.	starting salary	.208
2.	insurance benefits	.014
3.	retirement benefits	.009
4.	anticipated three-year salary	.064
5.	anticipated five-year salary	.035
6.	proximity to relatives	.068
7.	degree of urbanity	.068
8.	climate	.034
9.	daily commuting requirements	.034
10.	proportion of time away from home	.054
11.	duration of extended trips	.082
12.	immediate training requirements	.132
13.	personal interest in the technical content of the job	.099
14.	degree of variety implicit in the job	.059
15.	amount of training in management skills realizable from the job	.040
	Total	1.000

3.1.3 The Criterion Scores

Of the fifteen performance measures listed in Section 3.1.1 and displayed in Exhibit 1, only eleven were defined in such a manner as to require explicit scoring functions. In the remaining four instances, he decided to assign direct worth estimates to the relevant aspects of each alternative job offer. All eleven of the explicit scoring functions were sketched by a graphical technique similar to the one set forth in scoring procedure 20, Appendix XX of this thesis.

Exhibit 2 below shows the estimated performance of each of the four alternatives on his fifteen performance measures.

Exhibit 2

Estimated Performance

Performance Criterion	Alt. I	Alt. II	Alt. III	Alt. IV
starting salary	$ 8,100/yr.	$ 8,250/yr.	$ 8,733/yr.	$ 8,550/yr.
insurance benefits	$ 475/yr.	$ 550/yr.	$ 475/yr.	$ 400/yr.
retirement benefits	$ 750/yr.	$ 1,000/yr.	$ 1,100/yr.	$ 875/yr.
three-year salary	$11,250/yr.	$ 9,500/yr.	$10,500/yr.	$10,500/yr.
five-year salary	$15,000/yr.	$10,500/yr.	$11,500/yr.	$11,500/yr.
proximity to relatives	0 hrs.	0 hrs.	5 hrs.	1 hr.
degree of urbanity	2.5 million	2.5 million	1.0 million	15.0 million
climate	*	*	*	*
daily commuting	.50 hrs.	1.00 hrs.	.25 hrs.	1.25 hrs.
% time away	0 %	10 %	0 %	35 %
extended trip duration	0 days	5 days	0 days	20 days
required job training	9.0 months	.5 months	1.0 months	.5 months
interest in job	*	*	*	*
variety	*	*	*	*
training in management	*	*	*	*

* means direct worth estimate was made

Exhibit 3 below shows the worth scores assigned either by graphical scoring functions or by direct worth estimation to the performance data associated with each alternative.

Exhibit 3

Assigned Worth Scores

Performance Criterion	Alt. I	Alt. II	Alt. III	Alt. IV
starting salary	.68	.70	.75	.73
insurance benefits	.60	.70	.60	.50
retirement benefits	.60	.80	.90	.70
3-year salary	.75	.63	.70	.70
5-year salary	.75	.45	.53	.53
proximity to relatives	1.00	1.00	.10	.50
degree of urbanity	1.00	1.00	.70	.80
climate	.70 *	.70 *	.85 *	.60 *
daily commuting	.60	.50	.90	.40
% time away	1.00	.70	1.00	.35
extended trip duration	1.00	.70	1.00	.50
required job training	.50	.90	.80	.90
interest in job	.40 *	.60 *	.75 *	.85 *
variety	.50 *	.80 *	.70 *	.90 *
training in management	.70 *	.85 *	.75 *	.80 *

* means direct worth estimate was made

3.1.4 The Adjusted Effective Weights

His next step was to adjust the "effective" weights according to the perceived interpretive quality of each performance measure. This lead to a set of "adjusted effective" weights which could then be applied to the worth scores shown in Exhibit 3. The original "effective" weights, the adjusting factors, and the final set of "adjusted effective" weights are shown below in Exhibit 4.

Exhibit 4

"Effective" Weights, Adjusting Factors, and "Adjusted Effective" Weights

Performance Criterion	"Effective" Weights	Adjusting Factors	"Adjusted Effective" Weights
starting salary	.208	1.00	.268
insurance benefits	.014	.95	.017
retirement benefits	.009	.95	.012
3-year salary	.064	.75	.062
5-year salary	.035	.75	.034
proximity to relatives	.068	.80	.069
degree of urbanity	.068	.75	.066
climate	.034	.90	.040
daily commuting	.034	.85	.037
% time away	.054	.50	.035
extended trip duration	.082	.85	.090
required job training	.132	.70	.118
interest in job	.099	.60	.076
variety	.059	.60	.045
training in management	.040	.60	.031

3.1.5 The Total Worth Scores

His last step was to multiply the criterion scores by their "adjusted effective" weights and add the products to determine each alternative's total worth score. The results of these computations are shown in Exhibit 5 below.

Exhibit 5
Total Worth Scores

Performance Criterion	Alt. I	Alt. II	Alt. III	Alt. IV
starting salary	.182	.187	.201	.195
insurance benefits	.010	.012	.010	.009
retirement benefits	.007	.010	.011	.008
3-year salary	.047	.039	.043	.043
5-year salary	.026	.015	.018	.018
proximity to relatives	.069	.069	.007	.035
degree of urbanity	.066	.066	.046	.053
climate	.028	.028	.034	.024
daily commuting	.022	.019	.033	.015
% time away	.035	.025	.035	.012
extended trip duration	.090	.063	.090	.045
required job training	.059	.106	.094	.106
interest in job	.030	.046	.057	.065
variety	.023	.036	.032	.041
training in management	.022	.026	.023	.025
Total Worth	.716	.747	.734	.694

Inspection of Exhibit 5 shows that Alternative II achieved the

CHAPTER IV

THE PROCEDURE IN PERSPECTIVE

4.1.0 INTRODUCTION

Having developed and illustrated the assessment procedure, it will now be placed in perspective. This will be accomplished in two ways. First, its relationship to six separate areas of interest will be discussed and illustrated. These areas include:

1. current practices;
2. statistical decision theory;
3. operations research;
4. descriptive decision theory;
5. organization theory;
6. social psychology;

Second, a guide to related research will be presented, although the research itself will not be discussed in any detail.

4.1.1 The Procedure And Current Practice

It would be impossible to describe all of the assessment procedures currently used by practical decision makers. Moreover, it would serve no useful purpose. What will be provided instead is a description of important differences between the assessment procedure developed herein and various others.

Certainly the most striking difference between this procedure and most alternative approaches to worth assessment is its primary and explicit focus upon the decision maker himself. The more typical approach is to focus explicitly upon the decision context and the alternatives, but only implicitly upon the decision maker. In this manner, an aura of objectivity is created, which many consider to be safer, if not more desirable.

Now these two types of approaches are not inconsistent with one another. Quite to the contrary, they are mutually reinforcing, but with a distinct change in both focus and emphasis. Let us develop this theme more fully.

A traditional objective in making business decisions is to maximize dollar profits. Although it is not always clear whether this is the only, the primary, or just another objective, it is a very common one. One substantial gain realized from defining such an objective is that many of the performance consequences of various alternatives may be expressed in terms of a single unit -- dollars. This, in turn, permits comparing the alternatives themselves in terms of that same single index. Nevertheless, there are certain complications.

1. Not all important consequences of a decision can be measured meaningfully in dollar terms (e.g., consequences in terms of corporate image, good will, etc.).

2. Some consequences, although measurable in dollar terms, are important for entirely different reasons (e.g., the salary of an important executive would not reflect the true loss which would be suffered by a firm should he die - particularly if the firm could not continue without his unique skills).

3. Considering dollars both as a unit of worth and as a resource may lead to confusion. Resource dollars and worth dollars may not really be commensurable - particularly if decision makers in fact view them differently, as sometimes happens under complex budgeting and accounting systems.

Systems analysts obviate some of the above problems by defining a different unit of worth (usually called system effectiveness). It is recognized that profit maximization is not the objective at all (particularly for military and government organizations), and either a high-order performance measure (e.g., the kill probability of a missile) or another resource (e.g., equivalent number of man-hours of labor required) is used to render multiple performance consequences and, therefore, whole alternatives commensurable.

Now both of the above approaches are perfectly compatible with our assessment procedure. In the simplest case, there is logical equivalence. If either dollar profits or some single aspect of system effectiveness were considered the only objective in a decision, then a simple criterion hierarchy would be constructed containing only one criterion and only one performance measure. It is in more complicated cases, where both multiple objectives and multiple performance measures are considered relevant, that certain complications arise. It is also in

such cases that the mutually reinforcing aspects of our procedure and alternative approaches become most apparent.

Review of the assessment procedure will show a point of tangency between it and both kinds of traditional approaches at the moment where physical performance measures are selected (see Section 2.2.3). Conclusions drawn by a cost accountant or by a systems analyst would be relevant in selecting such measures. However, achieving commensurability is accomplished in our procedure by an entirely different technique. Rather than seeking a very high-order performance or resource measure, and thereby achieving commensurability in phenomenological terms, an even higher-order measure is sought. Decision makers are asked to produce that measure themselves on the basis of their objectives, their experience, and their judgment. The measure in terms of which all performance consequences are rendered commensurable is worth points, and our procedure is the means of generating it. This difference in both focus (i.e., upon decision makers themselves) and emphasis (i.e., in achieving commensurability in the non-objective worth sense) is particularly important when broad policy issues are at stake.

There are other differences as well. These are nowhere near as important as the one just discussed, but they deserve mention.

One aspect of our procedure which may not seem novel (but really is) is its linearity. Recall that linear weights are assigned to criteria throughout the hierarchy to indicate the relative importance of satisfying them. However, this does not imply linearity in the various performance measures. Since decision makers are free to define non-linear scoring functions linking performance measures to performance criteria, the overall index of worth can be non-linear in the measures themselves. Where linearity does seem appropriate, linear scoring

functions may be defined. but this is not essential. In this added flexibility our procedure is fairly unusual.

Another difference arises from the independent (as opposed to dependent or relative) technique of scoring alternatives. Whereas many assessment schemes score alternatives relative to each other, our procedure assigns scores relative to independently pre-assigned criteria. The criterion hierarchy, the weights, and the scoring functions define what is desired in the way of performance and distinguish this clearly from what is offered.[1]

Finally, our procedure is almost completely general - at least in principle. Only the judgement of decision makers serves to restrict the range of jobs, of alternatives, or of decision situations for which it is potentially applicable.

4.1.2 The Procedure and Statistical Decision Theory

Historically, one of the prime incentives for developing the assessment procedure sprang from statistical decision theory. Defining an explicit loss structure constitutes one of the important steps in applying this theory to practical problems. Our assessment procedure, in conjunction with utility-measuring procedures, could provide assistance in defining an explicit loss structure. If successful, decision theoretic approaches are thereby rendered more operational.[2]

---

1. The reader is referred to Section 1.4.5 for a discussion of the distinction between dependent and independent scoring procedures.

2. The reader is referred to Section 1.5.0 for a brief discussion of worth, risk, uncertainty, and classical utility. A more complete discussion of defining loss structures and applying statistical decision theory can be found in Schlaifer, R., Probability and Statistics for Business Decisions (1959), and in Raffa, H., and Schlaifer, R., Applied Statistical Decision Theory (1961). A practical application is reported in Grayson, C.J., Decisions Under Uncertainty: Drilling Decisions by Oil and Gas Operators (1960).

4.1.3 The Procedure and Operations Research

A similar relationship holds between the assessment procedure developed herein and operations research. Proper definition of an objective function to be optimized is critical to any useful application of operations research techniques, and our procedure could be helpful in this regard also.[3]

4.1.4 The Procedure and Descriptive Decision Theory

A great deal of attention has been paid in recent years to the empirical validation of what were originally normative decision theories. In particular, the principle of maximizing expected utility has been studied extensively. One of the problems associated with conducting such empirical research is to obtain independent measures of both utility and probability. Our assessment procedure could be helpful in generating (at least part of) a utility measure which, hopefully, would be independent of perceived probabilty. If successful, utility maximization and other descriptive decision rules could be more easily subjected to empirical test.[4]

4.1.5 The Procedure and Organization Theory

Several theories of organizational behavior deal similarly with decision making. One in particular is the Barnard-Simon theory of an individual's decision to participate in an organization and the implications of this decision for the organization's health and long-run survival. Included within this theory are such constructs as:

3. The reader is referred to Churchman, C.W., Prediction and Optimal Decision: Philosophical Issues in a Science of Values (1961), and to Churchman, C.W., et al., Introduction to Operations Research (1959) for a more complete discussion of these issues. A somewhat different approach can be found in Smith, N.M., Jr., "A Calculus for Ethics: A Theory of the Structure of Value", Behavioral Science, 1956 (1), pp. 111-142 and pp. 186-211.

4. A survey of empirical studies in decision making can be found in

1. inducements - payments made by or through an organization to its participants (e.g., monetary compensation, objective services, etc.) and measured in phenomenological terms;

2. contributions - payments made by participants to the organization (e.g., man-hours of labor, capital investments, etc.), and also measured in phenomenological terms;

3. inducement utilities - numerical measures assigned by a hypothesized individual utility function to each inducement and indicating the value of that inducement to the individual participant;

4. contribution utilities — numerical measures of the utility of alternative activities (e.g., leisure time a different job, etc.) which must be forgone in order to make the required contributions to the organization.

Participation behavior is then explained and predicted in terms of the net balance existing between inducement utilities and contribution utilities. The health and survival of an organization are explained and predicted in terms of the individual utility balances achieved by its members.[5]

4. (cont.) Edwards, W., "Behavioral Decision Theory", Annual Review of Psychology, 1961 (12), pp. 473-498.

5. A complete description of this and many other theories of organizational behavior can be found in March, J.G., and Simon, H.A., Organizations (1958).

Unfortunately, however, no explicit method for operationalizing these constructs (particularly the utilities) is presented. Only a general approach is sketched. The assessment procedure developed herein (with embellishments to account for risk/uncertainty) could be used to provide operational interpretations of both inducement and contribution utilities. Inducements would constitute what we have been calling performance measures, and contributions would play the role of resources. Overall utility would be defined (in part) by the overall worth measure. Reinterpreted in this manner, the Barnard-Simon theory would become operational and, therefore, subject to empirical validation.

4.1.6 The Procedure and Social Psychology

The assessment procedure could be used to operationalize various theories of social psychology in very much the same way. A specific case in point is Homans' theory of dyadic social interaction. Formally, this is quite similar to the Barnard-Simon theory, although the key constructs have different labels (i.e., "rewards" instead of "inducement utilities", "costs" instead of "contribution utilities", and "profit" instead of "net utility balance"). The same remarks apply here as were stated above.[6]

In a somewhat less direct manner, the assessment procedure could be used to operationalize such concepts as cognitive dissonance (Festinger) and rationalization (Freud). One important consequence of going through the complete procedure is the creation of cognitive dissonance (i.e., the

6. A complete description of Homans' theory can be found in Homans', G.C., Social Behavior: Its Elementary Forms (1961).

focusing of conscious attention upon seemingly contradictory elements in the worth structure), and one very common way of reducing this dissonance is by a process of rationalization.[7] The experiment to be reported in Chapter V brought these two points out very dramatically.

More generally, the procedure can be used as a kind of psychological measuring instrument. Two studies have already been performed using a simplified variation of the procedure to investigate the relative contribution of various sources of job satisfaction to overall satisfaction. One of these studies compared work groups in a profit-making research laboratory to groups working on similar problems in a government laboratory.[8] The second study made similar comparisons both between different kinds of groups at the same time and longitudinally over time.[9] Preliminary results of both studies were promising, and this research will be reported separately.

Finally, a completely descriptive application of the procedure to construct parametric predictive algorithms for social research is currently under development. One such application has already been made, but it is too early to assess its success.[10]

---

7. The writer is indebted to Peer O. Soelberg for pointing out the implications of the procedure in terms of inducing rationalization.

8. The writer is indebted to William M. Evan for providing research data on which to perform this analysis.

9. The results of this study will be reported by David Morrisroe as part of his work towards a Master's degree in Business Administration at the Harvard Business School.

10. Thanks are due to Richard S. Rosenbloom for providing research data and arranging the initial support of this project.

4.2.0 A GUIDE TO RELATED RESEARCH

In addition to the previous references, which constitute only a sampling of the various areas involved, a rather extensive review of assessment procedures in several fields is provided by Cronbach and Gleser (3). Beside describing many such procedures and their applications in detail, this book also directs the reader to similar reviews made by other authors. It would only serve to duplicate effort if the same material were reviewed again in this thesis. Hence, the interested reader is referred to the above mentioned book.

CHAPTER V

AN EXPERIMENTAL TEST OF THE PROCEDURE AND RELATED FACTORS INFLUENCING THE DECISION MAKING PROCESS

5.1.0 INTRODUCTION

In Chapters I, II, and III of this thesis, a systematic procedure to aid in the assessment of worth was first developed and then illustrated with a live example. The purpose of this procedure, it will be recalled, is to help decision makers formulate and articulate a consistent assessment structure (really a complex algorithm) for assessing the worth of specified alternatives in a definite choice situation. Once formulated, this assessment algorithm may be applied to each specified alternative so as to generate a numerical index of its overall worth.

The experiment to be reported in this chapter was designed with a dual objective in mind. First, it was designed to validate the assessment procedure - that is, to determine whether or not the procedure could be implemented by professional decision makers, and, if so, with what degree of success. Second, it was designed to test experimentally the impact of five specific factors upon the process of developing preferences for alternatives and eventually choosing one of them. These five factors will be described in Section 5.1.3.

5.1.1 A Brief Review Of The Worth Concept

In order to recall the conceptual foundations of the assessment procedure and to motivate discussion of the experiment, five critical

assumptions about the worth concept are restated below.[1]

1. Worth is an internal property of human beings. Worth notions exist within the perceptual and attitudinal apparatus of human decision makers - not as external properties of the physical objects and activities which human beings assess and to which they impute worth. To assess the worth of an object or activity, therefore, is to measure a decision maker's response (e.g., verbal assessments, behavioral choice, etc.) to that object or activity.

2. In general, human notions of worth are multidimensional rather than unidimensional. This means two things:
   a. A given physical object or activity is perceived as relevant simultaneously to more than one human objective;
   b. A given human objective may be satisfied by more than one alternative object or activity.

3. An individual's notions of worth need not necessarily be shared by others (i.e., consensual validation is not a <u>definitional</u> requirement of legitimate worth notions), although some consensus can be expected, particularly within his reference group.

---

1. The reader is referred to Section 1.3 of this thesis for a more complete discussion of the worth concept.

4. An individual's notions of worth need not necessarily be stable over time (i.e., temporal stability is not a definitional requirement of legitimate worth notions), although some stability can be expected, particularly where his more important values are concerned.

5. Worth notions do not usually exist in a conscious, clearly defined, and logically structured form within the minds of human decision makers. However, with some effort, a consistent assessment structure can be formulated to reflect an individual's notions of worth, so long as certain practical limitations on the ability to conceptualize are observed.

The above considerations suggest the need for a systematic and replicable procedure whose purpose is to help individual decision makers formulate, validate, and explicate a multidimensional assessment structure to aid them in choosing among complex alternatives.

5.1.2 A Brief Review Of The Assessment Procedure

The assessment procedure, it will be recalled, involves several sequential operations. These are outlined below.[2]

2. The reader is referred to Chapter II of this thesis for a more complete discussion of the assessment procedure.

1. Assuming that a job to be done and/or a set of activities to be performed has been described, formulate a list of overall job objectives by abstraction from the job description.

2. Refine each higher-level objective in terms of two or more lower-level performance criteria which define more precisely what is intended by or subsumed under the meaning of the higher-level objective. Generate thereby a complete criterion hierarchy.

3. Interpret lowest-level criteria in terms of physical performance measures.

4. Specify individual worth relationships perceived as holding between each lowest-level criterion and its linked performance measure.

5. Establish an overall index of worth, considering all of the previously listed objectives and sub-criteria simultaneously.

If a decision maker can successfully complete the above five operations he will have created an assessment structure (really a complex algorithm) by means of which a single cardinal worth number may be assigned to any specified alternative in a given choice situation. Inputs to this assessment algorithm consist of various physical performance measures selected by the decision maker as describing the relevant measurable attributes of an alternative. The output of this assessment

algorithm is a single cardinal number purporting to represent the worth imputed by the decision maker to that alternative.

5.1.3 A Formal Statement Of Purpose

As stated previously, one purpose of the experiment was to validate the assessment procedure on professional decision makers. A second purpose was to test experimentally the impact of each of the following five factors upon the process of developing preferences for specified alternatives and eventually choosing one of them:[3]

1. No information, no guidance -- the provision of neither information about the alternatives nor guidance on how to assess their worth or make a final choice;

2. Raw information -- the provision of raw information about the alternatives, but without any systematic guidance on how to utilize such information in assessing their worth or arriving at a final choice;

3. Ordinal guidance -- the provision of guidance toward formulating a lexicographic assessment structure which would utilize such information to rank-order the alternatives according to their perceived relative worth;

---

3. The writer is indebted to Donald G. Marquis not only for his helpful suggestions regarding the general design of this experiment, but also for his specific suggestion that the impact of raw information be subjected to explicit experimental test.

4. Cardinal guidance -- the provision of additional guidance toward formulating a more complex assessment structure (i.e., an assessment algorithm like the one described in Chapter II) which would assign single cardinal worth numbers to each alternative;

5. The act of making a final choice.[4]

None of the assessment procedure developed in Chapter II would be relevant to testing the impact of either the first factor -- no information, no guidance -- or of the second factor -- raw information without guidance. The first three operations of this procedure (as outlined in Section 5.1.2), plus an additional operation designed to induce a lexicographic assessment structure, could be utilized to test the impact of the third factor -- ordinal guidance. The complete assessment procedure, including all five operations and the assignment of cardinal worth numbers to alternatives, would serve to test the impact of the fourth factor -- cardinal guidance. The impact of the fifth factor -- the act of making a final choice -- must be tested outside the bounds of the assessment procedure.

Several experimental measures were defined, and the assessment procedure was utilized in the above ways to determine the impact of each of these factors upon individual decision processes. Military and Civil

4. The operational meaning of all five experimental factors and their descriptive terminology will be clarified by subsequent discussion.

Service personnel, whose professional duties include the making of continuing choices among alternative weapons systems, were used as experimental subjects.

The overall intent of this experiment was exploratory. It was not designed to validate existing decision theories (either normative or descriptive), nor was it designed to test specific hypotheses about individual decision processes. The writer did make several informal guesses about various experimental outcomes, but none of these were held with sufficient confidence before conducting the experiment to deserve the formal title of a hypothesis. In no instances were these informal guesses inferred deductively from existing theories. Consequently, describing the experiment and analyzing its outcomes in the classical language of hypothesis testing would be misleading. The intent of the experiment was to validate the assessment procedure and to test the impact (whatever it might be) of the five experimental factors previously discussed. It is in this spirit that results will be reported, that analyses will be discussed, and that conclusions will be drawn.

5.2.0 DESIGN OF THE EXPERIMENT

Several years ago, the Department of Defense established a school at Wright-Patterson Air Force Base to train military and Civil Service personnel in the intricacies of modern weapons systems management. Military officers from all three branches of the Armed Services and Civil Service personnel from various defense-oriented government agencies (e.g., the National Aeronautics and Space Administration) are selected four times each year to participate in an eleven-week training course. A class consists of approximately sixty such individuals holding the rank of Colonel, Lieutenant Colonel, Captain (Navy), Commander, GS-14, GS-15, or the equivalent, and with at least some (in most cases, substantial) prior experience managing government projects. Since the purpose of the course is to train project managers, a large part of the curriculum is devoted to new techniques in scientific management -- particularly those espoused by the Department of Defense. The eleventh and final week of the course consists of a computer-simulated game played by teams of five participants each. The computer is programmed to simulate contractor responses to various decisions made by each team as it progresses through the design, selection, installation, and eventual operation of a typical weapons system.

This eleven-week training course constituted the setting of the experiment. The sixty military and Civil Service personnel being trained for duty as project managers comprised the sample of experimental subjects.

5.2.1 Specific Design Objectives

In designing the experiment, the following specific objectives were set forth.

1. First, it seemed essential to select a sample of experimental subjects who regularly make important decisions among complex alternatives. After all, it is for precisely this kind of person that the assessment procedure was primarily (if not exclusively) designed. It is not clear that other kinds of people would be either willing or able to indertake such an arduous task.

2. Second, it seemed desirable to have each subject make a definite and clearly observable decision (i.e., choice among alternatives) concerning some issue which he would regard as meaningful and whose consequences would be directly and visibly related to his future well-being. By requiring each subject to make an observable choice, experimental measures of preliminary preferences (for the various alternatives) could be formulated and later tested for their ability to predict his final choice. By selecting an issue which he would perceive as both meaningful and bearing directly upon his future well-being, each subject could be expected to expend a reasonable amount of time and effort in formulating an assessment structure and applying this to the alternatives.

3. Third, to remain compatible with the assessment procedure, the choice had to be constrained to a fixed set of discrete and clearly specified alternatives.

4. Fourth, it seemed desirable to have all sixty subjects assess similar alternatives in an identical decision situ-

ation. This would permit making valid comparisons of results across subjects.

5. Fifth, it seemed desirable to have all sixty subjects make their assessments independently of one another. The focus of this experiment was upon individual (as opposed to group) decision making processes. In addition, maintaining independence across individual decision-makers would permit more powerful statistical analyses of results.

6. Sixth, it seemed desirable to make the decision situation relatively simple, relatively familiar, and restricted to a manageable number of alternatives. This would serve to economize time and effort both on the part of the experimenter and on the part of the experimental subjects (no prior training required).

7. Finally, in accordance with the second overall purpose of the experiment, experimental manipulations were designed in such a way as to obtain independent measures of the impact of each of the five factors:

    a. no information, no guidance;

    b. raw information, without guidance;

    c. ordinal guidance;

    d. cardinal guidance;

    e. the act of making a final choice.

5.2.2 The Decision Situation, The Alternatives, And The Final Choice

Recall that all sixty experimental subjects form teams of five participants each at the end of the training course. Through the medium

of a computer game against simulated defense contractors they then proceed to test their newly-acquired knowledge. For purposes of this experiment, the decision which each individual subject had to make was to choose partners and thereby form a team to play the computer game.

Assuming five-man teams (of which there would be twelve), an alternative would consist of a group of four other participants who, along with the individual making the choice, would constitute a complete five-man team.

If the individual subject were permitted to choose any four partners from among the entire remainder of the class, then he would have to consider over 455,000 alternative teams. This is obviously too many for any one person to handle. Consequently, a series of experimental devices had to be employed in order to reduce the alternatives to a manageable number.

The first device was to subdivide the sixty subjects into six sub-groups of ten each. Subdivision was to be performed prior to the beginning of the training course with the aid of a random number table. Then, each subject would be asked to peruse a list of ten names (including his own) and to subdivide the remaining list of nine others into two sub-lists. The first sub-list would contain six names of preferred candidates for inclusion in his final team, while the second sub-list would contain the three remaining names. Subjects would be asked to perform this latter subdivision after having had a few days to acquaint themselves with other participants in the training course. By means of these two devices, each subject would then have only six other candidates from whom to choose four team partners. This would serve to reduce the number of alternative teams which each individual must consider to fifteen.

However, despite these experimental devices, there still remained the problem of giving each subject an independent choice to make. Except by unlikely accident, not every individual in a ten-man sub-group could have his complete choice of partners fulfilled. If two or more individuals included the same third individual in their most preferred team, but failed to include each other, then somebody would have to lose. Consequently, a third experimental device had to be employed to obviate this difficulty and to maintain the prospect of an independent decision for all sixty subjects. It was decided to announce at the outset of the experiment that one subject in each of the ten-man sub-groups would have his choice of team partners honored. The remaining five subjects not chosen by him would be grouped to form a second team. Exactly whose choices were to be honored would remain undetermined until the end of the experiment, and a random number table would be used to make this determination at that time. Therefore, each subject might proceed on the assumption that he would be making the final choice, for his chances would be just as good as anyone else's of having his choice honored.

5.2.3 Satisfaction Of The Specific Design Objectives

Let us now review the seven design objectives set forth in Section 5.2.1 and see how the forgoing considerations give promise of satisfying them.

The first objective -- validating the assessment procedure on professional decision makers -- should be satisfied by the particular choice of experimental subjects and the experimental setting. All sixty participants in the training course are sent to the school for the express purpose of receiving education in decision making. Most of them have had extensive practical experience in assessing and choosing among complex alternatives prior to coming. The curriculum focuses heavily upon decision making techniques, and the work-pace is intensive. Students live on the Air Base throughout the eleven-week period and are required to attend six hours of class each day. Consequently, on the basis of these personal background and contextual factors, it seemed reasonable to hope that both the subjects and the setting would provide an appropriate vehicle for validating the assessment procedure.

The second objective -- having each subject make a definite and clearly observable decision -- should be satisfied by requiring everyone to choose four team partners at the end of the experiment, just prior to playing the computer game. The choice would be definite. It would be clearly observable by the experimenter (although not by the subject's fellow students). It could have an immediate impact upon his chosen team's performance in the game itself. Since the game was advertised in advance as competitive, and since previous participants in the game had demonstrated substantial personal committment and competitive zeal, it was reasonable to hope that subjects would take the experimental decision seriously.

The third objective -- providing a fixed set of discrete alternatives -- should be satisfied by means of the first two experimental devices. Each subject would have a fixed set of fifteen discrete teams from which to make a final choice. The number fifteen was decided upon by the writer as a compromise between two countervailing desires. On the one hand, the larger the number of alternatives presented to each subject, the greater would be the range of potential preference orderings he could assign thereto. Experimentally induced alterations in preference orderings could then vary over a much wider range which, in turn, would permit a clearer interpretation of results. On the other hand, as the number of alternatives increased, subjects would find it increasingly difficult to consider them all simultaneously and make meaningful comparisons. Semi-formal pre-testing of various numbers of alternatives suggested that fifteen would provide an appropriate compromise. Hence, fifteen were adopted, and the first two experimental devices were designed accordingly.

The fourth objective -- having all sixty subjects assess similar alternatives in an identical decision situation -- should likewise be satisfied by these two experimental devices.

The fifth objective -- inducing each subject to make an independent decision -- should be satisfied by the third experimental device. By announcing in advance that an individual's choice of team partners would be honored if and only if his name were selected by a completely random mechanism and without regard to whom he chose or who chose him, it was hoped to discourage the formation of coalitions and the adoption of competitive bidding strategies. In addition, it was decided to give continual instructions to the subjects requesting that they refrain from discussing with one another their preferences, their

assesment criteria, or their anticipated final choices.

The sixth objective -- presenting a relatively simple and familiar decision situation -- should be satisfied by the nature of the required choice. Choosing partners for some group enterprise is a familiar decision made many times in almost everyone's lifetime. Choosing up sides for an athletic contest or parlor game, selecting new members for a social or business organization, and choosing a marriage partner are common examples. Because of its familiarity and restriction to fifteen alternatives, the decision situation in this experiment should therefore satisfy the sixth objective.

Satisfaction of the seventh design objective -- testing the impact of the five experimental factors -- will be discussed in the context of introducing these factors (see Section 5.2.6). Also, as implementation of the design is described (in Section 5.3), comments will be made concerning the extent to which the above objectives were in fact satisfied.

5.2.4 Specific Effects To Be Observed

In testing the impact of the five experimental factors upon the decision process the writer was interested in observing several experimental effects. These might best be described in terms of the specific questions formulated prior to conducting the experiment and referring to each of the five experimental factors separately.

1. Will introduction of the factor into the decision making process serve to clarify, to confuse, or to have no noticeable impact upon an individual's preferences? If there is a noticeable impact, how great is it?

2. Will introduction of the factor increase, decrease, or have no noticeable impact upon the number of preference discriminations made? If there is a noticeable impact, how great is it?[5]

3. Will introduction of the factor increase, decrease, or have no noticeable impact upon an individual's confidence in the accuracy of his indicated preferences? Again, how much?

4. How satisfied will an individual be with whatever assessment structure results from introduction of the factor? Specifically, will he consider it helpful in improving the quality of his final choice. If so, by how much?

5. If given a choice between this experimentally induced assessment structure and alternative structures (spontaneously generated and/or induced by other experimental manipulations), to what extent will he choose this

5. The writer is indebted to Gordon M. Kaufman for suggesting that this effect be investigated and for suggesting a technique to do so (i.e., permitting subjects to indicate indifference among alternatives rather than forcing them always to indicate a distinct preference).

structure as more representative of his current preferences for alternatives?

6. Will introduction of the factor have any impact on altering preferences? If so, how much and in what ways?

7. How aware will an individual be of the extent to which his preferences become so altered?

8. To what extent will an individual feel that any gains made in clarification, confidence, satisfaction, and/or appropriate alteration were worth the additional costs to him in time and effort expended to realize these gains?

9. To what extent and in what ways will introduction of the factor serve to alter his attitudes toward formal assessment procedures?

10. To what extent will he adapt the formal assessment procedures presented (during introduction of the factor) to other decision situations lying outside of the experiment?

It should be noted that all of the above questions were asked about each of the five experimental factors separately. In addition, similar questions were formulated concerning the differential impact of these factors, and a procedure was designed so that both direct and differential effects could be observed simultaneouly. More will be said about this in Section 5.2.6.

Since the fourth experimental factor is really the complete assessment procedure described in Chapter II, the above questions also define the particular senses in which the writer chose to validate that procedure.

5.2.5 Experimental Measures Constructed

To answer the above questions, a battery of experimental measures was constructed. These measures and their relationship to each question will be discussed in the following paragraphs.

Five written questionnaire items were constructed to measure the separate clarifying effect of each experimental factor. In all cases, a five-point rating scale was defined, and subjects were asked to indicate which rating category best described the extent to which introduction of the factor had clarified their preferences for alternatives. The text of these items is reproduced below.

> (description of the factor) may have clarified your preferences for the various groups. Please indicate on the scale below whether or not your preferences have been clarified by virtue of (description of the factor) and, if so, to what extent. Circle the appropriate point on the scale below.

not at all clarified	slightly clarified	moderately clarified	substantially clarified	completely clarified

A request for a preference ordering over each subject's fifteen alternatives was utilized to measure the number of preference discriminations he made. This request was included as a single item in a written questionnaire. The text of this request is reproduced below.

> On the second page of this questionnaire you will find a list of fifteen (15) alternative groups of team partners. An alternative consists of four (4) other participants who, along with yourself, could comprise a five-man team. Please rank the fifteen (15) alternatives from most preferrred to least preferred. Place a "1" in front of the most preferred alternative, a "2" in front of the second most preferred alternative, and so on down the line. If you feel no preference between two or more alternatives on the list -- that is, if you feel genuinely indifferent about choosing between them - assign the same number to each. Then assign the next higher number to the next most preferred alternative.

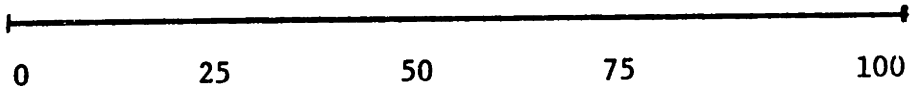
The second page of the referenced questionnaire displayed that subject's fifteen alternative teams, along with marked spaces to the left of each listed team in which he could place a rank number. The number of discriminations he made was then defined as the largest rank number assigned minus one.

A single written questionnaire item was constructed to measure the degree of confidence held by a subject in the accuracy of his indicated preferences (indicated by the rank numbers assigned to each alternative team). The text of this item is reproduced below, along with the confidence scale.

> Please indicate how confident you are that the rank numbers you just assigned to the fifteen (15) groups of team partners accurately represent you current preferences for them. Do this by placing an "X" at the appropriate place along the percentage confidence scale shown below.

NO CONFIDENCE WHATSOEVER

COMPLETE 100 % CONFIDENCE



PERCENTAGE CONFIDENCE SCALE

Five written questionnaire items were constructed to measure the perceived helping effect of each experimental factor. As with the clarification question, a five-point rating scale was defined in all cases. The text of these items is reproduced below.

> Did you find (description of the factor) at all helpful in improving your ability to make a better choice of team partners? Please indicate whether or not and to what extent (description of the factor) has helped you by circling the appropriate point on the scale shown below.

not at all helpful | slightly helpful | moderately helpful | substantially helpful | completely helpful

A single written questionnaire item was constructed to measure a subject's preference for experimentally induced assessment structures. Each subject reflected his current preferences by means of a ranking, and the question item requested that he compare and choose among alternative rankings. The text of this item is reproduced below.

> Now turn to the two ranking sheets. Please indicate which of these more accurately reflects your current preferences for the fifteen alternative groups of team partners by checking the appropriate statement below.
>
> ___ Sheet 1 is a more accurate representation of my current preferences.
>
> ___ Sheet 2 is a more accurate representation of my current preferences.
>
> ___ Both sheet 1 and sheet 2 reflect my current preferences equally accurately.

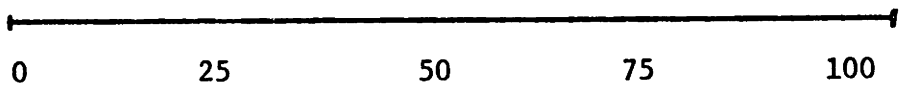
The two referenced sheets contained the subject's own rankings elicited under the influence of different experimental factors. Both sheets were provided for inspection along with this question item.

The extent to which introduction of an experimental factor served to invert rank numbers assigned to alternatives was utilized as a measure of that factor's preference altering effect. Specifically, the percentage of pair-wise comparisons showing a reversed ranking (when compared to a prior ranking elicited under a different experimental factor) was defined as a measure of alteration. The inverse of this measure indicates stability in preferences. In computing the percentage, only those pair-wise comparisons demonstrating a distinct preference order were considered, and instances of indicated indifference were ignored.

Five written questionnaire items were constructed to measure a subject's conscious awareness of changes in preference induced experimentally by each of the five factors. The text of these items is reproduced below.

Due to (description of the factor), you may have reversed some of your preferences for various groups of team partners. That is, you may now prefer one group to another, whereas you may have indicated just the reverse preference for those two groups on the previous ranking. Considering only those pair-wise comparisons on the previous ranking in which you indicated a definite preference (i.e., in which you assigned <u>different</u> rank numbers to the two groups being assessed), in about what percentage of those cases do you think you may have indicated a <u>reverse</u> preference on today's ranking? Please estimate the percentage of reversals by <u>placing an "X"</u> at the appropriate place along the scale shown below.

NO REVERSALS WHATSOEVER — ALL COMPARISONS COMPLETELY REVERSED



<u>PERCENTAGE REVERSAL SCALE</u>

Five written questionnaire items were also constructed to measure the extent to which a subject felt that any gains realized from introduction of each experimental factor were worth his time and effort. The text of these items is reproduced below.

Do you feel that any improvements made in your ability to make a better choice of team partners resulting from (description of the factor) were worth whatever time and effort you may have devoted to this problem? Please indicate whether or not and to what extent you think your time and effort were well spent by <u>checking</u> the appropriate statement below.

___ The improvements gained were worth substantially less than the time and effort expended.

___ The improvements gained were worth moderately less than the time and effort expended.

___ The improvements gained were worth slightly less than the time and effort expended.

___ The improvements gained were worth about the same amount as the time and effort expended.

___ The improvements gained were worth slightly more than the time and effort expended.

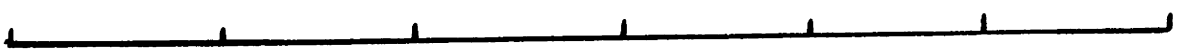
___ The improvements gained were worth moderately more than the time and effort expended.

___ The improvements gained were worth substantially more than the time and effort expended.

It was decided to measure changes in attitudes toward formal assessment procedures by means of a seven-point rating item, also contained in a written questionnaire. Attitude shifts could be measured by changes in individual responses to this item, which was to be answered before and after administration of the entire experiment. The text of this item is reproduced below.

> Please indicate your current attitude toward numerical and quantitative methods of evaluation as a useful tool in decision-making by CIRCLING the appropriate point on the scale

extremely unfavorable | moderately unfavorable | slightly unfavorable | neutral or indifferent | slightly favorable | moderately favorable | extremely favorable



As a measure of the extent to which subjects adopted the actual assessment procedures introduced during the course of the experiment, another written questionnaire item was constructed. This item asked each subject to indicate whether or not and to what extent he had borrowed various assessment procedures and used them in the computer simulation game played during the final week of the training course. The text of this item is reproduced below.

> During the course of our experiment, you were asked to rank alternative teams, to inspect biographical information concerning individual team members, to articulate your criteria for choosing among alternative teams, and, in some cases, to quantify your preferences for each team. Please indicate the extent to which you utilized any of these same evaluation and decision making methods in the course of making decisions during the simulation exercise. Do this by CIRCLING the appropriate point on the scale shown below.

no utlization	slight utilization	moderate utilization	substantial utilization	complete utilization

Three additional measures were constructed which combined the above information in various ways to obtain more refined indices. These were:

1. the number of confident discriminations made on a ranking (i.e., the number of raw discriminations multiplied by the subject's percentage confidence number);

2. the number of chosen confident discriminations made on a ranking (i.e., the number of confident discriminations made on whichever ranking was chosen as a better reflection of the subject's current preferences);

3. a prediction coefficient (i.e., Kendall's Tau coefficient indicating the degree of agreement between a given ranking and the final ranking produced at the moment of final decision).

5.2.6 Introduction Of The Five Experimental Factors, Along With Control Mechanisms

In order to obtain independent and valid measures of the impact of each experimental factor, three separate kinds of control mechanisms were designed into the procedure. These included:

1. using some subjects as their own controls to measure the incremental impact of the five factors as they were introduced sequentially over the course of the experiment;

2. assigning neutral or "placebo" tasks to other subjects to control for "Hawthorne" effects;

3. adopting several randomizing devices to achieve statistical control over other factors which were extraneous to the experiment, but which might otherwise have exerted a systematic influence on the results.

Perhaps the best way to describe the introduction of the five experimental factors, along with these control mechanisms, is by outlining and discussing the complete procedure.

Prior to the beginning of the experiment, all sixty subjects were assigned at random to three groups. Each group contained two of the ten-man sub-groups mentioned in Section 5.2.2, or a total of twenty subjects.

The first group (group I) then performed the following sequence of tasks at the indicated times.

1. 3rd Day
   - Fill out pre-experimental questionnaire concerning prior education/experience/attitudes toward numerical techniques of evaluation
   - Choose preferred list of 6 out of 9 possible team partners
   - Perform 1st ranking of alternatives

2. 12th Day
   - Perform 2nd ranking of alternatives
   - Answer 5 questions about impact of elapsed time with neither information nor guidance
   - Choose 1st or 2nd ranking as currently more applicable

3. 26th Day
   - Receive Raw information about alternatives, but no guidance
   - Perform 3rd ranking of alternatives
   - Answer 5 questions about impact of information
   - Choose currently most applicable ranking

4. 40th Day
   - Receive ordinal guidance
   - Perform 4th ranking of alternatives
   - Compute ranking implied by lexicographic assessment structure
   - Answer 5 questions about impact of ordinal guidance
   - Choose currently most applicable ranking

5. 54th Day - Receive cardinal guidance
   - Perform 5th ranking of alternatives
   - Compute ranking implied by cardinal assessment structure
   - Answer 5 questions about impact of cardinal guidance
   - Choose currently most applicable ranking

6. 68th Day - Make final choice of team partners
   - Perform 6th ranking of alternatives
   - Answer 5 questions about impact of making final choice
   - Indicate sociometric preferences

7. 71st - 75th Day
   - Play computer game

8. 75th Day - Answer questions about methods used in game, and indicate current attitudes toward numerical techniques of evaluation as a useful tool in decision making.

The pre-experimental questionnaire (3rd day) asked each subject to indicate his prior education concerning and practical experience using quantitative techniques of evaluation. These were two of the extraneous factors which the writer did not intend to manipulate experimentally, but which he felt should be controlled statistically by randomly assigning subjects to the three experimental groups. This information was collected to check whether or not randomization had been successful in these two respects. In addition, the writer suspected that many subjects would consider prior education and experience to be relevant assessment criteria in choosing team partners. Consequently, these same data were incorporated in the raw information given to subjects about their fifteen alternative teams.

The pre-experimental questionnaire also asked subjects to indicate their current attitudes toward quantitative evaluation techniques. Responses were used both to measure attitude shifts over the course of the experiment and as additional raw information about prospective team partners.

The preliminary preference questionnaire (3rd day) asked each subject to select six preferred candidates out of nine possible candidates for team partners. With this information, the writer was able to generate each subject's complete list of fifteen alternative groups (i.e., the fifteen logically possible combinations of six candidate partners taken four at a time). Having defined fifteen alternatives for every subject, it was then possible for them to perform their first ranking (3rd day).

Somewhat less than two weeks elapsed without any experimental manipulations. Then, each subject was asked to perform a second ranking of the same fifteen alternatives (12th day). The intervening passage of time without experimental intervention constituted introduction of the no information, no guidance factor. To measure the impact of this factor, a questionnaire was constructed containing the following five items -- referring specifically to the passage of time with neither information nor guidance:

1. clarification item;
2. estimated percentage reversal item;
3. confidence item;
4. helpfulness item;
5. gains realized versus effort expended item.

This five-item questionnaire was administered immediately after the second ranking. A day or two later, subjects were requested (in a separate questionnaire) to compare and choose either their first or their second ranking as more indicative of their current preferences.

Two more weeks elapsed. Subjects were then given printed biographical information about each of the six candidate team partners included in their fifteen alternatives. This constituted introduction of raw information without guidance. After a day to assimilate the biographical information, subjects were requested to perform a third ranking (26th day). This was followed by another five-item questionnaire (referring to the biographical information) and another request to compare and choose among rankings.

Again, two weeks elapsed. The writer then conducted interviews of approximately thirty minutes duration with each subject privately (40th - 45th days). The interviews were semi-directive in style. That is, they were designed to obtain specific kinds of information from each subject, but the subject was permitted to express himself in whatever manner he chose. The interview process was guided by the writer so that, by the end of the thirty minute time period, each subject would have:

1. articulated his major criteria for assessing and selecting a team to play the computer simulation game;

2. mapped out a criterion hierarchy;

3. purged the hierarchy of worth-interdependent members;

4. chosen specific measures (e.g., from the biographical information provided) to interpret his lowest-level criteria;

5. ranked whatever major criteria had been articulated in order of their relative importance;

6. ranked the fifteen alternatives according to the extent to which each satisfied his most important major criterion;

7. applied his second most important major criterion (and lower-ranked criteria, if necessary) to alternatives receiving identical rank numbers on the first criterion so as to break ties and obtain a more complete preference ordering over alternatives.[6]

These interviews were tape-recorded, and the writer subsequently played back the recordings to abstract a written record of each subject's assessment structure.

At the close of the interview, subjects were asked to perform afourth subjective ranking of their alternatives, to fill out the same five-item questionnaire (this time referring to the interview process), and to choose either their fourth subjective ranking or the lexicographic

6. Steps 5, 6, and 7 constituted the technique by which each subject was induced to form a lexicographic assessment structure. The result of this process constituted a lexicographic ordering of his alternatives.

ranking generated during the interview as currently more applicable. A day or two later, subjects were given the usual choice questionnaire asking that they compare their previously chosen ranking with whichever ranking they preferred at the end of the interview. The entire interview process constituted introduction of the ordinal guidance factor.

Two more weeks elapsed, and subjects were interviewed again (54th - 59th days). This second set of interviews took approximately one hour to complete. The purpose, here, was to introduce the cardinal guidance factor (i.e., quantitative portions of the complete assessment procedure developed in Chapter II). The interview process was again guided by the writer in a semi-directive manner so that, by the end of the hour, each subject would have:

1. reviewed the criterion hierarchy and interpretive measures articulated during the first interview:

2. made any additions, deletions, or alterations deemed appropriate thereto;[7]

3. attached numerical weights to all criteria in his hierarchy;

3. attached numerical scores to each alternative on each lowest-level criterion.

As the subject attached weights and scores, the writer made all necessary computations, with the aid of a desk calculator, to arrive at a total worth score for each alternative.

7. The written records abstracted from tape-recordings of the first interview were shown to each subject at the outset of the second interview.

At the close of the interview, subjects were asked to perform a fifth subjective ranking of their alternatives and to fill out another five-item questionnaire (referring this time to the second interview). These operations were performed while the writer was computing total worth scores on the desk calculator. Then, each subject was confronted with his computed total worth scores, converted into equivalent rank numbers, and asked to choose either his fifth subjective ranking or the computed ranking (computed on the basis of total worth scores) as currently more applicable. A day or so later, subjects were given the usual choice questionnaire asking that they compare their previously chosen ranking with whichever ranking they preferred at the end of the second interview. The second interview was also tape-recorded.

Two more weeks elapsed. Subjects then made a final choice of team partners from their list of fifteen alternatives (68th day). Immediately thereafter, they were requested to perform a sixth ranking and to fill out the usual five-item questionnaire concerning the impact of making a final decision. In addition, each subject was requested to indicate sociometric preferences for all of the other nine individuals contained within his ten-man sub-group (including the six previously chosen candidate partners). The purpose, here, was to check whether the random assignment procedure had been successful with respect to the structure of sociometric preferences within ten-man sub-groups. Differential sociometric preference structures constituted another one of the extraneous factors which the writer wished to control statistically.

All subjects participated in the computer simulation game during the next week (71st - 75th days). No experimental manipulations were made during this period. However, on the last day, the writer returned to administer a final battery of questionnaires. Subjects were asked to indicate their current attitudes toward numerical techniques of evaluation. They were also asked whether or not and to what extent they had spontaneously adopted one or more of the assessment methods (introduced during the first and second interviews) for use in playing the computer game. This completed the experiment for group I subjects.

The kernel of the experiment lies in the schedule followed by group I subjects. The five experimental factors were introduced sequentially, with approximately two weeks separating successive introductions. The impact of each factor was measured immediately thereafter. In this manner, it was possible to observe not only the total impact of all five factors, but also the incremental impact of each factor considered separately. In the case of incremental effects, subjects acted as their own controls.

The responses of group I subjects also served to validate the complete assessment procedure developed in Chapter II. When results are presented, analyses will be discussed and conclusions will be drawn from this point of view as well.

Let us now consider the roles played by group II and group III subjects. Recall that all sixty subjects were initially assigned at random to three groups containing twenty men each. The second group (group II) followed a schedule identical to group I through the ordinal guidance interviews (40th day). However, while group I subjects participated in cardinal guidance interviews (54th - 59th days), group II subjects were

given a neutral task to perform. This neutral task was designed to require approximately the same amount of time and effort as the cardinal guidance interview; but the nature of the task was completely unrelated to the problem of choosing team partners. Each group II subject was asked to estimate the ratios of sixty pairs of line lengths presented in a written questionnaire.

The purpose of introducing this neutral or "placebo" task was to control for possible response bias on the part of group I subjects. It was quite conceivable that group I subjects might record favorable reactions to the cardinal guidance interviews (i.e., favorable in terms of confidence, clarification, helpfulness, etc.) not because of the specific assessment methods they were induced to adopt, but just because they had committed a certain amount of time and effort to performing a task which they believed was related to and might help them make a final decision (cf. the "Hawthorne" effect). To control for this possibility, group II subjects were asked to estimate line length ratios. They were then asked to perform a fifth ranking, answer the five-item questionnaire, and choose a preferred ranking just like the group I subjects.[8]

Group III subjects played a similar role with respect to the ordinal guidance interviews. They followed the same schedule as group I

8. The line length ratio estimation task, although unrelated to the experimental decision, was itself designed as a separate experiment. Its purpose was to investigate certain subjective estimation phenomena, and the results obtained will be reported in a separate paper.

and II subjects through the introduction of raw information (26th day). However, while group I and II subjects participated in the ordinal guidance interviews, group III subjects estimated line length ratios. Then, while group I subjects participated in cardinal guidance interviews and group II subjects estimated line length ratios, group III subjects performed a second neutral task. They filled out a psychological questionnaire (The Remote Associates Test) designed to measure individual creativity. Following both neutral tasks, group III subjects went through the usual procedure of ranking alternatives, filling out the five-item questionnaire, and choosing a preferred ranking.

To maintain the fiction that both of these neutral tasks were somehow related to the final decision, the writer gave misleading instructions to group II and group III subjects. In addition, group II and group III subjects were treated in the same manner as group I subjects at final decision time (68th day).

The experiment was originally comtemplated in terms of four randomly assigned groups rather than three. The fourth group would have played a control role similar to the roles played by groups II and III, but this time with respect to raw information. That is, group IV would have received a neutral task, while the other three groups received biographical information. They would also have received neutral tasks during the ordinal and cardinal guidance periods. Unfortunately, however, this would have reduced the size of each group from twenty to fifteen. As a result, statistical comparisons of average group responses on the various experimental impact measures would have been less stable. Prior analysis of hypothetical data suggested that

strong conclusions could not be drawn from the kind of differences which would probably emerge. A trade-off decision was therefore made. The writer decided to sacrifice additional conclusions realizable from a fourth group in favor of stronger conclusions about differences among the remaining three.

5.3.0 IMPLEMENTATION

The purpose of this section is two-fold. First, more detailed information about implementing various design features will be presented than appeared in Section 5.2.6. Second, additional aspects of the procedure not directly relevant to the overall design (and therefore omitted from previous discussion) will be described.

5.3.1 Obtaining Approval To Conduct The Experiment

About four weeks before class exercises began, the writer travelled to Wright-Patterson Air Force Base to obtain approval for his proposed experiment. After several hours of discussion, approval was granted. Furthermore, both military and civilian personnel attached to the school pledged active support of the project. They decided to incorporate the experiment within the official curriculum which, in the writer's opinion, contributed immensely to its successful outcome.[9]

---

9. The writer is indebted to John P. Veasy for originally suggesting the Defense Weapons Systems Management Center at Wright-Patterson as an appropriate place to conduct his experiment. Thanks are also due to Donald C. Marquis for establishing personally the initial contact with officials at the school. Additional thanks are due to Bertrand L. Hansen and to Colonel James H. Schofield for championing the writer's cause during the course of the experiment, as well as for providing encouragement and constructive criticism regarding its implementation.

5.3.2 Laying Out A PERT Chart To Insure Effective Implementation

Following the decision to go ahead with the experiment, a detailed schedule was agreed upon. It became immediately apparent, however, that not all of the scheduled tasks could be performed in direct chronological sequence. Incorporation of the experiment within the official curriculum imposed too many time constraints, and it was difficult to make substantial changes thereto. In addition, all subjects were planning to leave the school immediately after completing the course. This made it practically impossible to recover from oversights in planning or missed deadlines. On the basis of these considerations, an integrated planning and scheduling device seemed essential.

A PERT analysis was performed, and a PERT chart was constructed to guide implementation of the entire experiment.[10] Critical path analysis suggested that several tasks originally scheduled for the end of the experiment should be moved forward in time and performed in parallel with earlier tasks. In particular, it was decided to pre-program all analytical routines before gathering any data and to analyze results under on-line computer control. The impact of this decision will be discussed in Section 5.3.3. It was also decided to pre-test all of the data-gathering instruments well in advance of their actual administration dates. Pre-testing was performed by professors,

10. PERT is an acronym for Program Evaluation and Review Technique.

graduate students, and secretaries at Massachusetts Institute of Technology.

5.3.3 Pre-programming The Analysis

Although the decision to pre-program analytical routines and to analyze results under on-line computer control was originally dictated by scheduling considerations, its impact upon the conduct and successful outcome of the experiment was much more far-reaching than anticipated. A brief discussion of these unanticipated consequences would seem appropriate.[11]

Pre-programming the analysis involved the following operations, most of which were performed prior to gathering any data:

1. deciding how to organize raw data for permanent storage within the computer;

2. coding computer programs to read in the data from an on-line teletype console, to check for illegal observations and obvious typing errors, and to print out an immediate error message diagnosing whatever errors were detected;

3. coding computer programs to form the experimental measures described in Section 5.2.5;

---

11. A detailed discussion of the advantages gained from analyzing these results under on-line computer control will be the subject of a separate paper.

4. formulating specific comparisons of these measures and selecting appropriate statistical tests to shed light on the questions raised in Section 5.2.4;

5. coding computer programs to carry out the above analyses;

6. formulating and coding "exception" analyses of individual data (i.e., analyses designed to establish trends in the responses of individual subjects on various measures and to detect sudden, extreme departures therefrom).

Perhaps the most important consequence of performing these operations before the experiment got underway was their substantial clarifying effect on the writer's concept of what he was really trying to accomplish. Having to plan the entire analytical process down to the level of detail required for an effective computer program pointed up numerous practical problems which might not otherwise have become apparent until after it was too late to do anything about them. This, in turn, permitted redesigning the experimental procedure in advance to remedy the specific problems which turned up. Several of the question items discussed in Section 5.2.5 were reworded on the basis of this exercise. In particular, the clarification, helpfulness and gains realized versus time and effort expended items were altered to obtain measures of impact (a dynamic concept). These changes were required to permit valid time series comparisons of individual observation values. More will be said about this later. In addition, several steps in the procedure outlined in Section 5.2.6 were resequenced for similar reasons.

A second consequence was the design trade-off decision between three and four experimental groups discussed in Section 5.2.6.

A third advantage was realized from the pre-programmed "exception" analyses. Five of the twenty group I subjects recorded confidence scores on the sixth ranking which departed substantially from the previous trend each had established. Upon questioning these five subjects, it turned out that the deviant scores indicated no radical shift in confidence at all, but rather an unanticipated change in the way each subject had interpreted the question. Fortunately, this confusion was cleared up on the spot.

A fourth advantage was realized from the ability to provide immediate feedback to the subjects at the end of the experiment. On the morning of the last day of classes major findings and conclusions drawn from the experiment were reported. Since it was announced in advance that a complete report would be made, and judging from the great interest shown by the subjects in these results (over ninety percent of the subjects arose an hour early that morning to hear the report), the prospect of immediate feedback would seem to have motivated subjects to participate more fully in the experiment. Other results, to be reported shortly, provide further evidence in support of this conclusion.

5.3.4 Assigning Subjects To the Three Experimental Groups

This experiment was originally designed for exactly sixty subjects. However, previous classes had ranged in size from fifty-six to sixty-four. By a stroke of good fortune, exactly sixty subjects arrived on the first day of classes, which made the random assignment procedure easy to implement, as well as assuring three groups of equal size (an advantage for purposes of statistical analysis).

A random number table was used to assign the sixty subjects to groups I, II, and III. The list of sixty names was arranged alphabetically and then sampled without replacement in such a way that all possible tri-partite divisions were equally likely. By a similar process, each of the three experimental groups were subdivided into two ten-man sub-groups. All of these operations were performed prior to the beginning of the experiment.

5.3.5 Introducing The Experiment To The Subjects

On the third day of classes, school officials introduced the writer to the students and announced that an experiment would be performed as a part of the curriculum. The writer then spoke for approximately thirty minutes. An outline of the topics covered in this introductory address is reproduced below:

1. introduction of the writer and the various organizations sponsoring the experiment;

2. statement of overall purpose;

3. description of the computer game;

4. description of the decision to be made in choosing a team to play the game;

5. outline of the experimental procedure;

6. justification of their participation;

7. request that subjects not discuss the experiment among themselves;

8. guarantee of anonymity and confidentiality with respect to their responses;

9. assurance that their performance in the experiment would not be graded by the school;

10. denial of any time limits associated with performing experimental tasks;

11. explanation of the three experimental devices, including the honoring of final choices;

12. request for cooperation and participation, even though the specific purpose of each experimental manipulation would not be disclosed until the experiment was over;

13. promise to disclose everything at the end, along with a complete report of results and conclusions;

14. additional administrative details;

15. advance thanks for their cooperation.

5.3.6 Obtaining Pre-Experimental Measures

Directly following the introductory address, the first battery of questionnaires was passed out. Questionnaires were contained in a sealed manila envelope with the subject's name printed on the outside. The purpose of the envelopes was to maintain an aura of confidentiality, as promised in the introductory address, as well as to assure proper identification of responses with the subject who gave them. These same envelopes were used throughout the course of the experiment.

The first battery of questionnaires included the pre-experimental questionnaire and the preliminary preference questionnaire. Reproductions of the actual instruments used appear in Appendices XXI and XXII at the end of this thesis.

5.3.7 Implementing The First Ranking

Using each subject's six preferred candidates as indicated on the preliminary preference questionnaire it was possible to construct a complete set of fifteen alternative groups of team partners. All sixty sets of alternatives were constructed and typed with the generous assistance of secretarial personnel provided by the school.[12] A sample sheet of alternatives appears in Appendix XXIII, using ficticious names.

---

12. The writer is indebted to Mrs. Jean Vogel for her assistance in this and many respects.

Each subject's list of fifteen alternatives was then appended to the standard ranking questionnaire. These were placed in the same manila envelopes and passed out during the third and fourth days. A reproduction of the standard ranking questionnaire appears in Appendix XXIV.

5.3.8 Implementing The Second Ranking

Data gathered from the pre-experimental questionnaire and the first ranking were typed directly into the computer and edited on-line.[13] In addition, all subsequent ranking sheets (i.e., for rankings 2 through 6) were created at that time and stored for later use. A few words about this latter procedure would seem appropriate.

It occurred to the writer that subjects might display a certain kind of response bias while assigning rank numbers to the fifteen alternative teams. Specifically, they might be influenced by the vertical ordering of alternatives on a ranking sheet and, perhaps, by the horizontal ordering of names within each alternative. To control for these possibilities, a computer program was written to randomize both the vertical and horizontal ordering of names on successive ranking sheets. By this device, any systematic position bias would be washed out of the analysis. Only systematic preferences for the alternatives (which is what the rankings were supposed to indicate) would remain.

---

13. The computer used was Project MAC's on-line, time-shared system. This same procedure of entering and editing data was performed after every ranking. Intermediate analyses were also performed simultaneously.

Naturally, a precise record had to be maintained of the position of each alternative on successive ranking sheets so that the results could be unscrambled prior to analysis. This was accomplished by creating a tape record of each sequence of alternatives and reading this tape into the analytical routines along with the actual ranking data.

On the twelfth day, the second ranking was performed. Subjects were then given a five-item questionnaire concerning the recent passage of time with neither information nor guidance. The next day, they were given the ranking choice questionnaire. The five-item questionnaire has been reproduced in Appendix XXV. The ranking choice questionnaire will be found in Appendix XXVI.

5.3.9 Implementing The Third Ranking

The school was kind enough to prepare background information on each of the sixty subjects participating in the training course.[14] This information, along with each individual's prior education, experience, and attitudes concerning numerical evaluation techniques was typed up on a single sheet. A sufficient number of Xerox copies were made of these biographical sketches so that each subject would have information on every one of his six candidate team partners.

14. The writer is indebted to Lt. Colonel Phillips for his cooperation in this regard.

Biographical information was given out on the evening of the twenty-fifth day, along with appropriate instructions. On the twenty-sixth day, subjects performed their third ranking and answered a five-item questionnaire about the impact of the information received. A second preference choice questionnaire was administered on the next day. A sample sheet of biographical information has been reproduced in Appendix XXVII. The five-item questionnaire referring to raw information without guidance has been reproduced in Appendix XXVIII.

5.3.10 Implementing The Fourth Ranking

By the fortieth day, it was time to administer ordinal guidance interviews to groups I and II and the first neutral task to group III. A series of forty interviews lasting approximately thirty minutes each was conducted in the writer's quarters. A beverage of the subject's choice was offerred at the beginning of each interview, and a brief warm-up discussion followed. Then the interview began in earnest.

All interviews were conducted between the hours of 4:30 and 11:30 P.M., with an hour out for supper. None of the subjects objected to having the interviews tape-recorded.

The line length ratio estimation exercise was handed out in the usual manila envelopes to group III subjects. They were free to complete this exercise at their convenience. The instruction page of this first neutral task has been reproduced in Appendix XXIX.

During the course of the ordinal guidance interviews, a number of interesting facts emerged. First, many subjects admitted that they had not initially taken a great deal of interest in the experimental decision, nor had they devoted much thought to ranking alternatives and answering questions. The major reasons given were that they had not gotten to know many of their prospective team partners personally, and that the decision seemed too distant in time and not sufficiently important to warrant a great deal of attention. A few informal interviews with group III subjects turned up the same results. Fortunately, however, subjects in all three groups reported a major change in this attitude following receipt of the biographical information. Their subsequent record of interest and cooperation confirmed that a distinct and permanent change in attitude had occurred at that time.

A second fact emerged concerning their discussions of the experiment with each other. Many subjects admitted to discussing the experiment in general, but none claimed to have divulged their specific responses. The major topic they did discuss was the intent of the various procedures. Since the writer had purposefully remained silent on this issue, the subjects were naturally quite curious. A certain amount of speculation as to the possible intent of the procedures had apparently taken place.

At the close of the interview, group I and II subjects were asked to perform a fourth ranking and to answer a five-item questionnaire -- this time referring to the impact of the interview process. Group III subjects were also asked to perform a fourth ranking and to answer a five-item questionnaire concerning the ratio estimation task. Several days later, all subjects were given another ranking choice questionnaire. The five-item questionnaire referring to the interview process has been reproduced in Appendix XXX. The one referring to the ratio estimation task has been reproduced in Appendix XXXI.

5.3.11 Implementing The Fifth Ranking

Cardinal guidance interviews were conducted with group I subjects between the fifty-fourth and fifty-ninth days. Simultaneously, group II subjects received the line length ratio estimation task, and group III subjects received a second neutral task -- The Remote Associates Test of Creativity.[15]

Cardinal guidance interviews lasted for approximately one hour. They were conducted in the same manner as the ordinal guidance interviews. The two neutral tasks required between thirty minutes and an hour to complete. All subjects then performed a fifth ranking, answered another five-item questionnaire, and later filled out another ranking choice questionnaire. The five-item questionnaire referring to the

15. Authors of the Remote Associates Test have requested that it not be published.

Remote Associates Test appears in Appendix XXXII.

Inspection of the various questionnaires reproduced in Appendices XXI through XXXVI will show that many of them contain both discrete and continuous rating scales. For purposes of analysis, it is important to know whether or not subjects perceived these scales as "linear". That is, did they perceive adjacent scale categories as "equally spaced" in terms of the attribute being measured? In the case of continuous scales, did they impute the same interval significance to line segments of equal length over the entire range of the scale? Only if answers to these questions are yes can their responses be treated and analyzed as genuine interval data. To determine whether or not subjects did perceive these scales as "linear", the writer asked the above questions during the cardinal guidance interviews. The first ten subjects all said yes, so the issue was happily dropped.

5.3.12 Implementing The Final Choice

Each subject made his final choice on the sixty-eighth day. Immediately thereafter, subjects performed a sixth ranking and filled out a five-item questionnaire referring to the act of making that choice. Included in the same manila envelope was a sociometric preference questionnaire designed to obtain information about mutual likes and dislikes within ten-man sub-groups. This latter questionnaire was filled out last. A reproduction of the final choice questionnaire will be found in Appendix XXXIII. The five-item questionnaire appears in Appendix XXXIV, and the sociometric preference questionnaire appears in Appendix XXXV.

5.3.13 Obtaining Post-Experimental Measures

All subjects participated in the computer simulation game between the seventy-first and seventy-fifth days. On the seventy-fifth day, a post-experimental questionnaire was passed out containing two final items. The first item asked subjects to indicate the extent to which they had utilized assessment methods introduced throughout the experiment in making decisions during the computer game. The second item asked for their current attitudes toward quantitative techniques of evaluation. A reproduction of the post-experimental questionnaire will be found in Appendix XXXVI.

5.3.14 Explaining The Experiment, Presenting Results, And Drawing Conclusions

Immediately following the post-experimental questionnaire, the writer spoke for approximately forty-five minutes about the experiment and its results. The complete design was divulged, and the purpose of each experimental procedure was discussed. Since all analyses had already been performed (except in the case of the two post-experimental questionnaire items), the writer was able to present results and state conclusions. This information was quite favorably received by the subjects, especially since it cleared up what had been for many a "great mystery". Substantial interest was also expressed in the conclusions drawn. A short, but animated discussion period followed. Subjects then received diplomas from the school and returned to their homes.

5.3.15 Epilogue

A few concluding remarks about the level of commitment and participation demonstrated by the subjects would seem appropriate.

There were a number of strong indications that almost all subjects took the experiment quite seriously. This was particularly true following receipt of the biographical information on the twenty-fifth day. Some of the evidence supporting this conclusion is outlined below.

1. Comments to this effect offered by various subjects during their interviews have already been reported.

2. The attendance record and interest demonstrated during the final discussion of results has also been reported.

3. Every subject voluntarily committed approximately four and one-half hours to performing the various experimental tasks. Some committed up to six hours.

4. Slightly over fourteen thousand pieces of data were gathered during the course of the experiment. Of these, the writer was unable to obtain only four. During the closing days of the final week, one subject was stricken with a mild heart attack, and another was called home for an illness in the family. Hence, their responses to the post-experimental questionnaire were unavailable.

5. The writer was made an honorary member of the class and presented with "The Mesmer Award for Mastery in the Art of Confusion" during a sham ceremony about three-quarters of the way through the experiment.

However, there were a few exceptions. Two subjects announced at the outset of the experiment that they had no intention of committing a great deal of thought and effort to making such a trivial decision. They promised nominal cooperation in performing various tasks, but assured the writer that their behavior would be nothing more than ritualistic. They kept their word. A third subject made no such declaration of intention at the outset, but his responses clearly demonstrated the same pattern. All three of these subjects were therefore deleted from the analysis.

5.4.0 RESULTS

Now comes the payoff -- the actual results obtained from the experiment. Basically, there are two kinds of results to be investigated. First, we shall assess the extent to which randomly assigning subjects to the three experimental groups did succeed in controlling for the various extraneous factors previously discussed. In addition, we shall see how effective randomization was in balancing the composition of the three groups in certain other ways. We shall also gain a more detailed picture of the subjects themselves.

Second, we shall investigate the impact of the five experimental factors in terms of the various measures previously defined. This will be accomplished on a measure-by-measure basis.

Section 5.4.1 will discuss the effectiveness of randomization. Section 5.4.2 will discuss results on each of the experimental measures. In Section 5.5, all results will be interpreted both in terms of the five experimental factors and in terms of the assessment procedure, integrated conclusions will be drawn, and several suggestions for further research will be presented.

5.4.1 The Effectiveness Of Randomization

Recall from Section 5.3 that four extraneous factors were singled our as possibly exerting a meddlesome influence upon experimental results. These were:

1. prior education in formal, quantitative techniques of analysis and evaluation;

2. prior practical experience using such techniques;

3. prior attitudes toward such techniques;

4. the structure of sociometric preferences within ten-man sub-groups.

By randomly assigning subjects to the three experimental groups, it was hoped to homogenize them in these respects and thereby neutralize any such meddlesome influences.

Table 1 below displays the mean scores of each group on the first three extraneous factors. Prior education is measured in number of semester-length courses taken (self-rated). Prior experience is measured in number of years (self-rated). Prior attitudes are measured on a seven-point rating scale linearly transformed to fall between minus one and plus one. Minus one indicates extremely unfavorable attitudes, zero indicates indifference, and plus one indicates extremely favorable attitudes (self-rated). The standard deviation associated with the entire sample of sixty subjects is presented along with each set of group means. This is to provide a basis for interpreting the magnitude of mean differences.

Table I

Mean Group Scores On
Three Extraneous Factors

FACTOR	Group I	Group II	Group III	Std. Dev.
Prior Education	.500	.600	.500	1.204
Prior Experience	.957	2.300	3.013	3.667
Prior Attitudes	.417	.500	.533	.432

Inspection of Table 1 shows that, in all three cases, the maximum difference is substantially less than the overall sample standard deviation. In the closest case (prior experience), the maximum mean difference (2.056) is still only fifty-six percent as

large as the associated standard deviation (3.667). These results suggest that randomization was successful in homogenizing the three groups with respect to all of these extraneous factors.

A similar picture emerges regarding the structure of sociometric preferences within ten-man sub-groups. Table 2 shows the relative frequencies of responses given on each response category of the sociometric preference question. Only responses relating to each subject's six candidate team-partners are considered, since they constituted the only possible choices. Also, the three subjects who did not participate fully in the experiment have been excluded to remain compatible with subsequent analyses of effects (as opposed to prior conditions) of the experiment.

Table 2

Relative Response Frequencies
on the Sociometric Preference Question

Response Category	Group I	Group II	Group III
Extreme Aversion	.009	.000	.028
Moderate Aversion	.037	.017	.028
Slight Aversion	.037	.058	.046
Indifference	.148	.100	.074
Slight Preference	.250	.158	.148
Moderate Preference	.287	.300	.306
Extreme Preference	.232	.367	.370

Table 3 shows the mean educational level achieved by subjects in the three experimental groups. The unit is number of years of formal education (self-rated). Also shown is the standard deviation associated with the entire sample of sixty.

Table 3

Mean Educational Level Achieved

FACTOR	Group I	Group II	Group III	Std. Dev.
Education Achieved	16.550	17.050	16.600	1.063

Once again, the maximum mean difference is swamped by the sample standard deviation, indicating reasonable homogeneity.

Tables 4 and 5 show similar results with respect to service affiliation and military rank (or government service grade). These tables contain compositional relative frequencies and should be self-explanatory. The only important departure from homogeneity occurs with respect to service affiliation. Table 4 shows a disproportionate concentration of Army officers in group III. However, the writer was unable to attribute any significance to this fact.

Table 4

Relative Composition
By Service Affiliation

Affiliation	Group I	Group II	Group III
Army	.200	.200	.500
Navy	.350	.350	.350
Air Force	.300	.300	.100
Non-military	.150	.150	.050

Table 5

Relative Composition
By Rank or Grade

Rank/Grade	Group I	Group II	Group III
Maj., Lt. Cmndr., or GS-13	.150	.000	.100
Lt. Col., Cmndr., or GS-14	.550	.650	.650
Col., Capt., or GS-15	.300	.350	.250

In summary, randomization appears to have been quite effective in terms of the four extraneous factors. It was also effective in terms of education achieved and rank or grade. It was not completely effective with respect to service affiliation, but this may not be important.

5.4.2 Results On Specific Measures

In this section we shall investigate experimental results on each of the measures previously defined. However, before launching into a detailed discussion, it would be well to describe the overall strategy which our investigation will follow.

Recall from the design of the experiment that all subjects were treated identically through introduction of raw information and performance of the third ranking (26th day). Following that, however, group III subjects, and then group II and group III subjects received neutral tasks, while group I subjects underwent the complete assessment procedure developed in Chapter II. These facts, coupled with the random assignment procedure carried out initially, suggest the following strategy of analysis.

1. Expect no significant differences between groups on any measures up through the introduction of raw information and the third ranking.[16] However, check to see

16. A single exception to this statement will occur when we discuss the prediction coefficient in Section 5.4.2.7.

if any significant differences did emerge, and be prepared to interpret subsequent results accordingly.

2. Expect differences between group III subjects and the combined sample of groups I and II to emerge following introduction of ordinal guidance and the fourth ranking. Analyze results accordingly.

3. Expect further differences between group I and group II subjects to emerge following introduction of cardinal guidance and the fifth ranking.

4. Look for cumulative effects of differential treatment at final decision time and the sixth ranking.

To implement this strategy, the following specific conventions will be adopted.

1. Unless otherwise noted, results obtained through the third ranking will be reported as total sample means. All three experimental groups will be lumped together, and total sample means will be displayed both numerically and graphically for each of the three time periods.

2. However, to permit visual inspection, individual group means associated with each of these early time periods will be displayed separately, and whatever inter-group differences did emerge will be tested statistically to validate the assumption of homogeneity.

3. The homogeneity assumption will be tested by a one-way analysis of variance (fixed effects model) applied to inter-group differences in mean responses. The result of this test will be encapsulated in a single conditional probability number indicating the likelihood that differences at least as large in any direction could have occurred by sheer chance, assuming that the three groups were really homogeneous. Naturally, the larger the probability number, the more credence we may attach to the homogeneity assumption.

4. The above conditional probability number will be referred to as the significance level achieved by observed mean differences.

5. Group III results will be compared with the combined results of groups I and II on the fourth ranking. As before, a one-way analysis of variance will be performed to validate the homogeneity of groups I and II. In addition, differences between group III results and the combined results of groups I and II will be tested by a T-test (of mean differences) to detect any systematic effects of their differential experimental treatment.

6. Results from all three groups will be displayed and analyzed separately on the fifth and sixth rankings. Inter-group comparisons will be made as appropriate.

7. In addition to the above analyses, which refer to comparisons between groups within a given time period, time series comparisons will also be made within each group. That is, average measures of change between time periods will be constructed for each group. The significance of such changes will be assessed by means of a T-test (assuming no real change), and results will again be summarized by means of a significance level.

8. Significance tests will sometimes be interpreted directionally, at which time one-tail significance levels will be reported. When this happens, it will not be because prior directional hypotheses were formulated (recall that no formal hypotheses of any kind were formulated), but rather because a striking directional pattern is obvious from the data, and, therefore, it would be silly to interpret results as if the pattern did not exist.[17]

With this overall strategy in mind, let us now proceed to the results themselves.

17. The writer is indebted to Merrill M. Flood for his counsel on the presentation and legitimate interpretation of statistical test results.

5.4.2.1 Raw Discriminations

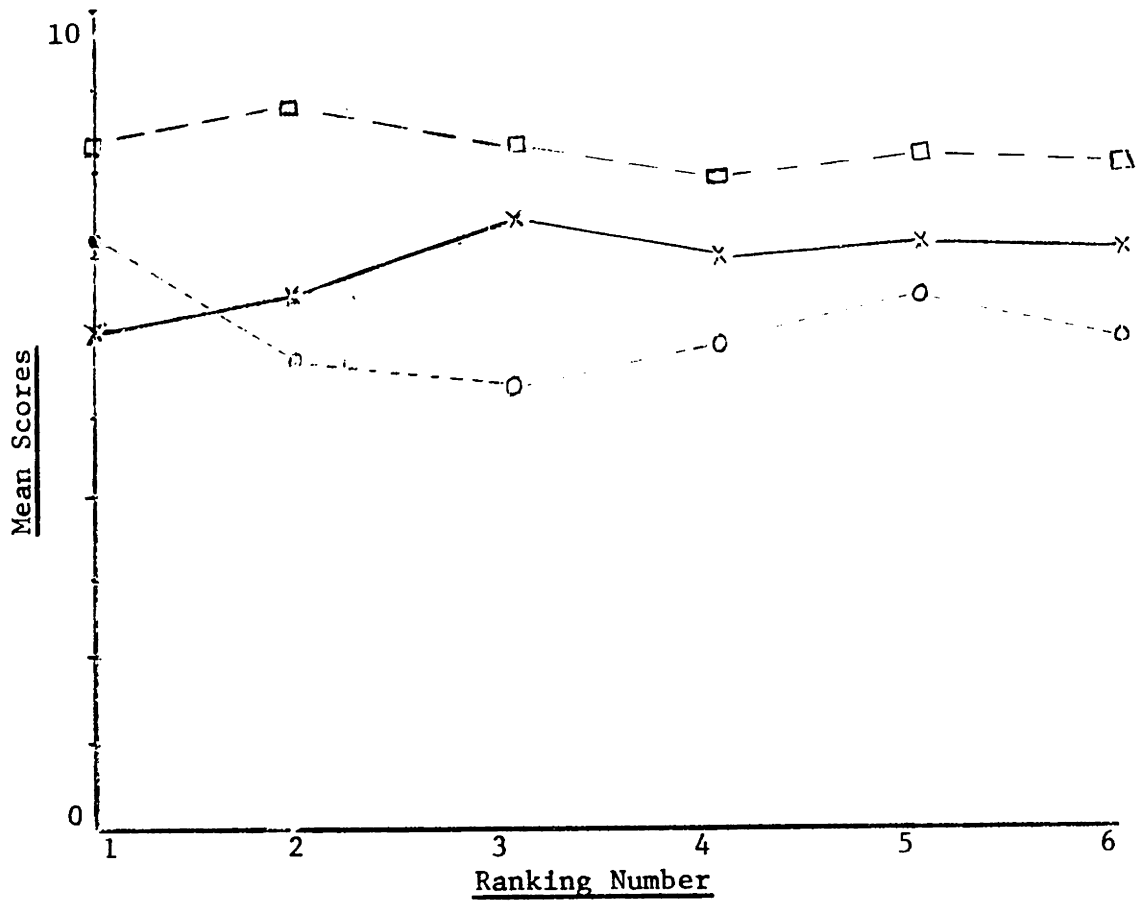
The most striking result obtained from analyzing the number of raw preference discriminations made on various subjective rankings is the apparent lack of experimental impact thereupon -- at least in terms of means. There occurred neither systematic differences between groups at any given time, nor were there any systematic differences within groups over separate time periods. In light of these results, we shall depart from the usual procedure of lumping groups together in the earlier time periods and display mean results for each group separately on each successive ranking. This is done in Figure 1.

Inspection of Figure 1 suggests a slight tendency for group III subjects to make consistently more discriminations then group I subjects and for group I subjects to make more than group II subjects. However, six one-way analyses of variance were performed (one on each ranking), and none of them yielded statistically significant results. The most significant difference occurred on the third ranking (significance level = .140), and the least significant difference occurred on the fifth ranking (significance level = .538).

Time series comparisons were equally inconclusive. Fifteen T-tests performed on contiguous time shifts yielded significance levels ranging from .085 (group II from fourth to fifth ranking) to .777 (group I from fifth to sixth ranking).

Figure 1

Mean Group Scores
Number of Raw Discriminations



X = Group I alone
O = Group II alone
□ = Group III alone

Mean Scores

Group Number	Ranking Number					
	1	2	3	4	5	6
I	6.000	6.526	7.947	7.053	7.263	7.158
II	7.100	5.750	5.450	5.950	6.600	6.150
III	8.444	8.889	8.333	8.000	8.389	8.222
TOTAL	7.156	7.000	7.193	6.965	7.386	7.140

From the above results, we must conclude that nothing interesthing occurred with respect to mean discriminations. However, an informal comment would seem appropriate. During the course of the experiment, the writer was aware of systematic trends occuring in the number of raw discriminations made over successive rankings by individual subjects. The problem is that some individuals displayed increasing trends, while other individuals displayed decreasing trends. Apparently these balanced each other out, since no consistent trends occurred in group means. It would be interesting to analyze these (and other) data from an individual point of view. Unfortunately, however, such analyses lie outside the immediate scope of this thesis.[18]

5.4.2.2 Confidence

In contrast to raw discriminations, the confidence measure (and all subsequent measures) did generate interesting results. Figure 2 displays these results in the standard format.

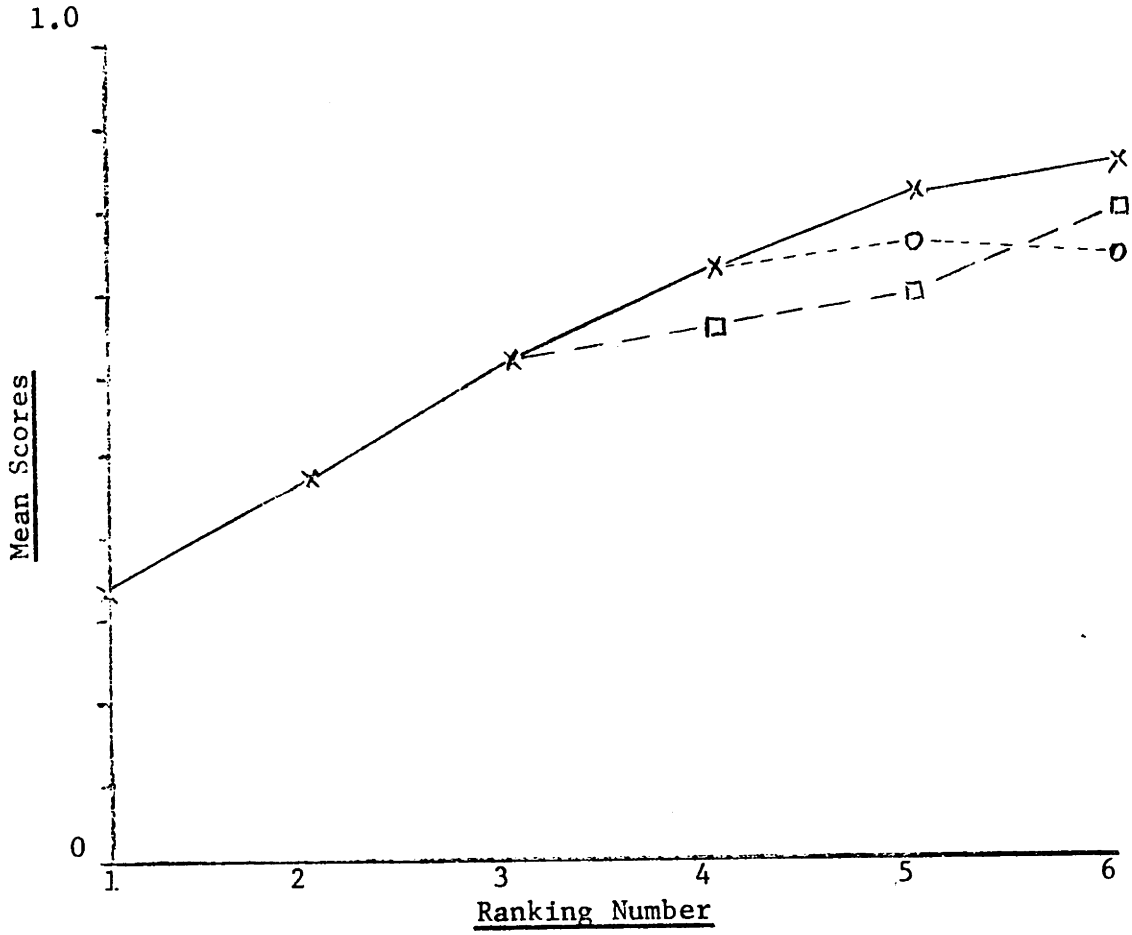
Inspection of Figure 2 shows a consistent upward trend in mean percentage confidence scores. This was true of all groups in all periods, except one. Group II subjects suffered a slight downward shift at final decision time (sixth ranking).

As expected, no significant differences emerged between any

18. The writer is indebted both to Donald G. Marquis and to Warren G. Bennis for pointing out the potential importance of analyzing results from an individual point of view.

Figure 2

Mean Group Scores
Percentage Confidence



X = Group I alone or combined
O = Group II alone
□ = Group III alone

Mean Scores

Group Number	Ranking Number					
	1	2	3	4	5	6
I	.245	.491	.644	.745	.822	.851
II	.415	.452	.595	.706	.764	.746
III	.343	.491	.625	.665	.698	.800
TOTAL	.336	.477	.621	.706	.762	.798

groups on the first three rankings. Three one-way analyses of variance were performed, generating significance levels of .156, .894, and .834, respectively.

In contrast, the upward trend of all three groups lumped together was highly significant. A T-test of the shift from the first to the second ranking (spanning the no information, no guidance treatment) was significant at .0006 (1-tail). A similar test on the shift from the second to the third ranking (directly following raw information) was significant at .001 (1-tail).

At the time of the fourth ranking (when groups I and II received ordinal guidance, while group III received their first neutral task), between-group differences began to emerge. Groups I and II taken together enjoyed a further increase in confidence significant at .0007 (1-tail), while group III registered only a slight increase significant at .242 (1-tail).

A comparison of mean differences between groups I and II on the fourth ranking yielded only a negligible spread significant at .532. On the other hand, lumping groups I and II together and comparing the combined sample with group III yielded a suggestive difference significant at .164 (1-tail). This suggestive difference widened to a clear difference during the next period.

At the time of the fifth ranking (when group I received cardinal guidance, while both group II and group III performed neutral tasks), between-group differences became even more pronounced. Group I enjoyed another noticeable increase significant at .074 (1-tail), while group III registered only a slight change significant at .236 (1-tail). However, contrary to what might have been expected on the basis of group

III's previous behavior under a neutral task, group II also enjoyed a noticeable increase significant at .050 (1-tail). A T-test comparing group I and group III showed a difference significant at .009 (1-tail), and a similar test comparing group I and group II was significant at .109 (1-tail).

Some interesting things occurred at final decision time. Whereas group I enjoyed another noticeable increase significant at .083 (1-tail), group II sufferred a slight but insignificant decrease. Group III, on the other hand, registered a dramatic increase significant at .013 (1-tail), which moved group III into second place, only slightly behind group I. The final difference between group I and group III became insignificant, but the difference between groups I and II achieved a significance of .060 (1-tail).

The above results, along with all subsequent results, will be interpreted in Section 5.5.2.

5.4.2.3 Confident Discriminations

The reader can undoubtedly guess how the pattern of confident discriminations turned out. Recall that this measure was constructed by multiplying each subject's number of raw discriminations on a ranking by the percentage confidence he placed in the accuracy of the rank numbers assigned. Recall also that the pattern of raw discriminations displayed nothing but random noise (see Figure 1), while the pattern of confidence scores displayed several systematic trends (see Figure 2).

Consequently, one could expect the pattern of confident discriminations to look very much like the confidence scores, but with some extra noise contributed by random fluctuations in the number of raw discriminations. This is exactly what happened (see Figure 3).

Since Figure 3 is a noisy replica of Figure 2, a detailed discussion will be omitted. Let it suffice to say that the pattern of differences and their significance levels were very similar to these associated with confidence scores, although significance levels were in general less conclusive due to the extra noise.

5.4.2.4 Chosen Confident Discriminations

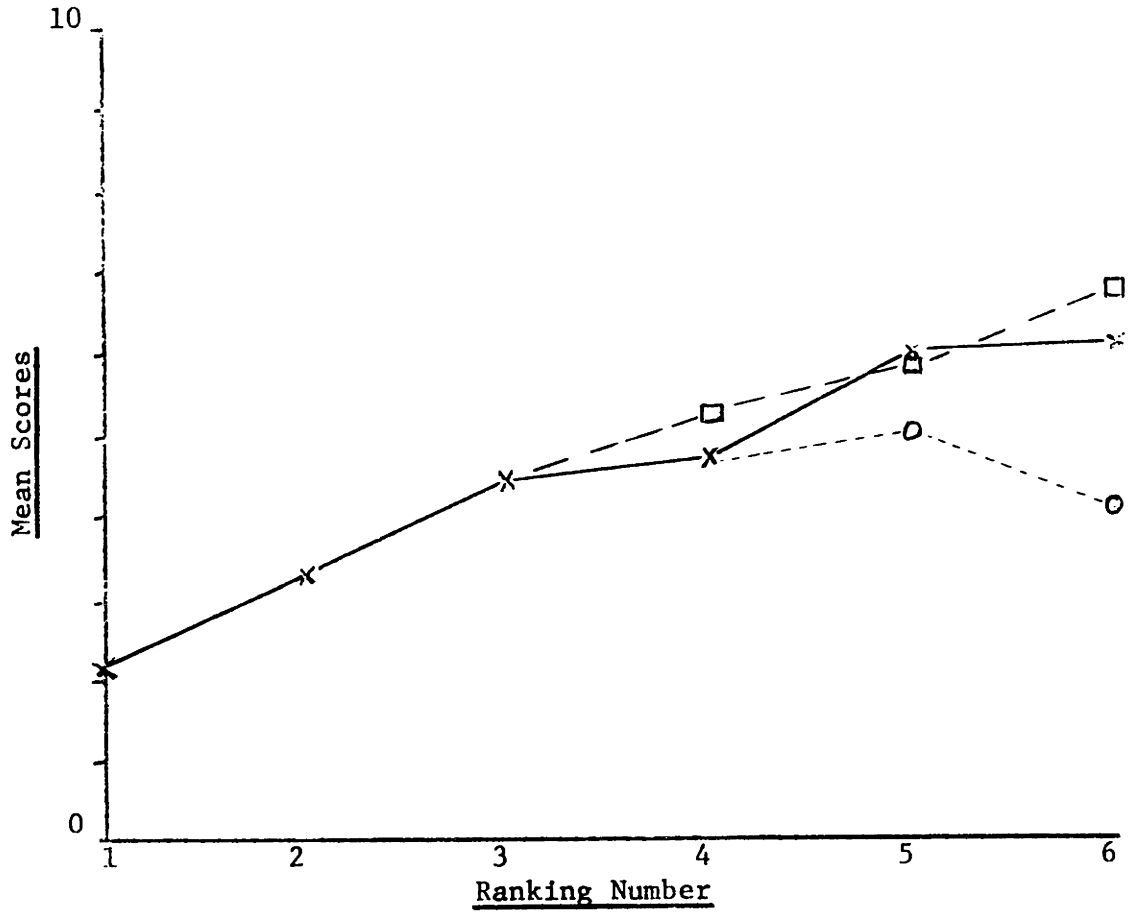
The pattern of chosen confident discriminations was once again similar both to confidence scores and to the number of confident discriminations. This can be seen from Figure 4.[19]

For the same reasons previously given, a detailed discussion of chosen confident discriminations will be omitted. However, one important result deserves special attention. Notice the dramatic upward shift displayed by group I subjects on the fifth ranking following cardinal guidance. This shift was significant at .057 (1-tail), and the resulting difference between group I and group II was significant at .021 (1-tail). The reason for this result will

19. Inspection of Figure 4 shows no entries associated with either the first or the sixth rankings. This is because no preference choice questions were asked at either of these times.

Figure 3

Mean Group Scores
Number of Confident Discriminations



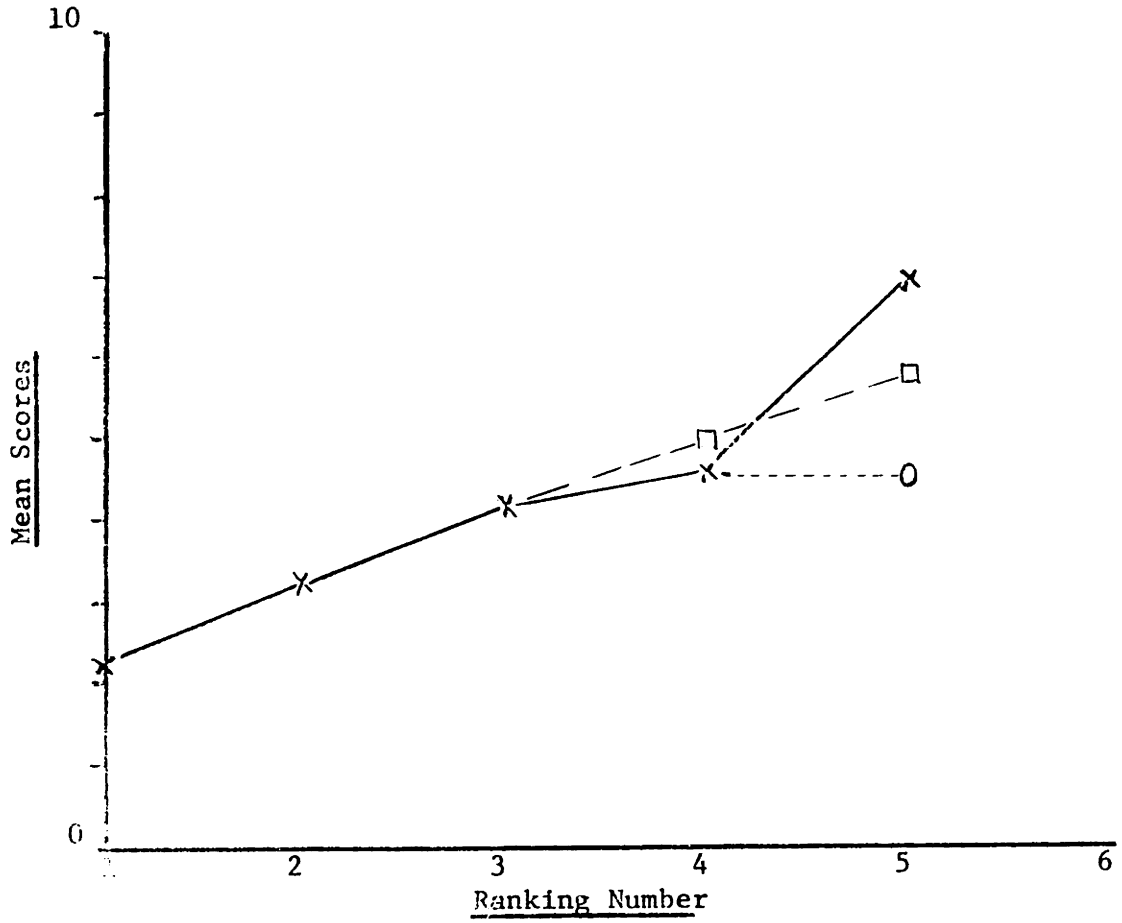
X = Group I alone or combined
O = Group II alone
□ = Group III alone

Mean Scores

Group Number	Ranking Number					
	1	2	3	4	5	6
I	1.763	3.404	5.113	5.166	6.077	6.058
II	2.342	2.283	3.150	4.255	5.006	4.107
III	2.271	4.235	5.090	5.336	5.958	6.796
TOTAL	2.126	3.273	4.417	4.900	5.664	5.607

Figure 4

Mean Group Scores
Number of Chosen Confident Discriminations



X = Group I alone or combined
O = Group II alone
□ = Group III alone

Mean Scores

Group Number	Ranking Number					
	1	2	3	4	5	6
I	-	3.323	5.067	5.108	6.945	-
II	-	3.018	2.937	4.025	4.588	-
III	-	3.342	4.573	5.053	5.801	-
TOTAL	-	3.222	4.164	4.711	5.757	-

become apparent in Section 5.4.2.11, wherein we shall investigate the way subjects chose among alternative rankings as better reflecting their preferences.

5.4.2.5 Actual Reversals

A clear pattern also developed regarding the percentage of reversals in preference orderings induced by successive experimental manipulations. This pattern is apparent in Figure 5, again organized according to the standard format.

Inspection of Figure 5 shows a consistent downward trend in mean percentage of actual reversals. This was true of practically every group on almost every ranking. As will be demonstrated shortly, the exceptions to this consistent downward trend were infrequent, small in magnitude, and statistically insignificant.

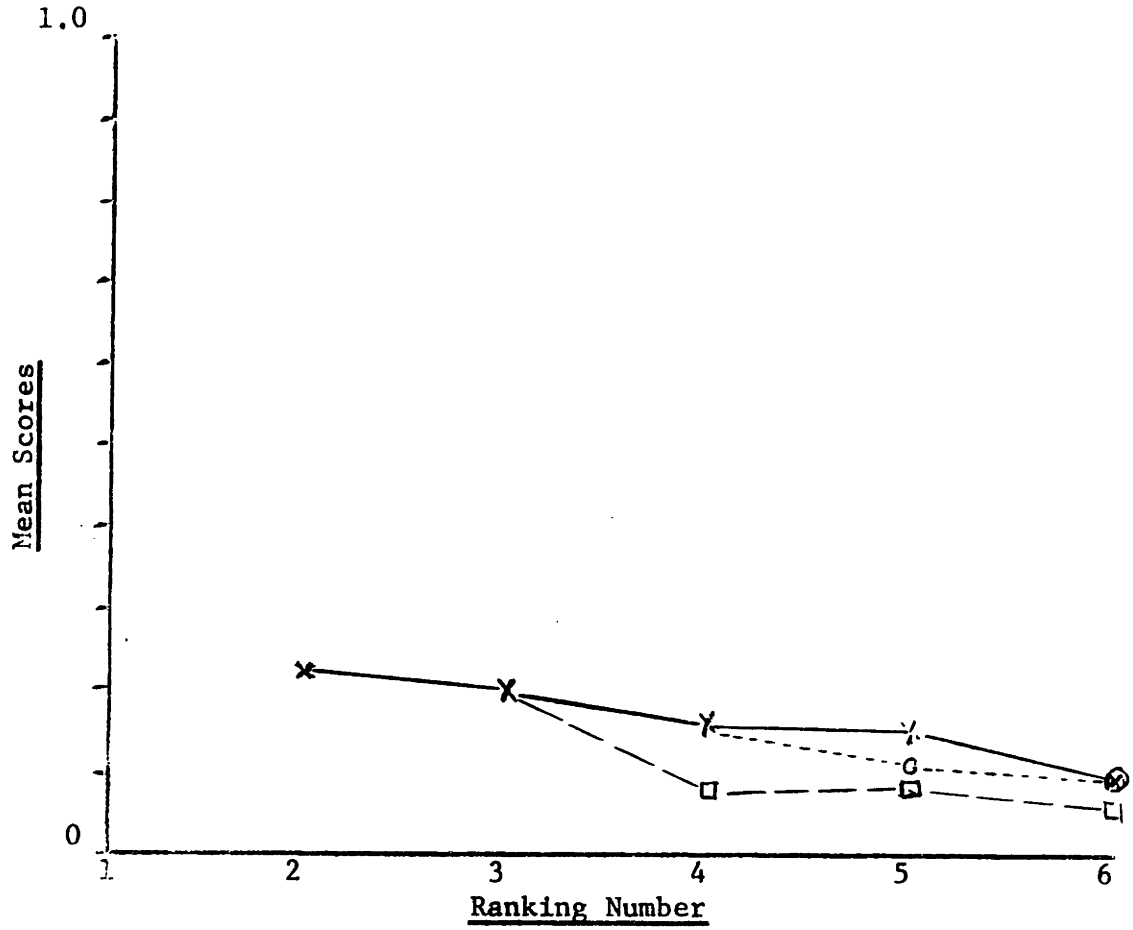
As expected, no significant differences emerged between any groups on the second and third rankings.[20] Two one-way analyses of variance were performed, generating significance levels of .399 and .969, respectively. Similarly, no significant differences emerged between groups I and II on the fourth ranking, as indicated by a significance level of .481.

At the time of the fourth ranking, anticipated differences

20. There were no reversals associated with the first ranking, since there were no previous rankings to serve as a basis of comparison.

Figure 5

Mean Group Scores
Actual Percentage of Reversals



X = Group I alone or combined
O = Group II alone
□ = Group III alone

Mean Scores

Group Number	Ranking Number					
	1	2	3	4	5	6
I	-	.186	.215	.146	.158	.098
II	-	.226	.205	.183	.122	.105
III	-	.280	.218	.081	.089	.072
TOTAL	-	.228	.212	.139	.124	.092

emerged between group III and groups I and II lumped together. While groups I and II received ordinal guidance, which served to reverse an average of 16.5 percent of their pair-wise preferences, group III received a neutral task. This neutral task appeared to have a substantially smaller impact upon altering preferences, since an average of only 8.1 percent of group III's pair-wise preferences were reversed. The significance of the difference between these means was .020 (1-tail). The significance of group III's drop in reversals between the third and fourth rankings was .003 (1-tail), while the significance of group I and II's drop was .071 (1-tail).

On the fifth ranking we see the familiar spread among the three group means. A relatively high reversal rate was maintained by introduction of the cardinal guidance factor to group I subjects. In fact, group I subjects displayed a slight, but insignificant increase. Group II displayed a decreased reversal rate following their first neutral task just as did group III in the previous period. The significance of group II's decrease was .066 (1-tail). Group III, on the other hand, showed an insignificant change (actually a small increase) from the previous period. Apparently the second neutral task had just as little impact as the first. The resultant pattern of mean differences showed group I undergoing more reversals than group III with significance .049 (1-tail) and more reversals than group II with significance .225 (1-tail).

The act of making a final decision had little impact on both group II and group III subjects. Both groups showed small reductions in reversal rates, but these were not particularly significant. In contrast, group I showed a substantial drop in percentage reversals similar to the drops shown by group III on the fourth ranking and group II on the fifth ranking. The significance of groups I's drop was .037 (1-tail). The result was convergence of all three groups on the sixth ranking. No significant differences among final mean reversal rates were apparent.

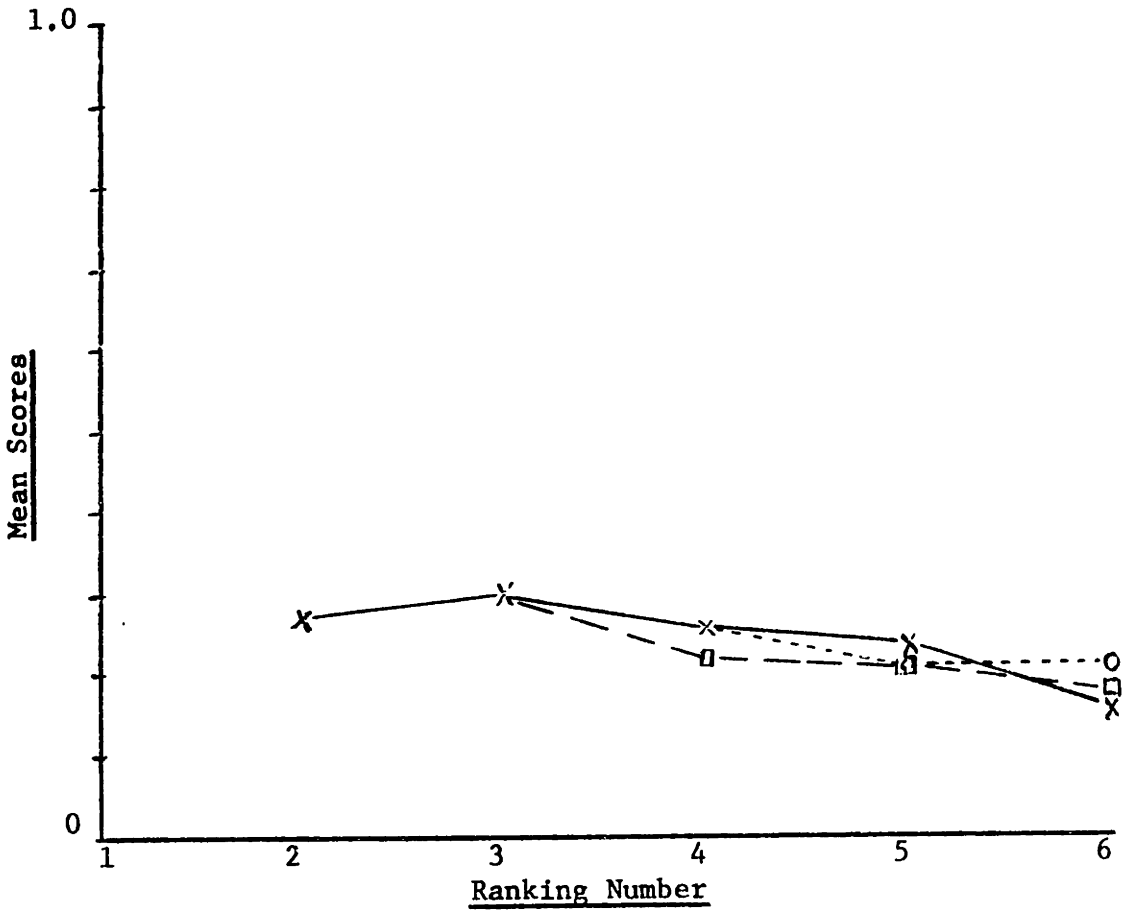
5.4.2.6 Estimated Reversals

The pattern of estimated percentage reversals was similar in many respects to the pattern of actual reversals. This can be seen from Figure 6.

However, three important differences should be noted. First, notice that the mean percentage of estimated reversals exceeded the mean percentage of actual reversals by every group on every ranking. This can be demonstrated by comparing the related numbers tabled at the bottom of Figures 5 and 6, respectively. These results provide strong evidence of a consistent perceptual bias. Subjects did not believe that their preferences were as stable over time as they actually were. Although no formal analyses were performed on the significance of these differences, it is quite clear from their magnitude (whose grand mean fell at approximately .09) and from the perfect uniformity of the pattern that a consistent perceptual bias did exist.

Figure 6

Mean Group Scores
Estimated Percentage of Reversals



X = Group I alone or combined
0 = Group II alone
□ = Group III alone

Mean Scores

Group Number	Ranking Number					
	1	2	3	4	5	6
I	-	.191	.297	.202	.237	.155
II	-	.308	.324	.305	.217	.218
III	-	.326	.286	.227	.216	.183
TOTAL	-	.274	.303	.246	.223	.186

Second, the drops in mean estimated reversal rates displayed by group III on the fourth ranking and by group II on the fifth ranking agreed in both timing and direction with the related actual drops. However, both the magnitude and significance of these estimated drops were less than their actual counterparts. This suggests a dampening mechanism, over and above the bias mechanism, existing within the subjects' perceptual apparatus. In fact, almost all of the differences (between groups and across time) were similarly dampened.

Finally, notice the relatively large drop displayed by group I between the fifth and sixth rankings. This was significant at .022 (1-tail), while the corresponding drop displayed by group II was significant at only .497 (1-tail). Group III displayed a small, but insignificant rise during the same period. This suggests that cardinal guidance may have had a particularly strong influence in counteracting both the bias and dampening mechanisms.

In other respects, mean estimated reversals behaved like mean actual reversals. Therefore, a detailed report of statistical results will be omitted.

5.4.2.7 Prediction Coefficient

Now we come to the issue of prediction. Recall that the measure of prediction selected was the net number of rank-order agreements between an early ranking and the final ranking (i.e., the number of instances where pairs of alternatives agreed in rank-order minus the number of instances where they disagreed). This measure was then adjusted for tied ranking and normalized to fall between minus one and

plus one. Minus one indicates a complete inversion between rank-orders, zero indicates an equal number of agreements and disagreements, and plus one indicates perfect agreement and, therefore, perfect prediction. Figure 7 shows mean group results on this coefficient.[21]

Before discussing Figure 7 in detail, a few preliminary statements would be helpful. In particular, the logical relationship between the prediction coefficient and the two reversal measures should be spelled out.

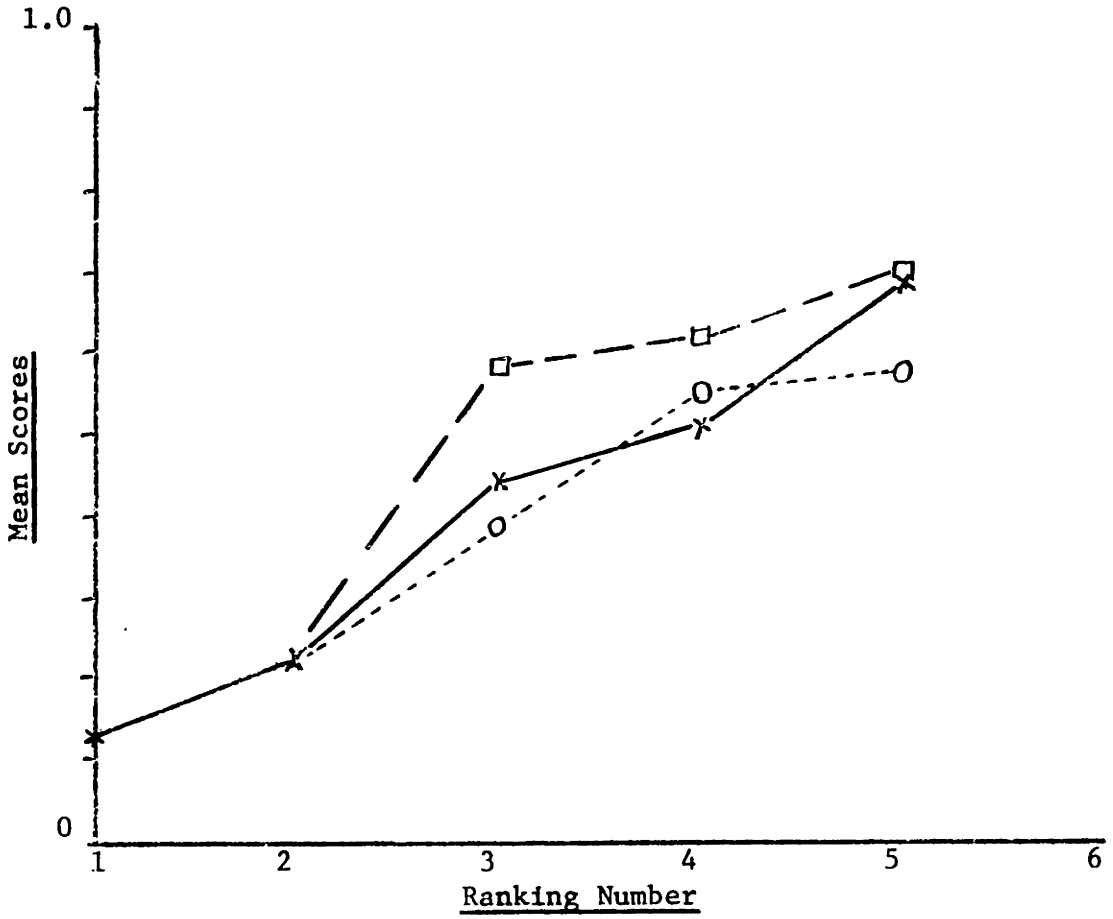
Whereas the reversal measures compare preference orderings on a given ranking with the immediately preceeding ranking, the prediction coefficient compares these uniformly with the sixth ranking performed directly after a final decision was made. Consequently, they provide somewhat independent information. Changes in the prediction coefficient can only occur if reversals occur, but the converse is not necessarily true. That is, a substantial number of reversals may occur over time, but, unless these are systematically related to the final decision, no changes need occur in the prediction coefficient. It is because of this semi-independent relationship that the two measures should be considered together. A pattern of reversals, unaccompanied by properly patterned changes in the prediction coefficient, indicates random instability. However, if a distinct pattern emerges from the prediction coefficient, and

---

21. The prediction coefficient is really Kendall's Tau rank correlation coefficient used as a purely descriptive measure. A complete description of it and the various statistical tests used in this thesis may be found in Hays, W.L., Statistics for Psychologists, (1963).

Figure 7

Mean Group Scores
Prediction Coefficient



X = Group I alone or combined
0 = Group II alone
□ = Group III alone

Mean Scores

Group Number	Ranking Number					
	1	2	3	4	5	6
I	.127	.185	.447	.513	.680	-
II	.081	.163	.389	.556	.577	-
III	.165	.357	.583	.625	.709	-
TOTAL	.123	.232	.469	.564	.653	-

if this pattern is properly related to the pattern of reversals, then we have evidence of directed changes in preference (i.e., directed with respect to the final decision). The next question is, what constitutes a pattern of prediction coefficients "properly related" to the reversal measures?

Recall from Figures 5 and 6 that the most salient changes in mean reversal rates occurred as follows.

1. Group III dropped between the third and fourth rankings following receipt of biographical information and simultaneously with their first neutral task. Small changes occurred thereafter.

2. Group II dropped between the fourth and fifth rankings following ordinal guidance and simultaneously with the ~~same~~ second neutral task. Small changes occurred thereafter.

3. Group I dropped between the fifth and sixth rankings following cardinal guidance and simultaneously with making a final choice.

If this is to be taken as evidence of directed rather than random behavior (i.e., preference alterations directed toward the final decision), then the prediction coefficient should display the following pattern of changes.

1. Group III should display a dramatic increase in the mean prediction coefficient between the second the third rankings and remain relatively stable thereafter.

2. Group II should display a similar increase between the third and fourth rankings and remain relatively stable thereafter.

3. Group I should display its large increase between the fourth and fifth rankings.[22]

Inspection of Figure 7 shows that almost exactly this pattern did occur.

Group means differed insignificantly on the first two rankings as usual. The two significance levels were .699 and .180, respectively. However, the increase in predictive ability displayed during the no information, no guidance time interval by all three groups lumped together was significant at .008 (1-tail).

Group III displayed the most dramatic increase on the third ranking significant at .017 (1-tail). Groups I and II displayed increases of almost identical significance during the same period. However, group III displayed a mean prediction coefficient substantially higher than groups I and II lumped together. The significance of this mean difference was .042 (1-tail).

On the fourth ranking, group II showed the most dramatic increase. This was significant at .010 (1-tail). Groups I and III showed less dramatic increases significant at .124 (1-tail) and .078

22. Naturally, all three groups would show a prediction coefficient of one on the sixth ranking, since any ranking agrees perfectly with itself. These results are omitted from Figure 7.

(1-tail), respectively. The three group means differed insignificantly at that time.

On the fifth ranking, group I showed the most dramatic increase. This was significant at .009 (1-tail). Group II showed practically no change at all, and group III showed a moderate increase significant at .025 (1-tail).

The only important departure from the expected pattern occurred with group III's moderate, but significant increase on the fifth ranking. This result, in combination with a similar result on the percentage confidence measure, will be interpreted in Section 5.5.2.

5.4.2.8 Clarification

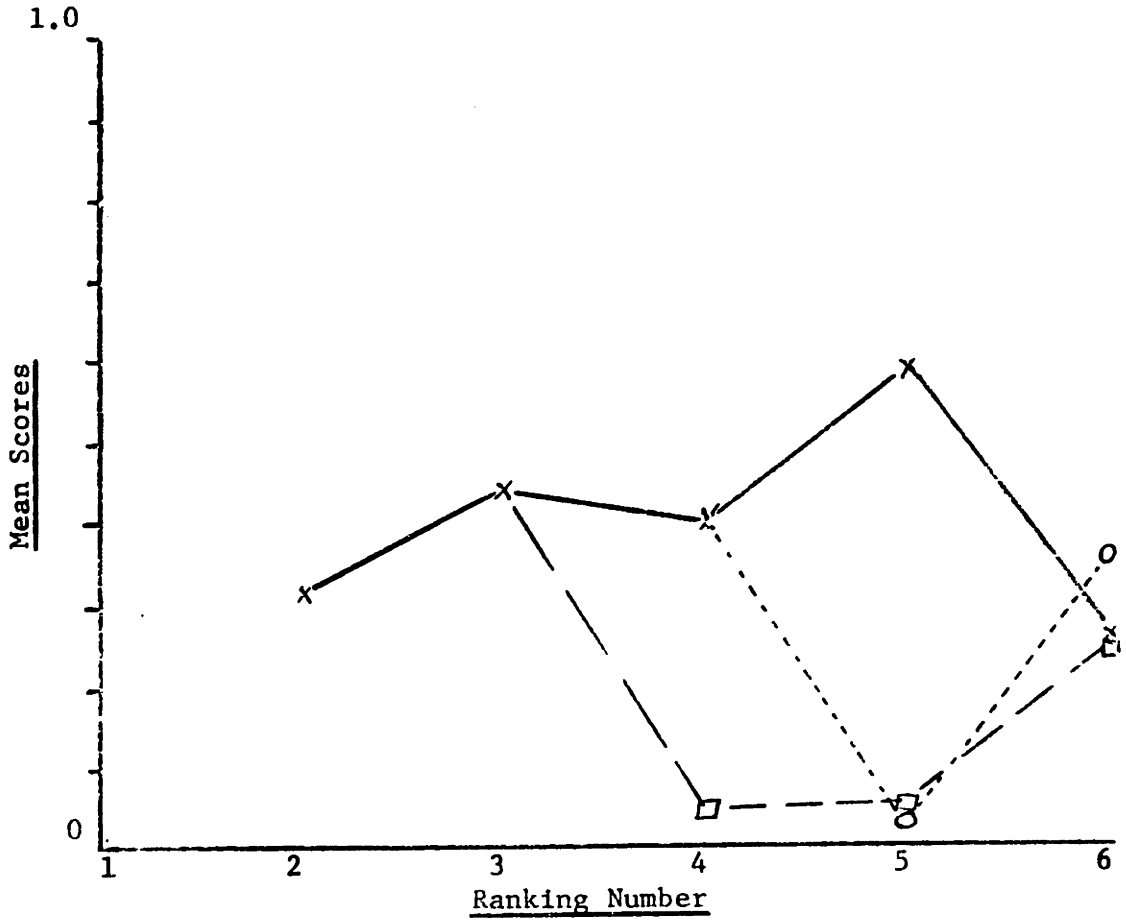
The most dramatic patterns of all emerged on clarification, helpfulness, and gains versus time and effort expended. Results on the clarification measure are shown in Figure 8.

However, before launching into a detailed discussion of these results, a few words about the clarification measure itself are in order. The clarification question, unlike the confidence question, referred to a dynamic process. That is, subjects were asked to assess the extent to which their preferences had been clarified (by virtue of an experimental factor) rather than to assess their current state of clarity. This is also true in the case of the helpfulness and gains versus time and effort questions. It will help to keep this in mind when interpreting results.[23]

23. This also explains why no results appear on any of these three measures at the time of the first ranking.

Figure 8

Mean Group Scores
Normalized Clarification



X = Group I alone or combined
0 = Group II alone
□ = Group III alone

Mean Scores

Group Number	Ranking Number					
	1	2	3	4	5	6
I	-	.329	.526	.395	.592	.263
II	-	.288	.438	.413	.038	.363
III	-	.347	.375	.056	.069	.250
TOTAL	-	.320	.447	.294	.232	.294

As usual, no significant differences emerged between any groups on the second and third rankings. Two one-way analyses of variance were performed, generating significance levels of .709 and .143, respectively. Similarly, no significant differences emerged between groups I and II on the fourth ranking, as indicated by a significance level of .835.

However, group III suffered a drastic reduction in clarification on the fourth ranking following their first neutral task. This reduction was significant at well beyond the .00001 level (1-tail). More important, the mean clarification attributed by group III subjects to their first neutral task was very close to zero, and this differed enormously from groups I and II lumped together. The significance of this mean difference was also significant at well beyond the .00001 level (1-tail).

Group II subjects responded almost identically when they received the same neutral task and performed their fifth ranking. An equally drastic decline in clarification was suffered. This was significant at well beyond the .00001 level (1-tail), and the mean difference between groups I and II ~~were~~ was equally significant. In contrast, group III displayed a negligible and insignificant change from the fourth ranking. Groups II and III differed insignificantly on the fifth ranking.

Following the final choice and sixth ranking, all three groups converged with respect to this measure. Final mean differences were significant at only the .499 level.

Let us now investigate the incremental clarifying impact of the various experimental factors. For this purpose, attention should be restricted to group I subjects, since they were the only ones who received all five factors. It is quite clear from Figure 8 that cardinal guidance was reported as the most clarifying factor of all. T-tests preformed on the mean differences between individual clarification scores (group I only) yielded the following significance levels:

1. Cardinal guidance (fifth ranking) versus no information, no guidance (second ranking) -- .0001 (1-tail);

2. Cardinal guidance versus raw information (third ranking) -- .115 (1-tail);

3. Cardinal guidance versus ordinal guidance (fourth ranking) -- .003 (1-tail);

4. Cardinal guidance versus making a final choice (sixth ranking) -- .0001 (1-tail).

From these results we can see that raw information constituted the only close competitor.

Similar analyses comparing raw information, the second most clarifying factor, with the remaining factors yielded the following significance levels:

1. Raw information (third ranking) versus no information, no guidance (second ranking) -- .004 (1-tail);

2. Raw information versus ordinal guidance (fourth ranking) -- .028 (1-tail);

3. Raw information versus making a final choice (sixth ranking) -- .019 (1-tail).

Ordinal guidance, the third most clarifying factor, differed insignificantly from no information, no guidance, but was more clarifying than making a final decision with significance .019 (1-tail).

The clarifying impact of no information, no guidance differed insignificantly from the impact of making a final choice.

Before moving to the next set of results, it is important to reiterate the dynamic aspects of the clarification measure. In particular, it should be pointed out that all five of the experimental factors were regarded by subjects as contributing some degree of clarification to the process of formulating preferences for alternatives. In contrast, the two neutral tasks contributed practically nothing, as intended. The previous discussion focused primarily upon various differences in the degrees of clarification provided, but this should not obscure the important differences between the impact of all five experimental factors on the one hand and the two neutral tasks on the other. The fact that group II and III subjects reported almost no clarification from the two neutral tasks casts strong doubt on the possibility of a "Hawthorne" effect operating among group I subjects. This, in turn, permits stronger conclusions to be drawn from group I responses.

5.4.2.9. Helpfulness

Results on the helpfulness measure were practically identical to results on the clarification measure. Inspection of Figure 9 demonstrates this.

Analyses were performed on this measure just like the ones performed on the clarification measure. Once again, the results were almost identical. Because there was such great similarity, a detailed presentation will be omitted. Let it suffice to say that virtually the same significance levels were achieved in all cases. Consequently, the same conclusions may be drawn concerning helpfulness as were drawn concerning clarification.

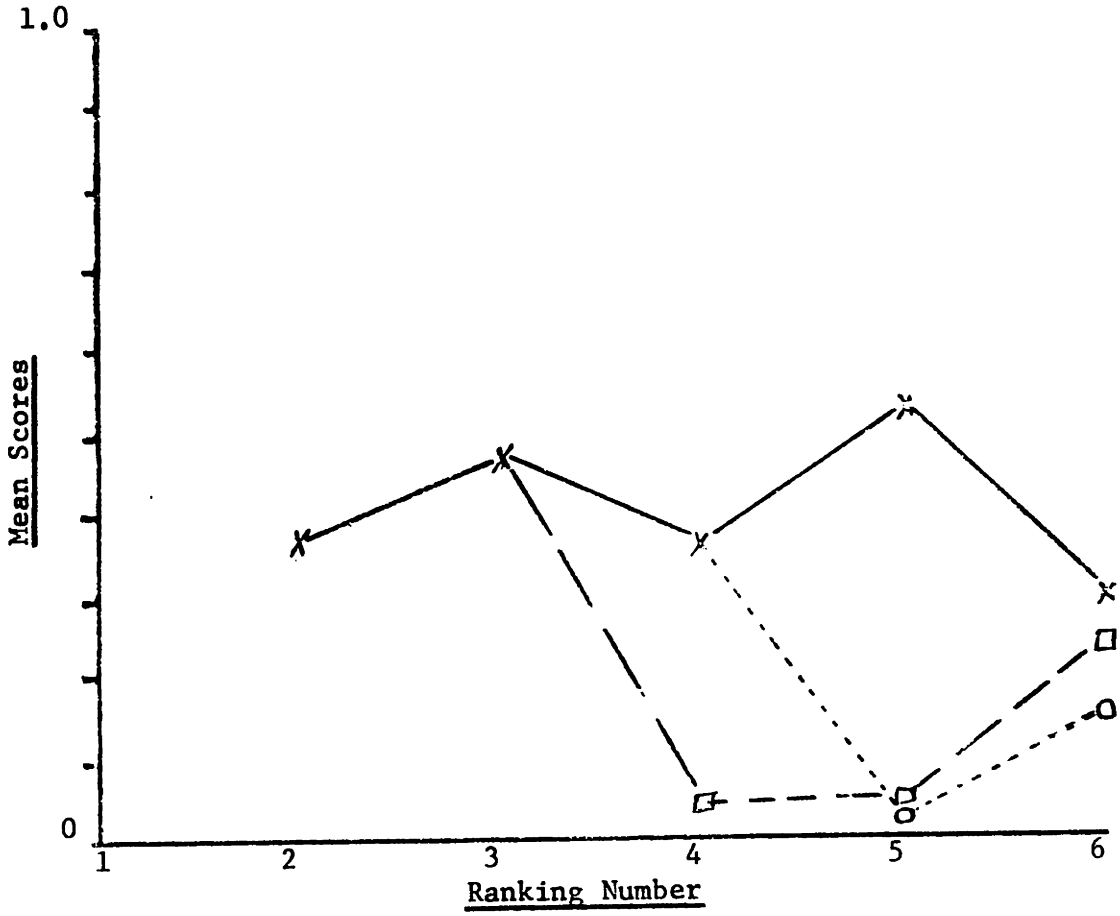
5.4.2.10 Gains Versus Time And Effort

Inspection of Figure 10 shows that responses to the gains versus time and effort question were extremely similar to responses on both the clarification and the helpfulness questions. Consequently, a detailed discussion will be omitted with the understanding that similar conclusions may be drawn concerning this measure too.

One difference should be pointed out, however, to facilitate the interpretation of Figure 10. Recall that both the clarification and helpfulness questions were five-point rating items. Responses on these items were linearly transformed to fall on a scale between zero (indicating "not at all") and one (indicating "completely"). The gains versus time and effort question was treated a bit differently. First it was a seven-point rating item ranging from "substantially less worthwhile" to "substantially more worthwhile". Therefore, to reflect

Figure 9

Mean Group Scores
Normalized Helpfulness



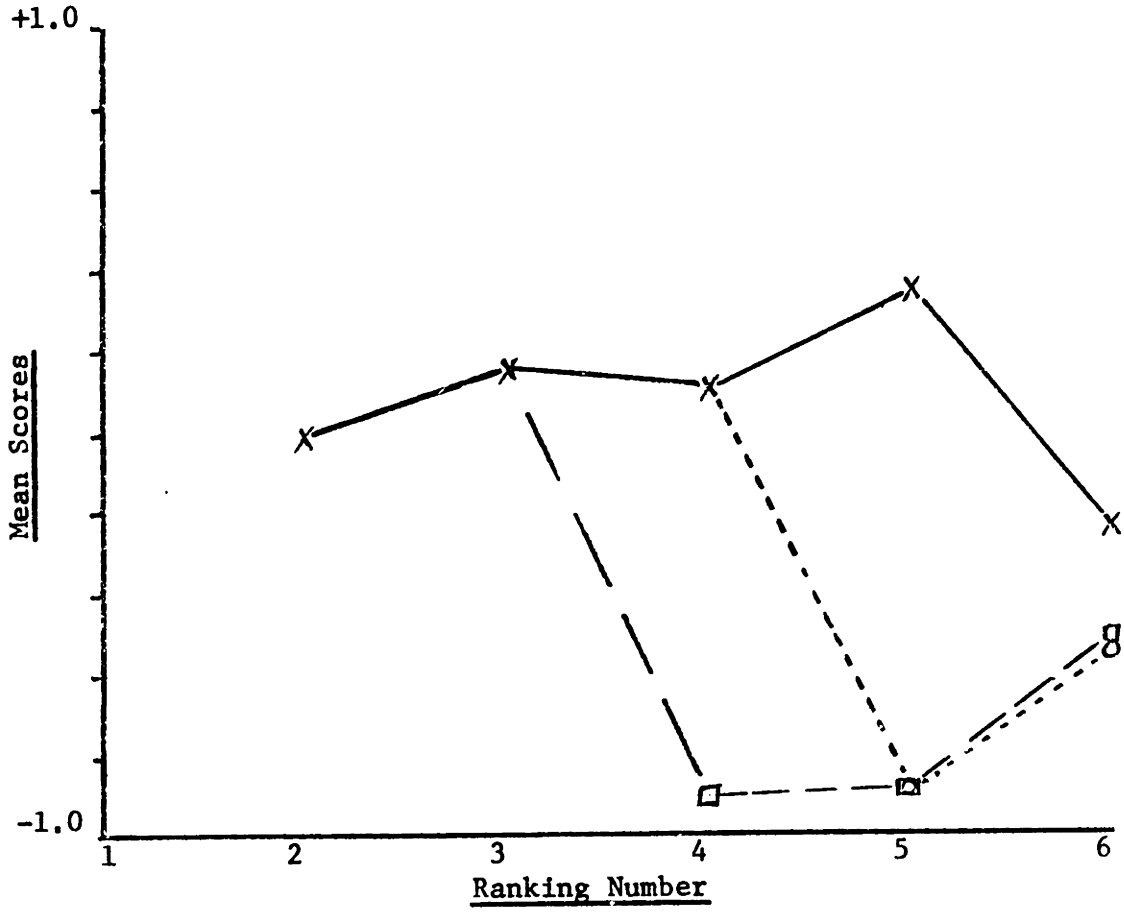
X = Group I alone or combined
0 = Group II alone
□ = Group III alone

Mean Scores

Group Number	Ranking Number					
	1	2	3	4	5	6
I	-	.434	.526	.395	.540	.303
II	-	.325	.450	.338	.038	.163
III	-	.361	.458	.056	.069	.250
TOTAL	-	.373	.478	.268	.215	.237

Figure 10

Mean Group Scores
Normalized Gains Vs. Time and Effort



X = Group I alone or combined
O = Group II alone
□ = Group III alone

Mean Scores

Group Number	Ranking Number					
	1	2	3	4	5	6
I	-	.088	.140	.123	.351	-.246
II	-	-.267	.183	.100	-.867	-.533
III	-	.111	.167	-.870	-.852	-.519
TOTAL	-	-.029	.164	-.199	-.456	-.433

this two-sidedness, responses were linearly transformed to fall on a scale between minus one and plus one. As a result, the interpretation of a zero score on this transformed scale is that the gains realized from introduction of an experimental factor were assessed as just equal in value to the time and effort expended (i.e., neither a "profit" nor a "loss" were incurred).

5.4.2.11 Choice Of Ranking

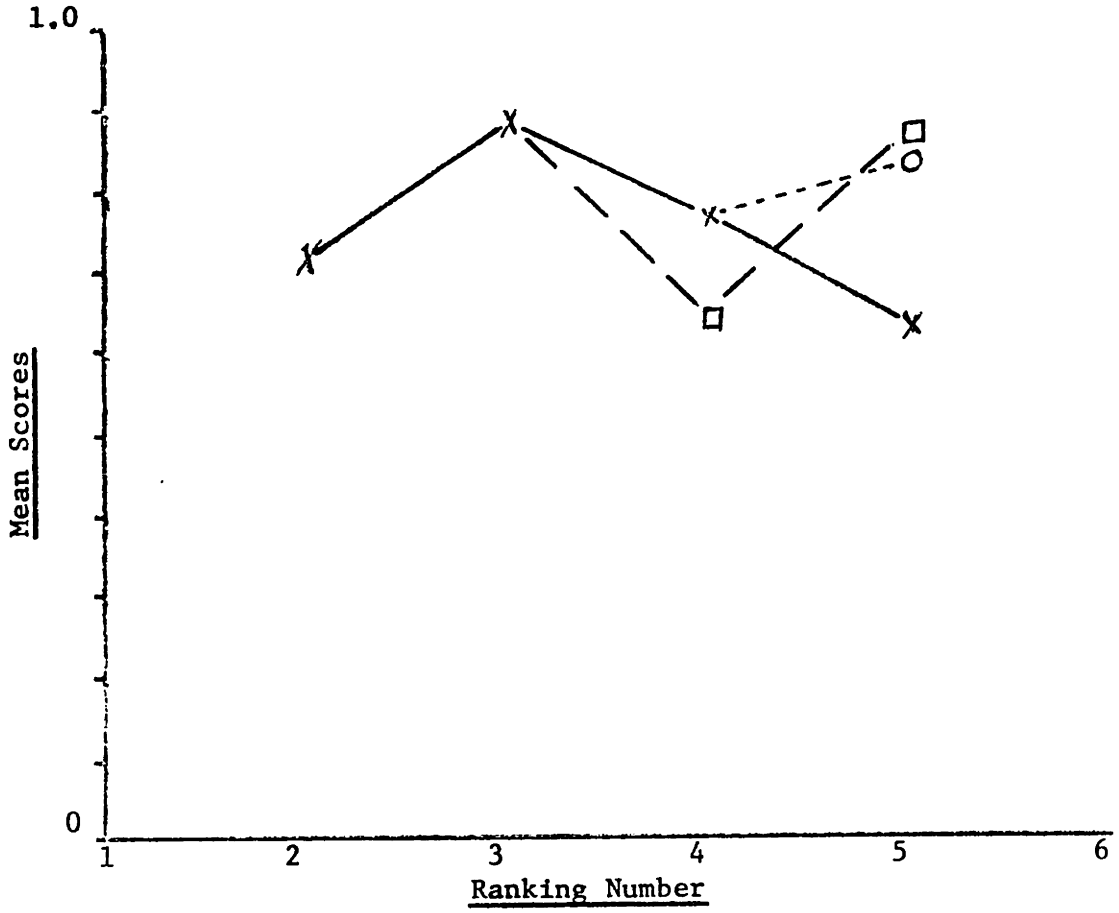
After performing the second, third, fourth, and fifth rankings, subjects were asked to compare their most recent ranking with earlier rankings and to indicate which one better reflected their current preferences. One use of this information (defining chosen confident discriminations) has already been reported in Section 5.4.2.4. Two other uses will be reported in this section, and additional light will be shed on the single unusual finding reported on chosen confident discriminations.

One question which might be raised concerns between-group differences and time trends in the frequency with which subjects chose recent rankings over earlier rankings. Group relative frequencies (ignoring indifference responses) on this measure are shown in Figure 11.[24]

24. Choices were not made on either the first or sixth rankings.

Figure 11

Mean Group Scores
Relative Frequency of Choosing
Most Recent Ranking



X = Group I alone or combined
0 = Group II alone
□ = Group III alone

Mean Scores

Group Number	Ranking Number					
	1	2	3	4	5	6
I	-	.735	.938	.848	.625	-
II	-	.707	.883	.764	.833	-
III	-	.714	.833	.643	.877	-
TOTAL	-	.717	.883	.728	.805	-

Inspection of Figure 11 shows a great deal of instability in these relative frequencies, but no particular pattern. Actually, none of the usual comparisons yielded significant differences.[25]

A second question, and by far the more interesting one, concerns the choice of a computed ranking on the part of group I subjects following their ordinal and cardinal guidance interviews, respectively. Whereas two out of the 19 group I subjects chose their computed lexicographic ranking over any past or present subjective ranking, 10 out of these same 19 subjects chose the computed ranking generated by the complete assessment procedure. This difference is certainly striking.[25]

In light of the above result we may now interpret the significant increase in group I's chosen confident discriminations on the fifth ranking (see Figure 4). Since more than half of these subjects chose their computed ranking following cardinal guidance interviews, whereas almost everyone had chosen subjective rankings in previous periods, and since the computed rankings ordinarily implied more preference discriminations than subjective rankings, a significant increase in chosen confident discriminations resulted.

5.4.2.12 Attitude Shift And Methods Utilization

There are two additional results to investigate before moving

25. Statistical analyses of shifts over time were not performed in this instance. Since we are dealing with frequency data, the previous device of eliminating serial correlation by averaging individual difference scores could not be used.

to an integrated interpretation of the entire experiment. One of these pertains to shifts in attitudes toward quantitative techniques of evaluation. The other pertains to spontaneous utilization of assessment methods introduced during the experment.

Recall that a single attitude question was included within the pre-experimental questionnaire. A similar item was included in the post-experimental questionnaire administered directly after all subjects had completed the computer game. Both of these items asked subjects to indicate how favorably they felt toward numerical techniques of evaluation as a useful tool in decision making. A seven-point rating scale was provided in each case. Responses to these items were then transformed linearly to fall on a numerical scale ranging from minus one (indicating extremely unfavorable attitudes) through zero (indicating neutrality or indifference) to plus one (indicating extremely favorable attitudes). Table 6 below displays group means on both prior and posterior attitudes, as well as mean attitude shifts.

Table 6

Mean Attitudes Toward Numerical Techniques Of Evaluation

	Group I	Group II	Group III	TOTAL
Prior Attitudes	.417	.500	.533	.483
Posterior Attitudes	.567	.283	.165	.338
Attitude Shifts	+.150	-.217	-.368	-.145

Inspection of Table 6 brings out two important facts. First, the total sample (all three groups lumped together) suffered a moderate negative shift in attitudes over the span of the entire eleven-week training course. A T-test (assuming no real shift) was performed on this total sample shift and showed it to be significant at .027 (2-tail).

However, the more interesting result from our point of view lies in the differential responses of the three experimental groups. Both group II and group III followed the predominant negative trend, but group I shifted in a positive direction (i.e., toward more favorable attitudes). Whereas the mean difference between group II's and group III's negative shifts was not significant (a T-test yielded a 2-tail level of .398), the mean difference between group I and groups II and III lumped together was significant at .002 (2-tail).

These results suggest that introduction of the cardinal guidance factor (i.e., the complete assessment procedure developed in Chapter II) had a critical impact on attitudes. Group I subjects were the only ones to receive this factor. On the other hand, differential treatment of groups II and III appeared <u>not</u> to influence their attitude shifts in any systematically different manner. Thus we may conclude that the total experience of the training course, including all of the experiment except cardinal guidance, had the general effect of inducing a moderate negative shift in attitudes. However, receipt of cardinal guidance served not only to neutralize this effect, but to reverse it.

Very similar results were obtained on the methods utilization question. This item also appeared on the post-experimental questionnaire and contained a five-point rating scale. As usual, responses were transformed linearly, but this time to fall on a numerical scale ranging from zero (indicating no utilization of experimentally induced assessment methods) to one (indicating complete utilization).

Table 7 displays groups means on this measure.

Table 7

Mean Utilization Of
Experimentally Induced Assessment
Methods For Purposes Of
Making Decisions In The Computer Game

MEASURE	Group I	Group II	Group III	TOTAL
Methods Utilization	.513	.400	.276	.394

Table 7 shows that group I spontaneously adopted assessment methods more than both group II and group III. The mean difference between groups II and III was again not particularly significant (a T-test produced a 2-tail significance of .139), but the difference between group I and groups II and III lumped together was significant at .014 (2-tail). These results serve to amplify the conclusions drawn concerning attitude shifts.

5.5.0 DISCUSSION

This section will be divided into four parts. First, the implications of sample homogeneity will be discussed. Second, results will be interpreted in terms of the five experimental factors. Third, results will be reinterpreted from the point of view of the assessment procedure. Finally, discussion will close with some suggestions for further research.

5.5.1 Implications Of Sample Homogeneity

In Section 5.4.1, we discussed the effectiveness of randomization with respect to:

1. prior education, experience, and attitudes regarding quantitative techniques of evaluation;
2. the structure of sociometric preferences within ten-man sub-groups;
3. level of formal education achieved;
4. rank or grade;
5. branch of service.

In all respects but the last, it was concluded that randomization had been highly effective in homogenizing the composition of the three experimental groups.

It should be pointed out, however, that in many respects, the entire sample was extremely homogeneous to begin with. The reason for this lies in the process by which subjects were originally selected to participate in the training course. They were recommended by their respective organizations (usually on the basis of merit) to learn better how to discharge the duties of military project management. Because of this prior selection mechanism, the total sample was quite homogeneous with respect to:

1. age (about thirty-five to fifty);
2. sex (uniformly male);
3. annual income (about twelve to eighteen thousand dollars);
4. professional orientation and experience (management of military and government projects);
5. professional competence (distinctly higher than the average of their respective organizations);
6. attitude toward learning (quite favorable).

Now the above observations have several very important implications. First, since the total sample was so homogeneous in so many respects, dramatic between-group differences which emerged following differential experimental manipulations cannot reasonably be attributed to unknown and uncontrolled random variations in the relative composition of the three groups. All subjects were just too much alike to

support such a contention. Even if extreme relative imbalances in composition did occur in various unknown ways, still, the range of possible variation was so restricted by the homogenizing effect of the prior selection mechanism as to render them probably unimportant.

Second, by undertaking a formal random assignment procedure, the likelihood of extreme compositional imbalances among the three groups was substantially reduced.

Third, on the several specific factors which the writer felt might have exerted a meddlesome influence on results, explicit homogeneity checks were made. Randomization was shown to have been quite effective in all of these cases.

The above considerations lead us to a rather happy conclusion. Whatever important and significant differences did emerge between groups must have been induced by the differential experimental manipulations performed. No other interpretation seems reasonable. In addition, a similar and even stronger conclusion may be drawn regarding within-group differences over time. Here, subjects acted as their own controls, which renders the compositional imbalance argument even less tenable.

Before moving to an interpretation of specific experimental effects, two additional comments should be made. These refer to the specific ways in which the sample was homogeneous. From our knowledge of the subjects themselves, their backgrounds, and the process by which they were selected for the training course, it is reasonable to conclude that almost all of them were quite favorably disposed to learning and to the whole concept of experimentation. The writer's personal impressions of the subjects support this conclusion. As a result, they were

motivated to cooperate, to participate, and to commit much more of their personal time and effort than the writer had originally anticipated (or even hoped). This may account, in part, for the clarity and consistency of results. In addition, they probably reacted more uniformly to uniform experimental treatment than if less interest and personal committment had been aroused. This may account, in part, for the high levels of statistical significance achieved on almost every result. For these reasons, both the subjects and the setting must be regarded as ideal for purposes of experimentation. On the other hand, for exactly the same reasons, great care must be taken in generalizing conclusions drawn from the experiment. It is not at all clear that other types of individuals, or even the same individuals in a different setting, would have reacted similarly. This issue of generalizability will be discussed more fully in Chapter VI.

5.5.2 Interpretations By Experimental Factor

Let us now attempt an integrated interpretation of results in terms of the five experimental factors. Fortunately, our task will be greatly simplified by the fact that most of these results were individually clear and collectively reinforcing. In addition, they regularly achieved a high degree of statistical significance.

Our strategy will be as follows. First, the important and significant results relating to each experimental factor will be summarized. Second, these results will be integrated and, in some cases, embellished with anecdotal evidence to provide a more complete

interpretation. Third, the incremental and cumulative impact of each factor will be assessed by making inter-factor and time series comparisons. This strategy will be implemented in the immediately following section.

5.5.2.1 The Impact Of No Information, No Guidance

Consider the first experimental factor -- no information, no guidance. The important results relating to its impact are summarized below.

1. There was no systematic impact on the number of raw preference discriminations made by any of the three groups (mean remained at around 7.1). However, substantial, but compensating, individual changes were known to have occurred.

2. Confidence in the accuracy of assigned rank numbers increased from a low (.33) to a moderate (.48) level for all three groups.

3. The number of confident discriminations was similarly, though less conclusively increased (from 2.1 to 3.3).

4. Preferences for alternatives were altered to a moderate degree (23% reversal rate).

5. Subjects estimated a somewhat higher rate of reversals (27%).

6. Nevertheless, alterations in preferences were not completely random. They were directed slightly toward the final choice (prediction coefficient increased from .12 to .23).

7. The lapse of time without formal information or guidance was regarded as slightly clarifying (.32) and slightly-to-moderately helpful (.37) in terms of formulating preferences and making a final choice.

8. On the other hand, these gains were considered insufficient to justify the time and effort expended (-.03).

9. All sixty subjects were treated identically, and no significant inter-group differences emerged in terms of these (or any other) measures.

From the above results we may conclude that individual decision processes were operative, but at a rather low level. This occurred without any formal information or guidance provided by the experiment. However, there were several experimental manipulations performed during the period which undoubtedly contributed in part to these results. Subjects were informed of the experiment (introductory address), and they were asked to answer two sets of questions and to perform two separate rankings. In addition, subsequent conversations (during the ordinal guidance interviews and informally) established that some attention was being paid to the experimental decision, albeit small. The major

contribution to these results probably sprang from spontaneous activities on the part of the subjects to acquaint themselves with one another. This conclusion received further support from subsequent conversations.

In summary, the mere introduction of the experiment and the requirement to answer questions and to rank alternatives probably served to motivate the above results. The actual mechanisms by which they were mediated however, sprang spontaneously, and perhaps semi-consciously, from within the subjects themselves.

5.5.2.2 The Impact Of Raw Information Without Guidance

Consider next the impact of raw information without guidance. The important results relating to this factor are summarized below.

1. Once again, there was no systematic impact on the number of raw preference discriminations made by any group.

2. Confidence again increased from a moderate (.48) to a high-moderate (.62) level. This was true for all three groups.

3. The number of confident discriminations was similarly, though less conclusively increased (from 3.3 to 4.4), and the number of chosen confident discriminations increased almost identically (from 3.2 to 4.2).

4. Apart from a slight and insignificant decline, the same moderate level of preference instability persisted (21% reversal rate).

5. However, subjects estimated a significantly higher reversal rate (30%). This estimated reversal rate was higher in two respects. First, it was nine percentage points higher than the actual rate, and, second, it marked a significant increase from the previous period, even though the actual reversal rate fell slightly.

6. A sharp and highly significant increase in the prediction coefficient occurred for all three groups (from .23 to .47). This was the most significant increase for group III subjects in all periods, and they achieved a significantly higher coefficient then groups I and II at that time (.58).

7. Raw information was regarded as providing low-moderate clarification (.45), as being moderately helpful (.48), and as being just a little more worthwhile than the time and effort expended (.16) in terms of formulating preferences and making a final choice. All three of these responses constituted dramatic and highly significant improvements over no information, no guidance.

8. All sixty subjects were treated identically, and, apart from the difference between group III and groups I and II combined on the prediction coefficient, no significant inter-group differences emerged.

From the above results, we may conclude that the introduction of biographical information, even without any formal guidance concerning what to do with it, had an important impact on individual decision processes. Since the biographical information was in the form of a personal resume, and since many of the subjects had substantial previous experience assessing resumes, we may infer that many of them already knew how to process the information. As such, a certain degree of guidance was provided spontaneously and internally. This process was carried out informally by most subjects, but with complete awareness and intent. At least this is what group I and II subjects later claimed during the ordinal guidance interviews. However, in one exceptional case, a formal processing mechanism was adopted, as evidenced by numerous notations, including quantitative assessment data, discovered on his third ranking sheet.

Direct evidence supporting the substantial incremental impact of raw information over no information comes from the persisting rate of reversals and the sharp increase in predictability. Additional confirmatory evidence comes from the confidence, estimated reversals, clarification, helpfulness, and gains versus time and effort measures.

One observation should be made concerning the relationship between actual and estimated reversals. Recall that actual reversals fell slightly, while estimated reversals increased markedly. This fact, in conjunction with the dramatic increases in predictability, confidence, clarification, helpfulness, and gains versus time and effort,

suggests that reversals occurred for different reasons under this and the previous factor. Under no information, no guidance, reversals may be attributed primarily to random instability based on little knowledge of the alternatives and low interest in the decision. However, following receipt of the biographical sketches, a persistently high rate of reversals probably indicates something else -- consciously intended changes in preference. This interpretation is rendered particularly plausible by the dramatic increase in estimated reversals. Here, and in the other measures tapping essentially conscious processes, we see strong evidence for a sudden increase in both conscious awareness of and organized attempts to solve the decision problem. Subsequent discussions, wherein subjects recalled a marked increase in interest following receipt of the biographical information, lends even further support to this conclusion.

In summary, the introduction of raw information seems to have been one of the critical moments of the experiment. It increased substantially the salience of the decision and motivated both interest and problem-solving activity. Most important of all, it placed the decision problem squarely within the realm of conscious concern, whereas for many it had previously received only casual attention.

5.5.2.3 The Impact Of Ordinal Guidance

Now let us investigate the impact of ordinal guidance. The important results relating to this factor are summarized below.

1. Again, no important changes in raw discriminations occurred.

2. Groups I and II combined increased their confidence from a high-moderate (.62) to a substantial (.72) level. However, group III registered a small and insignificant increase (to .67).

3. The number of confident discriminations and chosen confident discriminations moved similarly, though less conclusively.

4. Groups I and II combined showed a slight and marginally significant drop in their rate of reversals (from 21% to 17%), while group III showed a dramatic and highly significant drop (to 8%).

5. Estimated reversals followed the pattern of actual reversals, but with the same upward bias persisting. Also, the first evidence of a dampening mechanism appeared, since these drops were generally neither as dramatic nor as significant.

6. A sharp and highly significant increase in the prediction coefficient occurred for group II (from .39 to .56). Only slight and marginally significant increases occurred for groups I and III (from .45 to .51, and from .58 to .63, respectively).

7. Ordinal guidance was regarded as providing low-moderate clarification (.40), as providing low-moderate help (.37), and as being just a little more worthwhile than the time and effort expended (.11) by groups I and II combined. In general, these results constituted significant increases over no information, no guidance and significant decreases compared to raw information. However, the magnitudes of these changes were only moderate.

8. In sharp contrast, the neutral task performed by group III subjects was regarded as not at all clarifying (.06), not at all helpful (.06), and worth substantially less than the time and effort expended (-.87). All three of these results constituted massive and extremely significant declines compared to both of the previous factors.

9. Groups I and II were treated identically, and, apart from their different response on the prediction coefficient, no significant inter-group differences emerged on any measures.

10. However, all of the reported differences between group III and groups I and II combined were directionally consistent, most were at least marginally significant, and half of them were extremely significant.

11. Very few (only five) of the thirty-nine group I and II subjects chose their lexicographic rankings generated during the ordinal guidance interviews.

From the above results, we may draw four sets of conclusions. First, ordinal guidance did have a definite impact upon individual decision processes. This is evidenced by another increase in confidence, confident discriminations, chosen confident discriminations, and the prediction coefficient. It is also evidenced by maintenance of the actual and estimated reversal rates and by distinctly positive results on the clarification, helpfulness, and gains versus time and effort items.

Compared to other factors, however, the impact of ordinal guidance was mixed. The increase in confidence (and its associated measures) and in the prediction coefficient demonstrate positive incremental effects relative to both previous factors. The fact that group III subjects, who did not receive ordinal guidance, demonstrated insignificant changes on both of these measures provides further support to the same conclusion. However, the impact of ordinal guidance on actual and estimated reversals differed only slightly from raw information. In this context we must conclude that these two factors had approximately the same effect, subject to a general downward trend. Finally, ordinal guidance was regarded as decidedly less clarifying, less helpful, and less worthwhile in terms of time and effort expended than raw information; but decidedly more so than no information, no guidance. These latter conclusions will be amplified when we discuss the impact of cardinal guidance.

The third set of conclusions relates to introduction of the neutral task. From group III's responses we have very strong evidence against any "Hawthorne" effect. In particular, we can place much greater credence in group I's and group II's positive reactions to ordinal guidance. There is no reason to believe that they would have reacted any less honestly and incisively than group III, since the three groups were so homogeneous.

A fourth conclusion relates to the cumulative impact of ordinal guidance. On the basis of the low frequency with which group I and group II subjects chose their lexicographic rankings, the process of articulating and ranking criteria, and then using this assessment structure to rank alternatives, may not by itself exert a lasting impact on the decision process. However, let us suspend judgement on this issue until we can examine subsequent evidence.

In summary, ordinal guidance had a definite impact upon the decision process. Relative to previous factors its impact was mixed -- greater in some respects and about the same or less in others. Whether or not its impact was permanent is not yet clear.

5.5.2.4 The Impact Of Cardinal Guidance

By far the greatest number of results and the most exciting results occurred under cardinal guidance. These are summarized below.

1. As usual, nothing happened in terms of raw discriminations.

2. Group I increased from a substantial (.74) to a very substantial (.82) level of confidence. Group II increased to a substantial level (.76). Group III registered another

small and insignificant increase (to .70).

3. Once again, the number of confident discriminations moved similarly, though less conclusively.

4. The number of chosen confident discriminations also moved similarly in the case of groups II and III. However, group I showed a dramatic and significant increase.

5. Group I showed a slight and insignificant increase in its actual reversal rate (from 15% to 16%). Group II showed a substantial and significant reduction (from 18% to 12%). Group III remained virtually unchanged (at 9%).

6. Estimated reversals followed the same pattern as actual reversals. The previously noted bias and dampening effects persisted.

7. This time, group I demonstrated a sharp and highly significant increase in the prediction coefficient (from .51 to .68). Group II remained virtually stationary (at .58). Group III demonstrated a modest but significant increase (from .63 to .71).

8. Cardinal guidance was regarded as providing high-moderate clarification (.59), as providing moderate help (.54), and as being slightly more worthwhile than the time and effort expended (.35) by group I. These results constituted substantial and significant increases over both ordinal guidance and no information, no guidance. They constituted moderate and marginally significant increases over raw information.

9. In sharp contrast, the neutral task performed by group II subjects elicited massive and extremely significant reductions in all three of the above measures. Group II duplicated almost exactly the pattern set by group III in the previous period. Group II found the neutral task not at all clarifying (.04), not at all helpful (.04), and worth substantially less than their time and effort (-.87).

10. Group III evaluated their second neutral task in almost exactly the same manner as they had evaluated the first one.

11. More than half (eleven) of the nineteen group I subjects chose their computed rankings generated during the cardinal guidance interviews.

12. Group I showed a positive shift in attittudes toward quantitative techniques of evaluation (+.15), while groups II and III showed negative shifts (-.22 and -.37, respectively). This difference was highly significant.

13. Group I utilized assessment methods in the computer game significantly more than groups II and III (.51 versus .40 and .28, respectively).

From the above results, it is very clear that cardinal guidance was the single most important factor in the entire experiment. Every one of the above measures (save raw discriminations) bears witness to this conclusion. Not only did it have an important impact in all these respects, but it had the most important impact in terms of chosen confident discriminations, actual reversals, the prediction coefficient, clarification, helpfulness, gains versus time and effort, chosen rankings, attitude shifts, and methods utilization. To understand why this was so, some anecdotal evidence gleaned from tape recordings of the cardinal guidance interviews will be reported.

The major significance of cardinal guidance was its quantitative requirement. Ordinal guidance had required subjects to articulate their criteria and to make ordinal discriminations both between criteria and between alternatives. This was easy. But it did not seem to have either a clear or a lasting impact on preferences. On the other hand, cardinal guidance required a much greater mental effort. It required each group I subject to assign cardinal numbers both to his assessment criteria and to his alternatives. The consequences were startling.

When computed rankings were compared with subjective rankings at the close of the interview, whatever inconsistencies emerged were regarded with surprise and chagrin. Subjects felt compelled to eliminate these inconsistencies. This, in turn, motivated them to think more carefully about their criteria and their entire assessment structure. In addition, the particular inconsistencies which did emerge directed subjects to the underlying causes. As a result, substantial changes were made to the assessment structures in many cases.

In summary, cardinal guidance had the most important and lasting impact on preferences of any experimental factor. The reason for this lay in its quantitative requirement. Being forced to formulate a cardinal worth structure and to set down a quantitative assessment algorithm provided an independent check on the validity of subjective rankings. That is, subjects viewed the quantitative assessment procedure and the subjective ranking procedure as two independent processes which should generate identical results. Cardinal guidance was unique in providing such an independent validity check. In addition, formulating a cardinal asssessment structure served to detect hazy and inconsistent reasoning, it motivated subjects to remedy emergent inconsistencies, and it directed them specifically to the underlying cause of such inconsistencies. These results were quite favorably received by the subjects, and their impact persisted through the final decision into the computer game. Attribution of the above results to a "Hawthorne" effect was virtually ruled out by group II's sharply contrasting behavior.

5.5.2.5 The Impact Of Making A Final Choice

We now come to the last experimental factor, making a final choice. The important results relating to this factor are summarized below.

1. No important changes occurred in number of raw discriminations.

2. Group I increased from a very substantial (.82) to an even more substantial (.85) level of confidence. Group III declined slightly (to .75). Group III registered a large and quite significant increase (from .70 to .80).

3. Confident discriminations moved similarly, but less conclusively.

4. Group I showed a substantial and significant reduction in actual reversals (from 16% to 10%). Groups II and III showed slight and insignificant reductions (from 12% to 10%, and from 9% to 7%, respectively).

5. Estimated reversals followed the same pattern as actual reversals. However, group I showed evidence of overcoming both the bias and dampening phenomena previously displayed.

6. The most striking impact of making a final choice was to converge all three experimental groups on clarification, helpfulness, and gains versus time and effort. This final

act was regarded as slightly clarifying (.29) and slightly helpful (.24) by the three groups combined. No significant inter-group differences remained at this time.

7. All three groups regarded the act of making a final decision worth slightly-to-moderately less than the time and effort expended (-.25, -.53, and -.52, respectively). However, a noticeable difference emerged between group I and groups II and III combined, suggesting a cumulative effect of the cardinal guidance factor.

From the above results, we may conclude that making a final choice did have some impact on the various measures, but nowhere near as much as raw information, ordinal guidance, or cardinal guidance. Its impact was comparable in many respects to no information, no guidance.

The real importance of these results lies in their ability to indicate cumulative impacts of the previous experimental factors. Recall that all three groups were homogeneous initially and treated identically at final decision time. Therefore, any between-group differences remaining after the final choice had been made can best be interpreted as indicators of cumulative or lasting effects of intermediate factors. Let us investigate these residual between-group differences.

The strongest indicators of cumulative effects were confidence, estimated reversals, the prediction coefficient, gains versus time and effort, attitude shifts, and methods utilization. The lasting impact

of cardinal guidance in terms of attitude shifts and methods utilization has already been discussed. An additional result supporting this same conclusion can be found in group I's sudden drop in estimated reversals following the final choice. These subjects appear to have counteracted both the bias and dampening mechanisms. They made far more realistic estimates of their actual reversal rate at final decision time. Still another supporting result was the substantial difference between group I and groups II and III combined on gains versus time and effort. There appears to have been a positive cumulative effect of cardinal guidance which carried over into the final choice.

Group II seems to have suffered most from the experiment. Whereas the other two groups showed noticeable and significant increases in confidence and predictability (previous period), group II actually declined in confidence and failed to increase significantly in predictability (previous period). In addition, they later demonstrated a negative attitude shift. All of these results suggest that whatever gains were made by group II during the experiment were neither stable nor lasting. More will be said about this shortly.

Group III displayed unexpected increases in both confidence and predictability (previous period). In addition, these subjects converged toward the final levels attained by group I and II subjects on many measures. Despite the fact that group III had received nothing in the way of guidance from the experiment for six weeks, there was clear evidence that something had occurred, and with good results. To find out what had occurred, the writer asked several group III subjects to explain their responses. From these informal conversations a

valuable insight was gained. It seems that group III subjects, after receiving two neutral tasks, concluded that they were not going to receive any useful guidance from the experiment. They realized that final choices would have to be made on the basis of whatever information they already possessed and whatever assessment methods they might create on their own. Consequently, they directed their efforts accordingly, with a satisfactory result.

In contrast, both group I and group II subjects had come to expect guidance from the experimental manipulations. Whereas group I had received this guidance and judged it quite satisfactory, group II subjects had received only part of it. They were left with an assessment structure which was not entirely satisfactory. The cumulative result of this situation, followed by a demoralizing neutral task, was to arrest their prediction coefficient, to depress their confidence, and to induce a negative shift in attitudes at the close of the experiment.

5.5.3 Interpretations In Terms Of The Assessment Procedure

We are now in a position to evaluate the merits of the complete assessment procedure developed in Chapter II. Our judgements will be based on the results and conclusions relating to cardinal guidance. This factor, it will be recalled, was really the complete assessment procedure specially adapted to the experiment.

The reader is referred to Section 5.5.2.4 for specific results and conclusions regarding the descriptive impact of the assessment procedure. This section will attempt to summarize and integrate these conclusions from the normative point of view.

The most striking impact of the assessment procedure is to induce decision makers to formulate and to validate a consistent preference structure. Specifically, this means:

1. articulating more criteria than would otherwise have occurred;

2. taking greater pains to insure both completeness and internal consistency;

3. checking the implications of a systematically generated assessment structure (e.g. the assessment algorithm) against subjective intuition (e.g., a ranking). Critical to this process is the validation check. This is where the results of systematic reasoning and unsystematic intuition are compared. If inconsistencies emerge, decision makers are strongly motivated to eliminate them, and the assessment structure is such as to direct attention toward the underlying causes thereof.

A second consequence of the assessment procedure is to create a map of the decision maker's worth structure. As such, the procedure acts as a measuring instrument. Alternatively, the procedure may be used as an operational definition of the worth concept. This would be useful for scientific purposes.

Quantitative aspects of the procedure are critical. Performing only the non-quantitative operations (e.g., through articulation of a criterion hierarchy, but not to the point of assigning cardinal weights and scores) is not very useful in terms of either formulating or measuring worth notions. This was clear from the differential reactions of group I and group II subjects.

The assessment algorithm generated by the procedure may actually be used to assess alternatives. The numerical output of this algorithm may be used in the context of a complete, formal decision methodology. However, this should not be attempted unless or until the decision maker has satisfied himself that his algorithm correctly reflects what he want it to reflect.

Finally, use of the procedure will undoubtedly have an important effect on the decision maker himself. In particular, it will alter both his preferences for alternatives and his attitudes toward the task of assessment. On the basis of the experiment we know that these can be highly desirable changes from the decision maker's point of view. Whether or not they are desirable from anybody else's point of view has been defined (by the worth concept) as an irrelevant question.

5.5.4 Suggestions For Further Research

In light of the rather clear results which emerged from the experiment, the writer would not recommend a complete replication. This would not appear to serve any useful purpose. However, there were several factors omitted from the design of the experiment which probably

deserve additional study. These are discussed below.

First, there was the omission of a fourth experimental group discussed in Section 5.2.6. It is not clear what would have happened if subjects were merely asked to perform six successive rankings with neither information nor guidance. Group III's spontaneous invention of an assessment methodology at the end of the experiment suggests that some kind of spontaneous mechanisms would become operative. However, group III did receive raw information, and it is clear that raw information had a substantial impact on all subjects Therefore, we cannot be sure just how a fourth group might have behaved. It would be interesting to find out.

Second, and in the same spirit, it would be interesting to induce various other kinds of assessment structures beside the lexicograpic and cardinal structures induced herein. These two certainly do <u>not</u> exhaust the possibilities.

Third, it is very likely that additional interesting results would emerge if the same sort of experiment were conducted focusing on group (as opposed to individual) decision processes. Some specific suggestions were made in the assessment procedure concerning this kind of situation, and these suggestions have been tested informally on a very limited sample. Nevertheless, no large-scale experimentation has been performed on group processes.

Fourth, a great deal of additional research could usefully be performed on the specific assessment procedures presented in Chapter II. In their current form, they reflect the trial-and-error experience of several dozen decision makers during the last three years. But there is plenty of room for improvement.

Fifth, one issue which was not treated at all during the experiment concerns long-term effects. Although the short-term impact of the various factors was quite clear, we really do not know how the subjects felt in retrospect long after the consequences of their final choices had become apparent. We do know from the attitude and methods utilization questions how they felt shortly (i.e., one week) after making a final decision. Those who completed the assessment procedure reacted quite favorably, while those who performed only part of it reacted unfavorably. Nevertheless, we do not know for sure whether or for how long this differential attitude endured.

Sixth, in light of the favorable reaction to the assessment procedure, it would seem worthwhile to develop it further. Specifically, it would be most exciting to automate the procedure and to operate on-line and interactively with a computer. Preliminary investigation has shown that all of the hardware and most of the software required to automate the procedure already exists. It should not be a terribly difficult task, and, if successfully accomplished, it would ease considerably the burden of formulating and validating an assessment algorithm.

Finally, the various applications of the assessment procedure suggested in Chapter IV might be investigated formally.

CHAPTER VI

CRITICISMS, OVERALL CONCLUSIONS, AND FINAL INTERPRETATIONS

6.1.0 INTRODUCTION

The purpose of this chapter is to bind together previous results, conclusions, and interpretations. Both the experiment and the complete assessment procedure will be subjected to critical review. Overall conclusions will then be drawn. The thesis will close with a final interpretation.

6.1.1 A Critical Review Of The Experiment

A rather detailed discussion of the experimental procedure was carried out in Chapter V. Analyses were described, results were presented, interpretations were made, and conclusions were drawn. None of these things will be repeated in any detail here. However, the discussion in Chapter V was pointed toward coming up with positive statements about the questions originally raised and tested. Little attention was paid to possible design flaws, errors in implementation, unexplainable results, or restrictions on conclusions. In this section we shall take a more negative point of view and attempt to point out some of the weaknesses inherent in the experiment.

Consider first the design. The omission of a fourth experimental group to control for raw information has already been discussed. The absence of any long-term impact measures has also been discussed. However, there are two other design problems which deserve attention before drawing final conclusions. These involve:

1. procedural error control;

2. differential treatment of the various groups during introduction of both ordinal and cardinal guidance.

These two issues will be discussed in order.

Whenever any empirical research is undertaken, there always lurks the danger of obtaining erroneous observations, transcribing data incorrectly from one medium to another and applying invalid computational routines to generate results. If a large amount of data is involved, and if manipulation of these data is entrusted to an electronic computer, these problems can become quite serious. Several devices were used to control for errors, and these will be described below.

1. Every single observation obtained from subjects over the course of the entire experiment was inspected visually by the writer minutes after it had been generated. Whenever missing or logically impossible observations were detected, the particular subject involved was approached personally and requested to amend his response accordingly. This device, along with the extremely cooperative attitude on the part of all subjects, accounted for the low number of missing observations. It also contributed to the cleanliness of results and the validity of conclusions. However, it did not provide any direct protection against observations which were present and logically possible, but just not what the subject intended.

2. Partial control over unintended observations was provided by the "exception analyses" described in Section 5.3.3. However, this was only partial.

3. Transcription errors were partially controlled by reducing the number of transfers involved to two (i.e., written questionnaire or tape recording to teletype console, and teletype console to permanent tape record), by having a single individual perform all of these operations personally, and by pre-programming on-line editing routines. However, these devices were effective primarily against logically impossible observations. The only evidence of their effectiveness against logically possible, but incorrect transcriptions is a brief, quasi-random error check performed by the writer after all analyses had been performed. On the basis of this, the writer would estimate a transcription error rate of something less than one percent.

4. Faulty programming constitutes another source of possible error. Several thousand lines of computer code were written to manipulate data and compute results. In every case, test data were used to validate program segments, but there still remains the possibility of undetected bugs.

A second design problem involves the differential treatment of various groups during introduction of ordinal and cardinal guidance. Recall that neutral tasks were given to certain groups as a control against possible "Hawthorne" effects. But these neutral tasks were in the form of written questionnaires, while both ordinal and cardinal guidance were administered through personal interviews. It is conceivable that some of the differences which emerged between groups can be attributed to this differential mode of administration rather than to the differential content of the tasks. Were it not for the stringent scheduling constraints, this might have been controlled by administering neutral tasks under identical interview conditions. Unfortunately, however, this was not feasible. It is possible that this made an important difference, although there were no direct indications to support such a conclusion.

Consider next the issue of generalizing results and conclusions. It has already been pointed out that the total sample was quite homogeneous in many respects. For purposes of obtaining and analyzing results, this was a blessing. However, for purposes of drawing conclusions, this imposes an important restriction. It is not at all clear that other kinds of people would respond as these subjects did to the various experimental manipulations. Furthermore, it is not even clear that these same subjects would respond similarly outside of the idyllic atmosphere provided by the training course. In particular, it is not clear that real decision makers would react as favorably to the complete assessment procedure if it were introduced to them "on the job". This is because "real" and "important" decisions are

made in an atmosphere rife with political overtones. There frequently exist strong interest groups, and there is always a large stake riding on "important" outcomes. Since none of these factors were present during the experiment, no conclusions can be drawn concerning their impact. In Section 5.5.4, it was suggested that group decision processes be investigated using essentially the same techniques applied in this experiment. If this were done, and if the investigation were carried out "in the field" (as opposed to the laboratory atmosphere characterizing this experiment), the impact of political factors could be determined.

On the other hand, the writer is not particularly concerned with generalizing conclusions to cover either people or situations involving only trivial decisions. From the point of view of the assessment procedure, this is of no concern whatsoever. The time and effort required to create a complex assessment algorithm would not justify its meager rewards in the case of trivial decisions. From the point of view of learning about and describing decision processes, this would be interesting; but the writer's primary research interest lies with "important" decisions and with decision makers predisposed to using "rational" techniques.

6.1.2 Overall Conclusions

Subject to the preceeding criticisms, the writer would draw the following overall conclusions from the experiment.

1. Almost any conscious activity which a decision maker perceives as relevant to making a choice will serve to clarify his preferences for alternatives. In particular, providing factual information, requiring that he articulate and structure assessment criteria, requiring that he quantify his preferences, and the act of making a final choice all have this effect. The mere realization that a choice must be made, accompanied by preliminary efforts to structure the alternatives, has the same effect. However, if a decision maker does not perceive such activities to be relevant, even though they are alleged to be, clarification will not occur. When clarification does occur, its magnitude varies with the particular type of activity engaged in. Of critical importance is any kind of activity which challenges and thereby tests the validity of existing preferences.

2. The number of preference discriminations spontaneously made by decision makers among well-defined alternatives depends primarily upon individual factors. Changes thereto induced by various conscious activities also depend upon individual factors.

3. Almost any conscious activity perceived as relevant to the decision will also increase a decision maker's confidence in the accuracy of his preferences. In particular, the five types of activity mentioned above (i.e., the five experimental factors) will have this effect. Irrelevant

activities will not have this effect, apart from a slight "momentum" phenomenon. Again, the magnitude of this effect depends upon the particular type of activity.

4. The same conclusions concerning clarification apply to the satisfaction derived by decision makers from undertaking various conscious activities. Satisfaction, here, refers to the degree to which such activities are perceived as helpful to improving the quality of the final choice.

5. Although decision makers may receive clarification, satisfaction, and additional confidence from undertaking various conscious and relevant activities, this does not guarantee that they will overtly alter prior preference committments in light of newly-perceived implications. Once again, provision of a challenge or validity check is of critical importance. If such a check is performed, then overt committment will generally follow.

6. On the other hand, changes in preference will occur covertly following almost any conscious and relevant activity, but will not occur (apart from random instability) unless such activity is perceived as relevant. The magnitude of such changes decreases steadily as confidence and clarification increases and as the moment of final decision draws near.

7. Without knowing precisely what his previous preferences were, a decision-maker will tend to underestimate their temporal stability. He will also tend to underestimate the magnitude of changes in stability over time. Both of these phenomena become less pronounced if he makes a definite and overt committment to a particular preference structure.

8. The perceived value of engaging in various conscious activities compared to the time and effort expended depends critically upon the type of activity engaged in. If the activity is perceived as irrelevant, it is considered a great waste of time. However, even if the activity is perceived as relevant, it may not be considered sufficiently valuable to justify the time and effort expended. Once again, providing a challenge or validity check is particularly important in this respect.

9. The complete assessment procedure developed in Chapter II has four important impacts upon decision processes.

    a. Its primary impact is to induce decision makers to formulate and validate a consistent assessment structure. Validation is provided by comparing computed with subjective preferences, and the quantitative aspects of the procedure are critical in this respect.

7. Without knowing precisely what his previous preferences were, a decision-maker will tend to underestimate their temporal stability. He will also tend to underestimate the magnitude of changes in stability over time. Both of these phenomena become less pronounced if he makes a definite and overt committment to a particular preference structure.

8. The perceived value of engaging in various conscious activities compared to the time and effort expended depends critically upon the type of activity engaged in. If the activity is perceived as irrelevant, it is considered a great waste of time. However, even if the activity is perceived as relevant, it may not be considered sufficiently valuable to justify the time and effort expended. Once again, providing a challenge or validity check is particularly important in this respect.

9. The complete assessment procedure developed in Chapter II has four important impacts upon decision processes.

    a. Its primary impact is to induce decision makers to formulate and validate a consistent assessment structure. Validation is provided by comparing computed with subjective preferences, and the quantitative aspects of the procedure are critical in this respect.

b. In the process of formulating an assessment structure, preferences for alternatives are significantly altered. However, they are not altered randomly, but rather in a manner directed toward the final choice.

c. If the entire procedure is followed, particularly the final steps of quantitative assessment, a mechanism is provided to validate preferences. This, in turn, is largely responsible for inducing favorable reactions on the part of decision makers. It is also responsible for inducing at least intermediate-term changes both in attitudes toward the procedure and in preferences for alternatives. On the other hand, if only part of the procedure is followed, the reaction of decision makers is nowhere near as favorable nor as permanent.

d. Another important impact is to measure and display assessment criteria, which can be useful both for purposes of normative decision making and for purposes of scientific description.

10. These conclusions are applicable to adult, highly educated, highly competent, highly motivated, and well-trained professional decision makers. Furthermore, they apply to individual decision processes conducted under ideal laboratory conditions. Further research is required to extend the range of applicability either to different kinds of people or to different situations. The writer would be particularly skeptical about generalizing these conclusions to situations characterized by substantial conflict, strife, or political combat. These conclusions may also be inapplicable to individuals who do not characteristically possess the courage of their convictions or, perhaps, no convictions at all.

6.1.3 A Critical Review Of The Assessment Procedure

In this section, we shall return briefly to the assessment procedure. Critical scrutiny will be directed toward the methodology in an attempt to pinpoint "soft spots". Three types of criticisms will be made. First, important substantive issues omitted or accorded only a cursory treatment will be recalled. Second, attention will be focused upon underlying assumptions which seem particularly questionable. Finally, some of the more difficult procedures will be reviewed.

The reader may have noticed that several important aspects of the overall task of assessment were either ignored completely or else given only a cursory treatment. Methodological issues falling into this category include:

1. the problem of describing adequately and accurately the job to be performed by whichever alternative is finally selected -- this was ignored completely;

2. the problem of producing alternatives to accomplish the stated job -- this was also completely ignored;

3. the problem of predicting both performance and resource consequences associated with each produced alternative -- very little was said about this issue;

4. the problem of validating the descriptive accuracy of performance and resource estimates -- this was ignored completely;

5. the problem of establishing feasibility constraints (i.e., mandatory performance and/or resource requirements) on alternatives -- this issue was also ignored;

6. the problem of incorporating risk/uncertainty considerations -- this was discussed only briefly;

7. the problem of defining a decision rule -- this also was discussed only briefly;

8. the problem of selecting appropriate personnel to assess alternatives and to make a final choice -- except to point out that final results could depend critically upon both the identity of decision makers and the point in time when an assessment is made, this issue was largely ignored.

Now it is not claimed that the above issues are unimportant. Quite to the contrary, they are all very important, and they deserve the same amount of attention accorded to worth assessment. However, the scope of this thesis was not intended to cover these issues, except insofar as they provided a context in which to discuss worth assessment.

Three critical assumptions about the manner in which decision makers can be induced to conceptualize worth notions deserve special attention. First, is it easy for decision makers to formulate objectives and to map out a complete criterion hierarchy? From the experiment we know that high-calibre, professional decision makers can, without much difficulty, but this does not mean that everybody can.

Second, can decision makers conceive of and articulate meaningfully the notion of proportional satisfaction of a stated objective or criterion? Specifically, can they distinguish easily between this problem and the problem of comparing alternatives with one another and assessing them on a purely relative scale? The experiment demonstrated that this is a somewhat difficult distinction to make at the outset. It also demonstrated that the first problem is a more difficult one to solve than the second, even after they have been distinguished. However, at least for the kinds of decision makers studied, this is possible.

Third, can decision makers be comfortable with a linear weighting scheme? A typical first reaction to this question is negative, on the grounds that strict linearity is too simple and too restrictive. However, after realizing that linearly weighted performance criteria do not necessarily imply an assessment algorithm linear in performance measures (recall that scoring functions can assume any desirable non-linear shape), decision makers will generally retract their objection. At least this is what occurred during the experiment.

This brings us to the question of difficult assessment procedures. The most difficult one by far was shown (by the experiment) to be identification of worth-interdependence among performance criteria. Without independence, the linear weighting assumption is highly suspect. The writer will readily admit that this is a difficult question to understand and an even more difficult one to answer. The current procedure is not completely satisfactory in this regard. Additional research might profitably be directed toward reformulating and illustrating this issue more clearly.

Another difficult procedure involves definition and selection of physical performance measures. This requires some ingenuity and, therefore, may restrict the procedure's applicability to sophisticated decision makers.. However, provision of a master list does help a great deal. This was demonstrated in the experiment (biographical information constituted the master list).

Adjusting the "effective" weights to account for differential quality among performance measures is not a difficult procedure to implement, but there is some question in the writer's mind whether this is the proper way to handle the problem. Critics of the procedure have been similarly skeptical. Additional research might profitably be directed toward this problem also.

6.1.4 Overall Conclusions

Subject to the preceeding criticisms, the writer would draw the following conclusions about the assessment procedure.

1. It can be carried out successfully by professional decision makers in an ideal laboratory setting. Whether it can or should be carried out by other types of people or in other types of settings is not yet clear.

2. Although reactions to the procedure on the part of decision makers were shown to be quite favorable, no claim is made that alternative procedures could not produce the same effects. In other words, the particular procedure developed herein is not claimed to be unique. However, it is claimed to be sufficient in these respects.

3. The reader should bear in mind that the complete procedure has two distinct phases. The first phase induces decision makers to formulate, to articulate, to validate, and to map out their assessment structure. The primary advantages of the procedure lie in this first phase -- particularly in

its quantitative aspects, which permit internal validation of prior preferences. The second phase involves application of the assessment algorithm so generated to specified alternatives. This can be useful, although it is not essential, to guiding a final choice.

4. The procedure is not completely free of conceptual problems or difficult questions. Further research would probably alleviate some of these difficulties.

5. The procedure is definitely not "objective" in the sense of eliminating the need for human judgement. Quite to the contrary, its basic purpose is to systematize subjective judgement in ways which decision makers will find helpful.

6. A second purpose is to provide an instrument for measuring worth notions. This can be useful in scientific research.

7. The procedure is completely general with respect to type of decision problem and type of objective, criterion, and performance measure.

8. However, in its present form, the procedure is restricted to assessing the worth of discrete alternatives. It was not intended to facilitate the design of alternatives.

6.1.5 A Final Interpretation

This thesis opened with a lengthy discussion of decision making in general and the assessment of worth in particular. A procedure was devleoped to aid in this process. A live example was presented to illustrate the procedure, and its relationship to various disciplines was discussed. A dual-purpose experiment was then described, results were reported, and conclusions were drawn about the normative efficacy of the procedure and about its descriptive impact on individual decision processes. Before closing, it would be appropriate to address a most controversial and, therefore, a most difficult issue -- the issue of relevance.

The issue is simply this. How relevant are the questions posed by the assessment procedure and tested by the experiment? Should the effectiveness of a decision and, therefore, of a decision making technique be evaluated on the basis of the decision maker's (or somebody else's) personal reaction thereto? With one exception, the writer's answer is yes! The design of the assessment procedure and the questions tested by the experiment clearly reflect this position. However, let us investigate some alternative points of view.

The position taken implicitly by many organizations (explicitly by religious organizations) is that the "right" values, the "right" objectives, and the "right" assessment criteria have already been determined (e.g., the Ten Commandments). From their essentially practical point of view, the "problem" facing members of the organization is to become aware of these already determined objectives and to interpret

them correctly in a particular choice situation. Formulating objectives is a problem (really a privilege) reserved for higher authority. In business organizations, the "right" values are encapsulated in "corporate objectives". For government organizations "national policy" plays a similar role. In all cases, however, the basic position remains unchanged. If this position is adopted, then the effectiveness of decisions and decision making techniques would be assessed according to how well decision consequences do in fact (and after the fact) satisfy these "right" objectives. This appears to be inconsistent with the writer's point of view. But let us take a closer look.

First, all objectives (either "right" or "wrong") exist in the minds of human beings. There may be some argument concerning where they came from originally (a philisophical and religious problem), how they are to be maintained and applied (a legal problem), or whose objectives will prevail at any given time (a political problem), but it is meaningless to talk of objectives apart from their human possessors. Therefore, evaluating a decision in terms of its consequences amounts ultimately to assessing the responses of some person or group of persons. Defining a set of "right" objectives serves as a convenient fiction both to expedite the decision making process and, much more importantly, to facilitate social control.

Second, even if we wish to accept the above fiction, there still remains the task of operationalizing these "right" objectives in a specific decision context. The assessment procedure developed herein is perfectly compatible with this point of view, since the "right" objectives could be inserted mechanically at the top level of the criterion hierarchy, and the procedure could be followed from there. To make the procedure more palatable, the concept of formulating objectives would be replaced by the concept of interpreting already established objectives in terms of a particular choice situation. However, apart from relabelling, no essential changes in the procedure itself would be required.

Third, whether or not we accept the fiction of "right" objectives, real organizations and real decisions require that somebody (usually called a policy maker) establish and continually update broad guidelines for practical decision making. The survival of the organization depends, in part, upon the success of this process. The assessment procedure developed herein is particularly well suited for policy makers; and, the writer suspects, these people would be the first to claim that their personal reactions provide the best means of assessing its efficacy.

A somewhat different point of view is taken by normative decision theory. Here, concern is neither with the people who make decisions nor with the particular objectives they attempt to satisfy. Attention is focused primarily upon the formal process by which

conclusions are drawn from assumptions. The assumptions include all facts relating to the decision situation, as well as the objectives which are to be satisfied. Conclusions relate to choosing a particular alternative. A set of procedures is formulated to guide the movement from assumptions to conclusions (i.e., from objectives and assumed factual relationships to a final choice). The purpose of these procedures is to obtain a result which is optimum (or at least satisfactory) in terms of the objectives. The effectiveness of decisions is then assessed according to how "rationally" they were carried out (i.e., how closely they followed these optimizing procedures).

Now the writer's position is not inconsistent with the normative point of view either. Quite to the contrary, they are closely interrelated. The assessment procedure was originally designed with a decidedly normative purpose -- to provide inputs into normative decision models. For statistical decision theory, it establishes, in part, the decision maker's loss structure. For operations research techniques, it helps define the objective function which is to be optimized. As such, the assessment procedure serves to operationalize some of the assumptions of normative decision models by inducing decision makers to formulate explicitly what their objectives really are. Without such information, normative models are inapplicable to practical problems.

Then there is the descriptive point of view. In its purest form, it attempts only to understand, to explain, and to predict actual choice behavior. To aid in this process, concepts are formed, hypotheses are formulated, and whole theories are constructed. The descriptive decision theorist is concerned with testing these theories, and he assesses the effectiveness of a decision model according to how well it provides insight into the decision process and how well it generates predictions which agree with actual behavior. Concern with objectives is purely descriptive. That is, can they be described with sufficient adequacy and accuracy to permit correct predictions of choice behavior? If not, the entire concept of an objective may be discarded as useless. Concern with reasoning processes is similarly proscribed. Whether or not these processes violate rules of correct reasoning is only important as a descriptive fact and as an aid to understanding and predicting behavior.

The writer's position is not inconsistent with this point of view either; but there are distinct differences in emphasis. First, the writer has a broader objective than mere prediction. He would like to provide decision makers with a tool of some practical value from <u>their</u> point of view. Descriptive methods (i.e., scientific research and experimentation) are used heavily to determine what their point of view is and whether it has been satisfied. In addition, substantial efforts are made to understand, to explain, and to predict choice behavior. But an important and independent purpose is to generate a procedure which possesses practical value.

Second, the writer has a narrower objective than traditional theory testing. Rather than attempting to confirm or deny an existing theory, the writer formulated a set of questions specifically related to the assessment procedure itself. Answers to these questions are of substantive interest from the descriptive point of view, but not solely (or even primarily) because of their relationship to traditional theory.

Third, there is descriptive value in the assessment procedure as a psychological measuring instrument. In particular, it may be used to define operationally the concept of worth. It could then be used to test descriptive decision theories (e.g., expected utility maximization) and to conduct social research (e.g., comparative studies of cultures, organizations, small groups, etc.).[1] Herein lies the single exception previously mentioned. In its role as a psychological measuring instrument, predictive efficacy -- not personal reactions on the part of decision makers -- would be the appropriate criterion for assessing the procedure.

In short, the writer has adopted none of these three points of view to the exclusion of the others; but he has attempted to remain consistent with all of them. Without normative guidance and descriptively accurate information, practical decisions are frequently ineffective. Without practical problems to solve and descriptive devices to gather facts and predict consequences, normative guidance

---

1. A variant of the assessment procedure has been used quite successfully by the writer and one of his colleagues to study comparative sources of job satisfaction in various work groups. The results of this research will be reported in a separate paper.

is just plain vacuous. Without important applications and a normative incentive, descriptive results are uninteresting. In light of these observations, the writer has attempted to synthesize a product combining the strengths of all three points of view without falling prey to the excesses of any one. Whether or not and to what extent this has been accomplished the reader must judge for himself.

REFERENCES

1. Churchman, C.W., Prediction and Optimal Decision: Philosophical Issues of a Science of Values, Prentice Hall, Englewood Cliffs, 1961.

2. Churchman, C.W., et al., Introduction To Operations Research, John Wiley and Sons, New York, 1959.

3. Cronbach, L.J., and Gleser, G.C., Psychological Tests and Personnel Decisions, University of Illinois Press, Urbana, 1965.

4. Grayson, C.J., Jr., Decisions Under Uncertainty: Drilling Decisions by Oil and Gas Operators, Division of Research, Graduate School of Business Administration, Harvard University, Boston, 1960.

5. Hays, W.L., Statistics for Psychologists, Holt, Rinehart and Winston, New York, 1963.

6. Homans, G.C., Social Behavior: Its Elementary Forms, Harcourt, Brace and World, New York, 1961.

7. March, J.G., and Simon, H.A., Organizations, John Wiley and Sons, New York, 1958.

8. Miller, J.R., A Conceptual Framework for Analyzing Military Systems, Working Paper Number 6682, The MITRE Corporation, 1963.

9. Miller, J.R., The Problem of Selecting Electronic Data Processing Equipment, Working Paper Number 7724, The MITRE Corporation, 1965.

10. Miller, J.R., Two Theories of Social Interaction, unpublished seminar report, 1964.

11. Miller, J.R., and Wolfe, C.F., Making Cost Effectiveness Trade-Offs in EDP System Selection, Technical Memorandum Number 4032, The MITRE Corporation, 1964.

12. Raiffa, H., and Schlaifer, R., Applied Statistical Decision Theory , Division of Research, Graduate School of Business Administration, Harvard University, Boston, 1961.

13. Rosenthal, S., "Analytical Technique For Automatic Data Processing Equipment Acquisition", 1964 Spring Joint Computer Conference, 1964.

14. Schlaifer, R., Probability and Statistics for Business Decisions, McGraw-Hill Book Company, Inc., New York, 1959.

15. Smith, N.M., Jr., "A Calculus for Ethics: A Theory of the Structure of Value", Behavioral Science, 1956 (1), pp. 111-142 and pp.186-211.

16. Von Neumann, J., and Morgenstern, O., Theory of Games and Economic Behavior, Princeton University Press, Princeton, 1944.

APPENDIX I

SCORING PROCEDURE 1

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. discrete scale;
2. two-level scale;
3. level 1 = absence of some desirable performance attribute;
4. level 2 = presence of that desirable attribute.

Step 1. Assign zero worth points to absence of the desirable attribute.

Step 2. Assign one worth point to presence of that desirable attribute.

This completes the procedure.

APPENDIX II

SCORING PROCEDURE 2

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. discrete scale;
2. two-level scale;
3. level 1 = absence of some desirable performance attribute;
4. level 2 = presence of that desirable attribute in conjunction with some qualitative measure of relative worth, when present.

Step 1. Assign zero worth points to absence of the desirable attribute.

Step 2. Identify all feasible alternatives which promise the desirable attribute.

Step 3. Assemble one or more decision makers.

Step 4. After discussing collectively the various merits of the desirable attribute - why it is important and what benefits its presence provides - have each decision maker make a separate and independent judgment of the extent to which each feasible alternative's promised attribute satisfies the related lowest-level performance criterion. All judgments will be recorded by assigning a number between zero and one indicating the proportional satisfaction provided by each feasible alternative.

Step 5. To determine each feasible alternative's score on this performance measure, assign either zero points (if the attribute is absent) or the arithmetic mean (possibly weighted) of the individual scores assigned judgmentally by separate decision makers.

This completes the procedure.

APPENDIX III

SCORING PROCEDURE 3

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. discrete scale;
2. three, four, or five levels on the scale;
3. all levels constitute strictly nominal categories.

Step 1. Since the scale of the physical performance measure is strictly nominal, a preference or worth ordering must be placed directly on the nominal categories. This will be accomplished by means of the same ranking procedure used in defining weights and presented in section 2.2.4.

Step 2. Assemble one or more decision makers.

Step 3. After discussing collectively the various merits of nominal categories, have each decision maker perform a separate and independent rank-ordering of the various categories. This may be accomplished by performing Steps 4 through 7 below.

Step 4. List the nominal categories in approximate order of decreasing worth, starting with the category perceived as most valuable at the top of the list. The category perceived as least valuable should appear at the bottom of the list. It is not necessary to have the categories perfectly ranked or ordered on this first pass, since subsequent operations will be performed to guarantee complete ordering.

Step 5. Compare the first two categories on the list.

a. If the first category is perceived as more valuable than the second, proceed directly to Step 6.

b. If both categories are perceived as roughly equal in worth, proceed directly to Step 6.

c. If the second category is perceived as more valuable than the first, invert their positions on the list (i.e., place the first category where the second used to be on the list, and vice versa), and then proceed to Step 6.

Step 6. Compare the lower-ranked category from Step 5 with the next-lower category on the list. Repeat the comparisons and stipulated operations in Step 5 on this new pair of categories. Continue in this manner all the way down to the end of the list until pair-wise comparisons have been made between all contiguous criteria.

Step 7. After the list has been completely exhausted, go back and determine whether any inversions (position changes) occurred. If none occurred, proceed directly to Step 8. If one or more occurred, return to the head of the list, and repeat the entire procedure described in Steps 5 and 6.

Step 8. Next, have each decision maker make a separate and independent judgment of the extent to which each ranked category satisfies the related lowest-level criterion. All judgments will be recorded by assigning a number between zero and one indicating the proportional satisfaction provided by each scale category.

Step 9. Finally, to determine each nominal category's point score, compute and record the (possibly weighted) arithmetic mean of the individual category scores assigned by separate decision makers in Step 8.

This completes the procedure.

APPENDIX IV

SCORING PROCEDURE 4

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. discrete scale;
2. three, four, or five levels on the scale;
3. the scale is ordered.

Step 1. List the ordered levels in a single column.

Step 2. Inspect the level appearing at the head of the column. Is that the most preferred or the least preferred level? If most preferred, proceed to Step 4. If least preferred, proceed to Step 3.

Step 3. Invert the column, and list the levels again - this time in reverse order. Proceed to Step 4.

Step 4. The discrete levels should now be listed in perfect order of descending relative worth. Inspect to verify that this is true. If so, proceed to Step 5. If not, check earlier steps to insure that no errors occurred. If no errors occurred, this particular performance measure cannot be handled by the scoring procedures presented herein. Define a new measure, and return to Step 1 of the questioning procedure in section 2.2.5.1.3.

Step 5. Assemble one or more decision makers.

Step 6. After discussing collectively the various merits of nominal categories and (hopefully) agreeing on their rank-order, have each decision maker record a separate and independent judgment of the extent to

which each nominal category satisfies the related lowest-level performance criterion. All judgments will be recorded by assigning a number between zero and one indicating the proportional satisfaction provided by each scale category.

Step 7. To determine each nominal category's score, compute and record the (possibly weighted) arithmetic mean of the individual category scores assigned by separate decision makers.

This completes the procedure.

APPENDIX V

SCORING PROCEDURE 5

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. discrete scale;
2. more than five levels on the scale;
3. all levels constitute strictly nominal categories.

Step 1. Since the scale of the physical performance measure is strictly nominal, a preference or worth ordering must be placed directly on the nominal categories. This will be accomplished by means of the same ranking procedure used in defining weights and presented in section 2.2.4.

Step 2. Assemble one or more decision makers.

Step 3. After discussing collectively the various merits of nominal categories, have each decision maker perform a separate and independent rank-ordering of the various categories. This may be accomplished by performing Steps 4 through 7 below.

Step 4. List the nominal categories in approximate order of decreasing worth, starting with the category perceived as most valuable at the top of the list. The category perceived as least valuable should appear at the bottom of the list. It is not necessary to have the categories perfectly ranked or ordered on this first pass, since subsequent operations will be performed to guarantee complete ordering.

Step 5. Compare the first two categories on the list.

a. If the first category is perceived as more valuable than the second, proceed directly to Step 6.

b. If both categories are perceived as roughly equal in worth, proceed directly to Step 6.

c. If the second category is perceived as more valuable than the first, invert their positions on the list (i.e., place the first category where the second used to be on the list, and vice versa), and then proceed to Step 6.

Step 6. Compare the lower-ranked category from Step 5 with the next-lower category on the list. Repeat the comparisons and stipulated operations in Step 5 on this new pair of categories. Continue in this manner all the way down to the end of the list until pair-wise comparisons have been made between all contiguous criteria.

Step 7. After the list has been completely exhausted, go back and determine whether any inversions (position changes) occurred. If none occurred, proceed directly to Step 8. If one or more occurred, return to the head of the list, and repeat the entire procedure described in Steps 5 and 6.

Step 8. Inspect adjacent pairs of ranked scale categories. Locate that adjacent pair of scale categories which seem closest to one another in terms of their perceived worth (i.e., locate the most equally valuable adjacent pair of scale categories). Collapse these two categories into a single category.

Step 9. Repeat Step 8 as many times as is required to reduce the number of resulting categories to five. Then proceed to Step 10.

Step 10. Next, have each decision maker record a separate and independent judgment of the extent to which each ranked category satisfies the related lowest-level criterion. All judgments will be recorded by assigning a number between zero and one indicating the proportional satisfaction provided by each scale category.

Step 11. Finally, to determine each nominal category's point score, compute and record the (possibly weighted) arithmetic mean of the individual category scores assigned by separate decision makers in Step 10.

This completes the procedure.

APPENDIX VI

SCORING PROCEDURE 6

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. bounded from above by some finite positive number;
4. direct preference relationship;
5. worth score zero assigned to zero performance;
6. worth score one assigned to performance at the logical upper bound;
7. constant rate of change of worth with increases in performance.

The above seven characteristics describe completely a linear scoring function passing through the origin and with positive slope equal to the reciprocal of the logical upper bound. The equation of this scoring function is

$$\text{worth score} = \frac{\text{Measured Performance}}{\text{Logical Upper Bound}}$$

A graphical picture of this scoring function appears below in Figure 1.

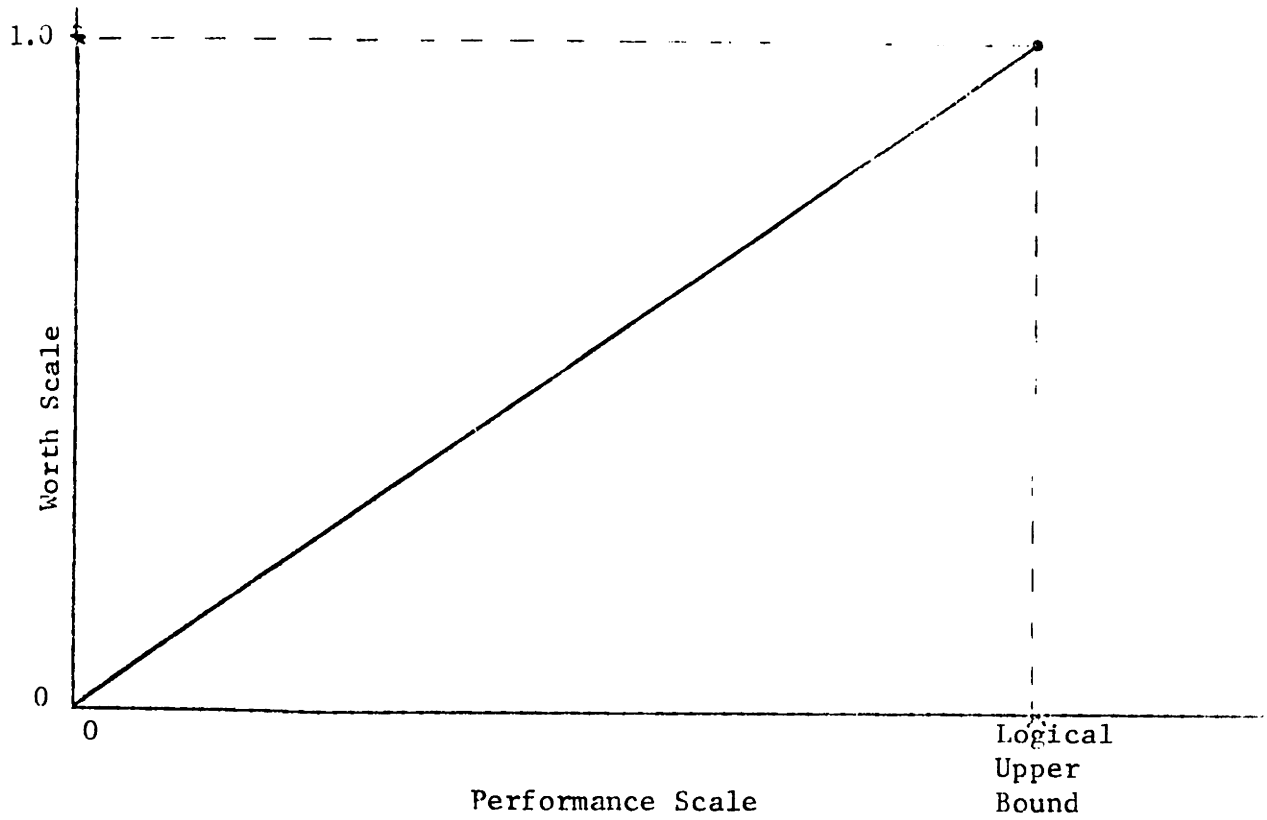


FIGURE 1

This completes the procedure.

APPENDIX VII

SCORING PROCEDURE 7

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. bounded from above by some finite positive number;
4. direct preference relationship;
5. worth score zero assigned to zero performance;
6. worth score one assigned to performance at the logical upper bound;
7. uniformly accelerating rate of change of worth with increases in performance.

A graphical picture of this general shape of scoring function appears below in Figure 2.

Step 1. At this point, decision makers have two choices. The simplest procedure would be to fit a standardized quadratic scoring function to the performance measure under the following stipulated assumptions.

1. The scoring function is quadratic with positive second derivative (indicating uniform acceleration).
2. The minimum of the quadratic function falls exactly at the origin.
3. The upper tail of the scoring function passes through the point whose coordinates are (performance = logical upper bound, worth score = one).

These three assumptions completely determine a scoring function (see Figure 2) whose equation is

$$\text{worth score} = \left(\frac{\text{Measured Performance}}{\text{Logical Upper Bound}}\right)^2.$$

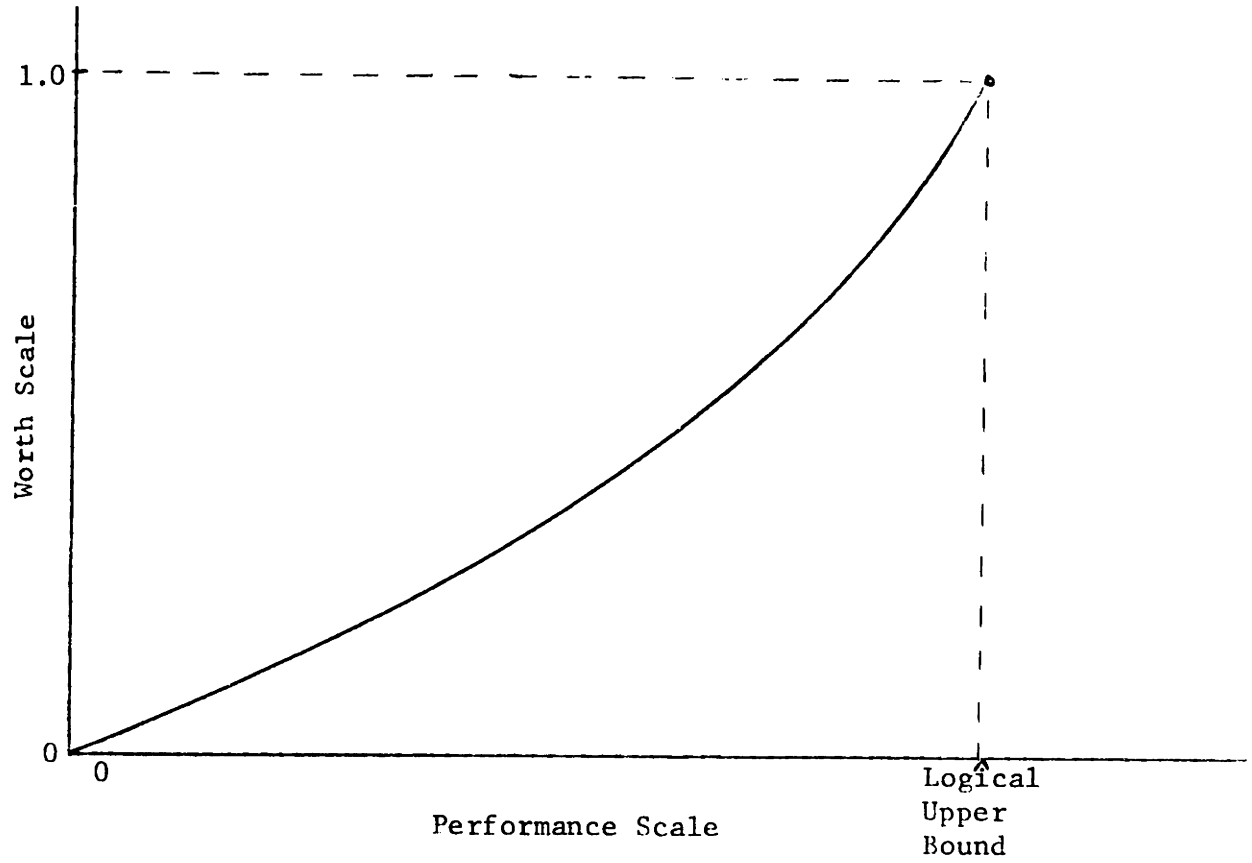


FIGURE 2

To determine whether or not this looks like an appropriate scoring function, it is suggested that a sheet of standard graph paper be procured and that the above equation be plotted thereupon. Five or six representative points should be sufficient to grasp the exact shape of the function and to decide whether or not it seems appropriate. If yes, this completes the procedure. If no, proceed to scoring procedure 20.

APPENDIX VIII

SCORING PROCEDURE 8

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. bounded from above by some finite positive number;
4. direct preference relationship;
5. worth score zero assigned to zero performance;
6. worth score one assigned to performance at the logical upper bound;
7. uniformly decelerating rate of change of worth with increases in performance.

A graphical picture of this general shape of scoring function appears below in Figure 3.

Step 1. At this point, decision makers have two choices. The simplest procedure would be to fit a standardized quadratic scoring function to the performance measure under the following stipulated assumptions.

1. The scoring function is quadratic with negative second derivative (indicating uniform deceleration).
2. The maximum of the quadratic function falls exactly at the point whose coordinates are (performance = logical upper bound, worth score = one).

3. The quadratic function passes through the origin.

These three assumptions completely determine a scoring function (see Figure 3) whose equation is

$$\text{worth score} = 2\left(\frac{\text{Measured Performance}}{\text{Logical Upper Bound}}\right) - \left(\frac{\text{Measured Performance}}{\text{Logical Upper Bound}}\right)^2.$$

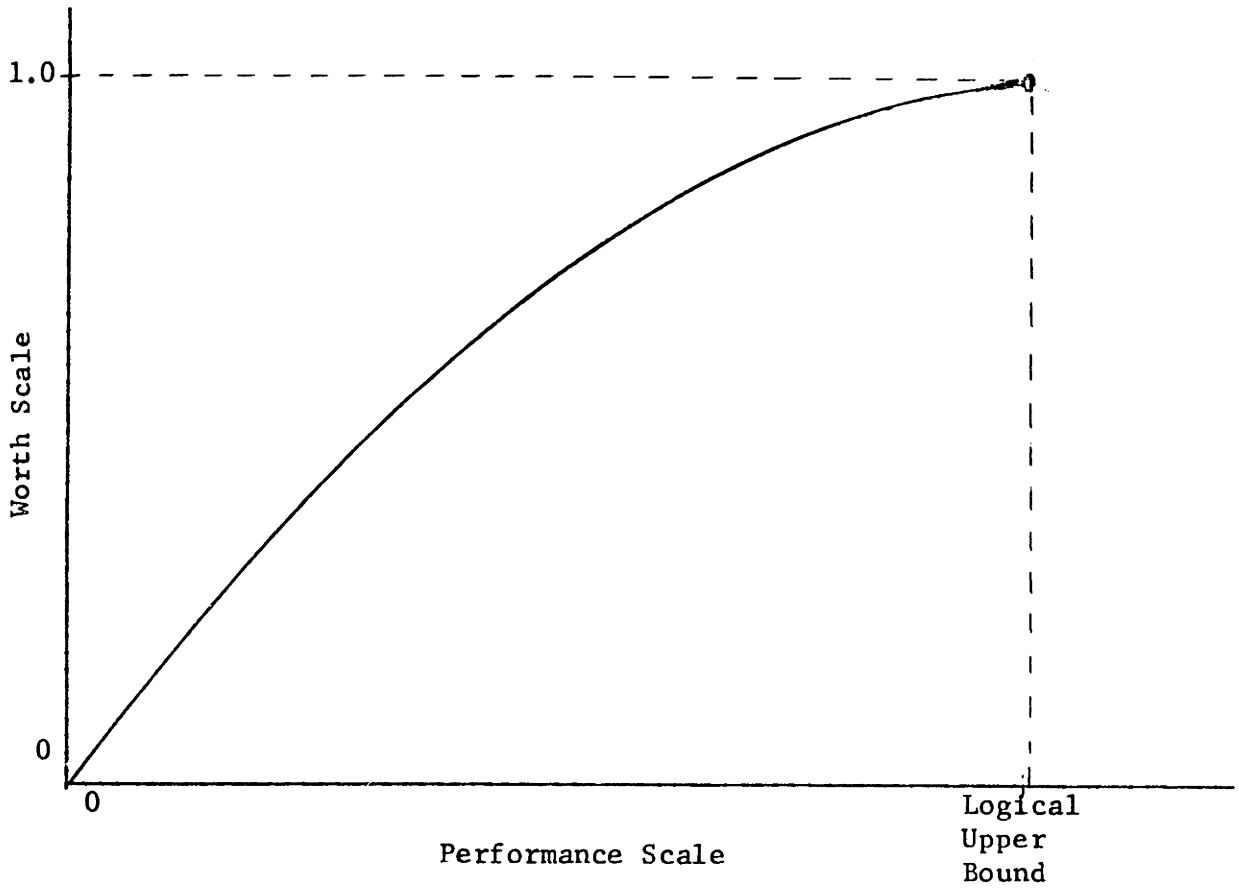


FIGURE 3

To determine whether or not this looks like an appropriate scoring function, it is suggested that a sheet of standard graph paper be procured and that the above equation be plotted thereupon. Five or six representative points should be sufficient to grasp the exact shape of the function and to decide whether or not it seems appropriate. If yes, this completes the procedure. If no, proceed to scoring procedure 20.

APPENDIX IX

SCORING PROCEDURE 9

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. bounded from above by some finite positive number;
4. direct preference relationship;
5. worth score zero assigned to zero performance;
6. worth score one assigned to performance at the logical upper bound;
7. first accelerating, then decelerating rate of change of worth with increases in performance.

A graphical picture of this general shape of scoring function appears below in Figure 4.

Step 1. At this point, decision makers have two choices. The simplest procedure would be to fit a standardized cosine function to the performance measure whose equation is

$$\text{worth score} = \frac{1}{2} - \frac{1}{2} \text{ cosine } \left[\pi \left(\frac{\text{Measured Performance}}{\text{Logical Upper Bound}}\right)\right],$$

where $\pi = 3.1416$,

and cosine values may be looked up in a trigonometric table (function expressed in terms of radians) or computed on an engineering slide rule.

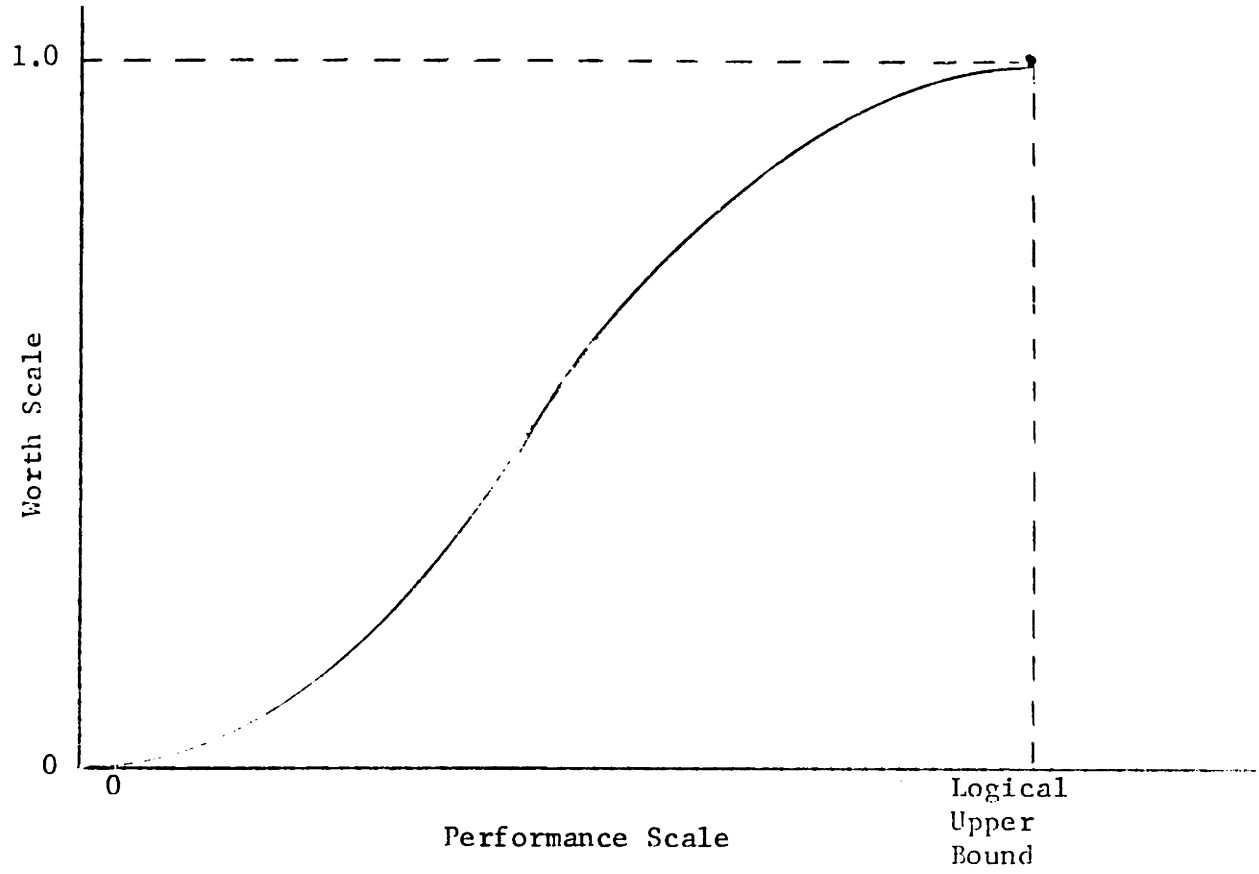


FIGURE 4

To determine whether or not this looks like an appropriate scoring function, it is suggested that a sheet of standard graph paper be procured and that the above equation be plotted thereupon. Five or six representative points should be sufficient to grasp the exact shape of the function and to decide whether or not it seems appropriate. If yes, this completes the procedure. If no, proceed to scoring procedure 20.

APPENDIX X

SCORING PROCEDURE 10

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. bounded from above by some finite positive number;
4. direct preference relationship;
5. worth score zero assigned to zero performance;
6. worth score one assigned to performance at the logical upper bound;
7. first decelerating, then accelerating rate of change of worth with increases in performance.

A graphical picture of this general shape of scoring function appears below in Figure 5.

Step 1. At this point, decision makers have two choices. The simplest procedure would be to fit a standardized cosine function to the performance measure whose equation is

$$\text{worth score} = 2\left(\frac{\text{Measured Performance}}{\text{Logical Upper Bound}}\right) + \frac{1}{2}\cos\text{ine}\left[\pi\left(\frac{\text{Measured Performance}}{\text{Logical Upper Bound}}\right)\right] - \frac{1}{2},$$

where $\pi = 3.1416$,

and cosine values may be looked up in a trigonometric table (function expressed in terms of radians) or computed on an engineering slide rule.

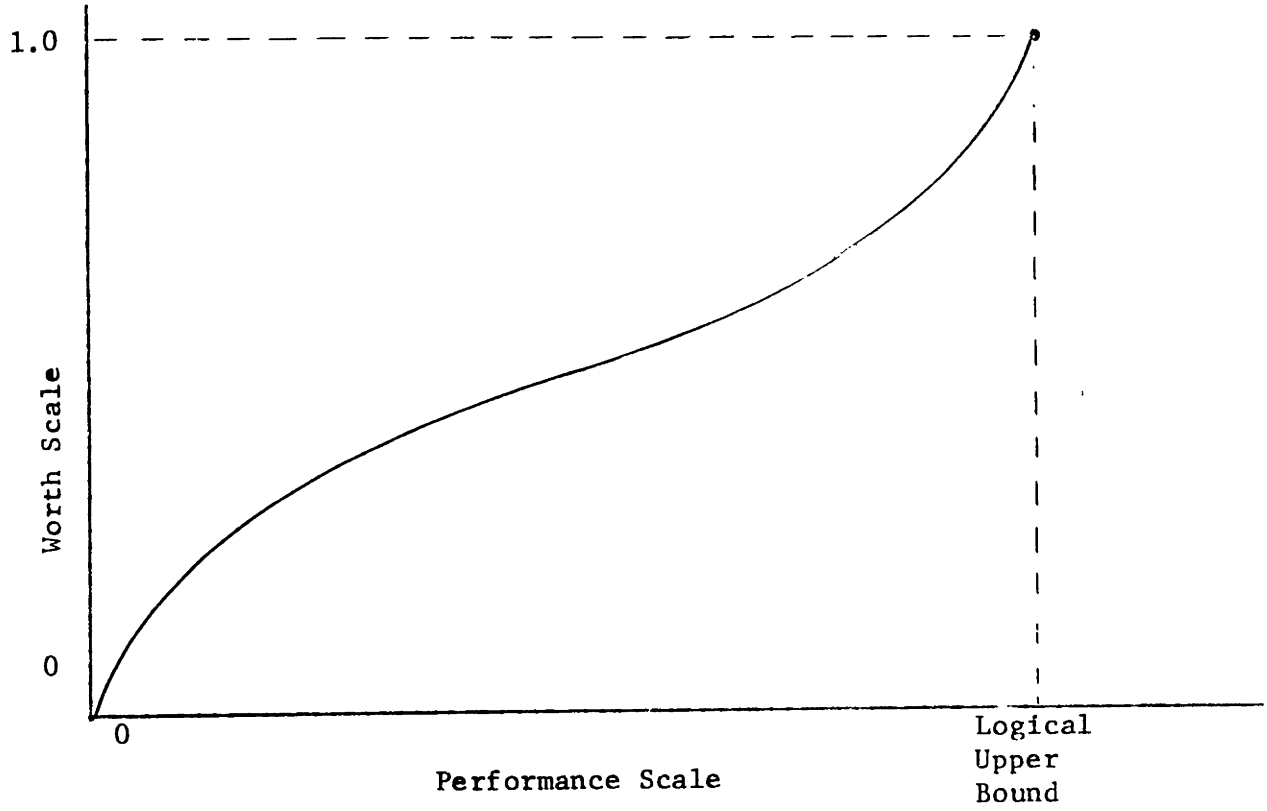


FIGURE 5

To determine whether or not this looks like an appropriate scoring function, it is suggested that a sheet of standard graph paper be procured and that the above equation be plotted thereupon. Five or six representative points should be sufficient to grasp the exact shape of the function and to decide whether or not it seems appropriate. If yes, this completes the procedure. If no, proceed to scoring procedure 20.

APPENDIX XI

SCORING PROCEDURE 11

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. bounded from above by some finite positive number;
4. reverse preference relationship;
5. worth score zero assigned to performance at the logical upper bound;
6. worth score one assigned to zero performance;
7. constant rate of change of worth with increases in performance.

The above seven characteristics describe completely a linear scoring function passing through the point whose coordinates are (performance = zero, worth score = one) and with negative slope equal to minus the reciprocal of the logical upper bound. The equation of this scoring function is

$$\text{worth score} = 1 - \frac{\text{Measured Performance}}{\text{Logical Upper Bound}} .$$

A graphical picture of this scoring function appears below in Figure 6.

APPENDIX XI

SCORING PROCEDURE 11

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. bounded from above by some finite positive number;
4. reverse preference relationship;
5. worth score zero assigned to performance at the logical upper bound;
6. worth score one assigned to zero performance;
7. constant rate of change of worth with increases in performance.

The above seven characteristics describe completely a linear scoring function passing through the point whose coordinates are (performance = zero, worth score = one) and with negative slope equal to minus the reciprocal of the logical upper bound. The equation of this scoring function is

$$\text{worth score} = 1 - \frac{\text{Measured Performance}}{\text{Logical Upper Bound}} .$$

A graphical picture of this scoring function appears below in Figure 6.

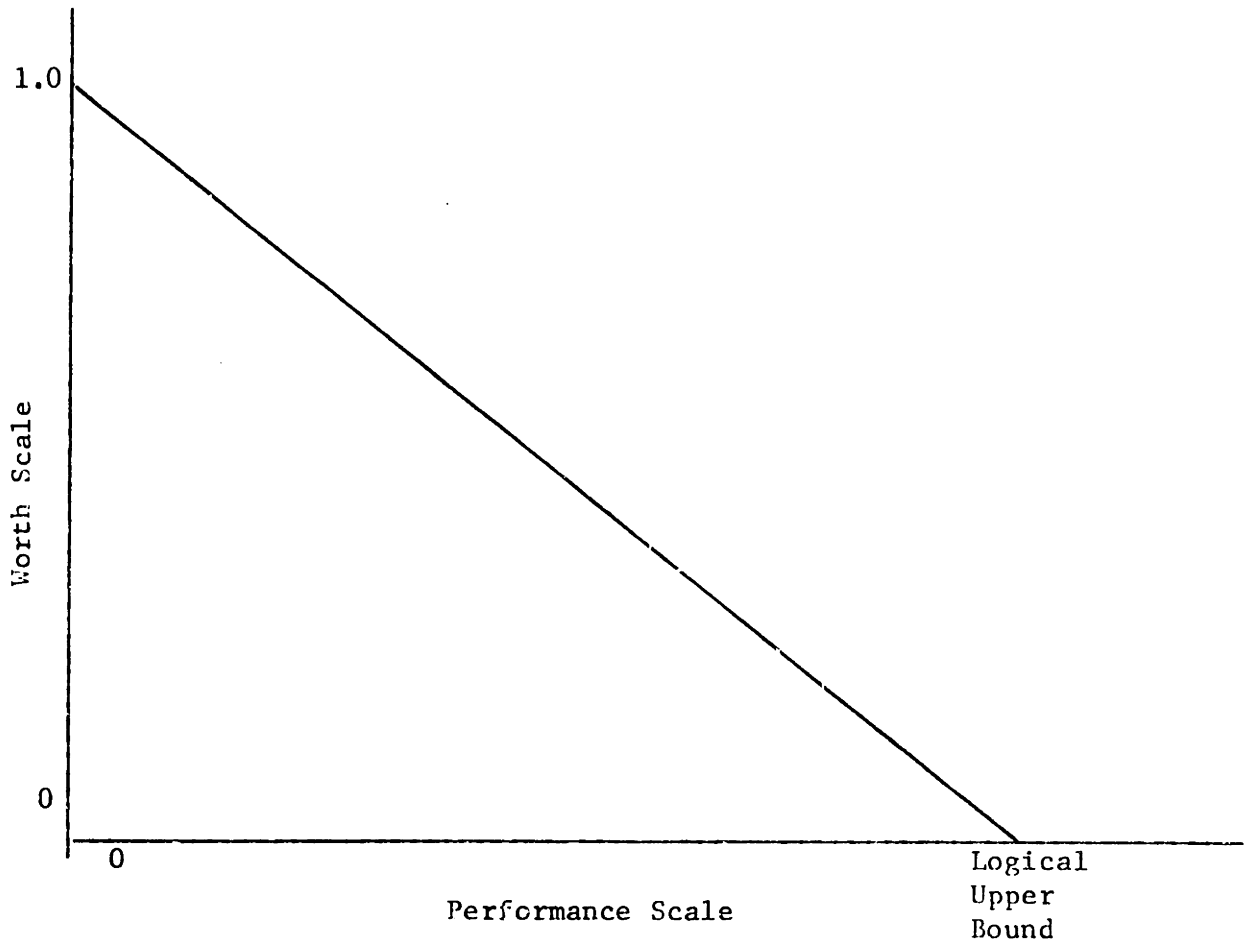


FIGURE 6

This completes the procedure.

APPENDIX XII

SCORING PROCEDURE 12

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. bounded from above by some finite positive number;
4. reverse preference relationship;
5. worth score zero assigned to performance at the logical upper bound;
6. worth score one assigned to zero performance;
7. uniformly accelerating rate of change of worth with increases in performance.

A graphical picture of this general shape of scoring function appears below in Figure 7.

Step 1. At this point, decision makers have two choices. The simplest procedure would be to fit a standardized quadratic scoring function to the performance measure under the following stipulated assumptions.

1. The scoring function is quadratic with negative second derivative (indicating uniform acceleration).
2. The maximum of the quadratic function falls exactly at the point whose coordinates are (performance = zero, worth score = one).

3. The quadratic function falls to the point whose coordinates are (performance = logical upper bound, worth score = zero).

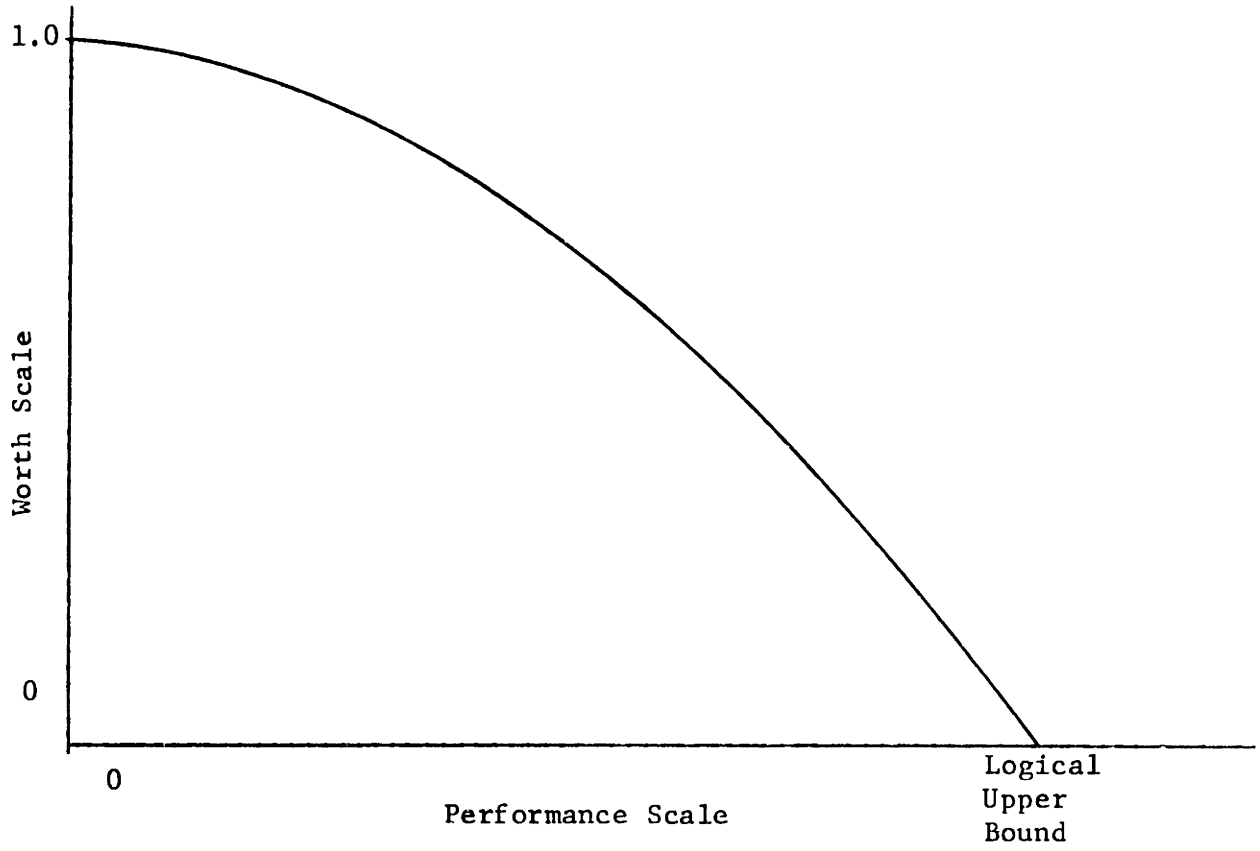


FIGURE 7

These three assumptions completely determine a scoring function (see Figure 7) whose equation is

$$\text{worth score} = 1 - \left(\frac{\text{Measured Performance}}{\text{Logical Upper Bound}}\right)^2 .$$

To determine whether or not this looks like an appropriate scoring function, it is suggested that a sheet of standard graph paper be procured and that the above equation be plotted thereupon. Five or six representative points should be sufficient to grasp the exact shape of the function and to decide whether or not it seems appropriate. If yes, this completes

APPENDIX XIII

SCORING PROCEDURE 13

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. bounded from above by some finite positive number;
4. reverse preference relationship;
5. worth score zero assigned to performance at the logical upper bound;
6. worth score one assigned to zero performance;
7. uniformly accelerating rate of change of worth with increases in performance.

A graphical picture of this general shape of scoring function appears below in Figure 8.

Step 1. At this point, decision makers have two choices. The simplest procedure would be to fit a standardized quadratic scoring function to the performance measure under the following stipulated assumptions.

1. The scoring function is quadratic with positive second derivative (indicating uniform acceleration).
2. The minimum of the quadratic function falls exactly at the point whose coordinates are (performance = logical upper bound, worth score = zero).

3. The upper left-tail of the function passes through the point whose coordinates are (performance = zero, worth score = one).

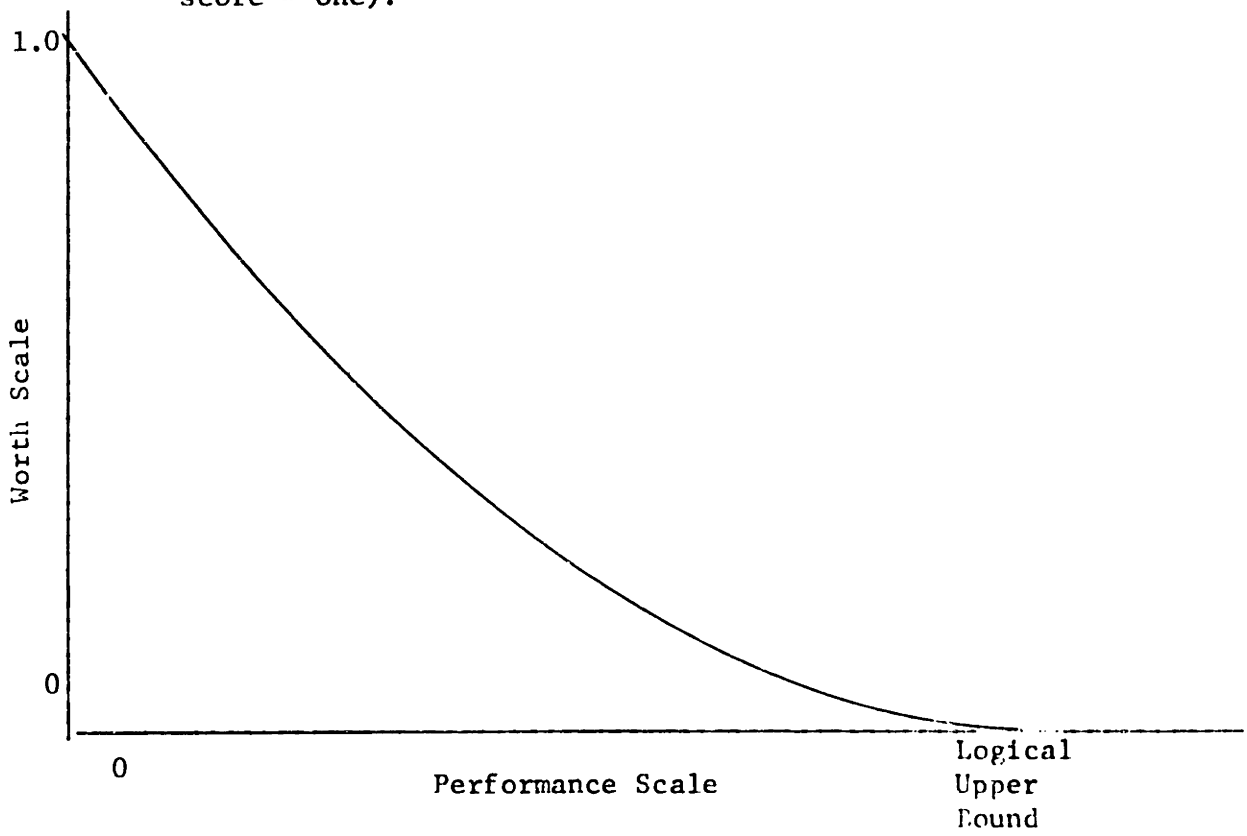


FIGURE 8

These three assumptions completely determine a scoring function (see Figure 8) whose equation is

$$\text{worth score} = 1 - 2\left(\frac{\text{Measured Performance}}{\text{Logical Upper Bound}}\right) + \left(\frac{\text{Measured Performance}}{\text{Logical Upper Bound}}\right)^2 .$$

To determine whether or not this looks like an appropriate scoring function, it is suggested that a sheet of standard graph paper be procured and that the above equation be plotted thereupon. Five or six representative points should be sufficient to grasp the exact shape of the function and to decide whether or not it seems appropriate. If yes, this completes the procedure. If no, proceed to scoring procedure 20.

APPENDIX XIV

SCORING PROCEDURE 14

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. bounded from above by some finite positive number;
4. reverse preference relationship;
5. worth score zero assigned to performance at the logical upper bound;
6. worth score one assigned to zero performance;
7. first accelerating, then decelerating rate of change of worth with increases in performance.

A graphical picture of this general shape of scoring function appears below in Figure 9.

Step 1. At this point, decision makers have two choices. The simplest procedure would be to fit a standardized cosine function to the performance measure whose equation is

$$\text{worth score} = \frac{1}{2} + \frac{1}{2} \text{ cosine} \left[\pi \left(\frac{\text{Measured Performance}}{\text{Logical Upper Bound}} \right) \right],$$

where $\pi = 3.1416$,

and cosine values may be looked up in a trigonometric table (function expressed in terms of radians) or computed on an engineering slide rule.

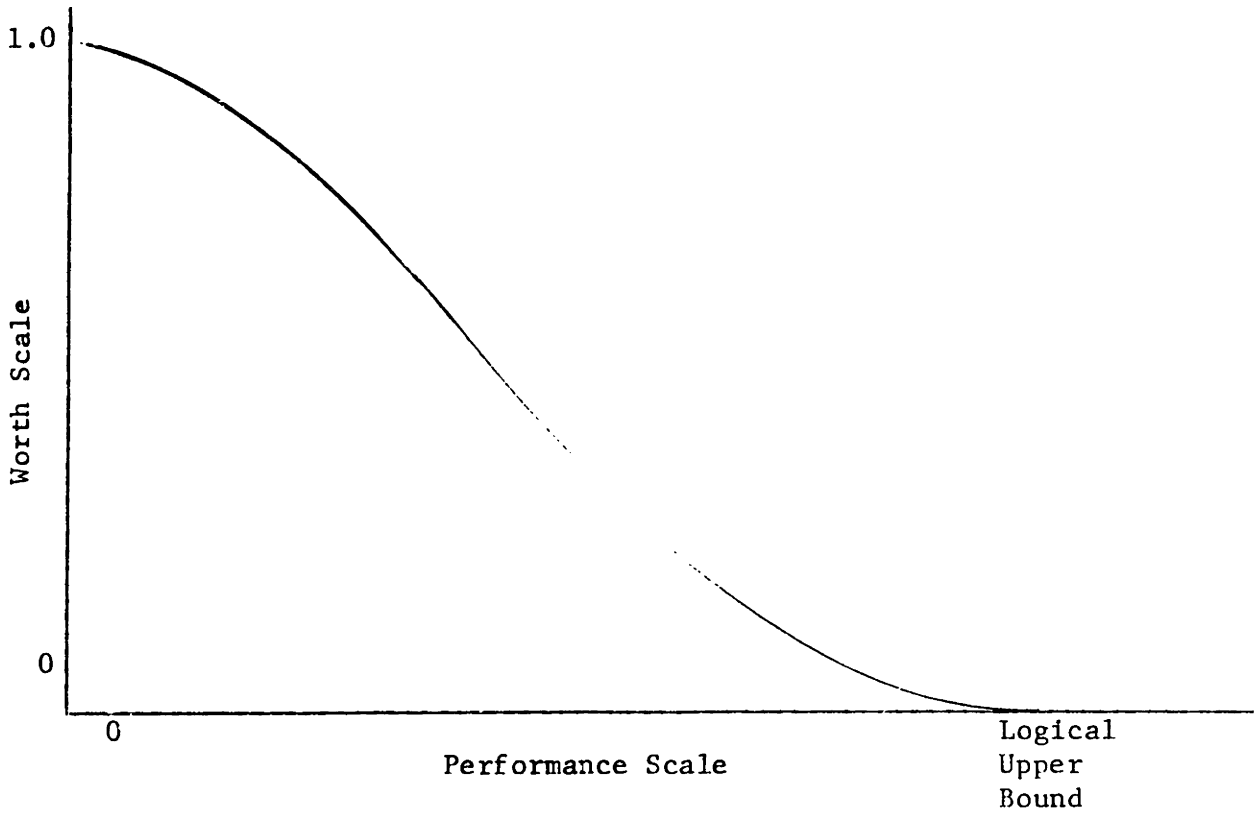


FIGURE 9

To determine whether or not this looks like an appropriate scoring function, it is suggested that a sheet of standard graph paper be procured and that the above equation be plotted thereupon. Five or six representative points should be sufficient to grasp the exact shape of the function and to decide whether or not it seems appropriate. If yes, this completes the procedure. If no, proceed to scoring procedure 20.

APPENDIX XV

SCORING PROCEDURE 15

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. bounded from above by some finite positive number;
4. reverse preference relationship;
5. worth score zero assigned to performance at the logical upper bound;
6. worth score one assigned to zero performance;
7. first decelerating, then accelerating rate of change of worth with increases in performance.

A graphical picture of this general shape of scoring function appears below in Figure 10.

Step 1. At this point, decision makers have two choices. The simplest procedure would be to fit a standardized cosine function to the performance measure whose equation is

$$\text{worth score} = \frac{3}{2} - 2\left(\frac{\text{Measured Performance}}{\text{Logical Upper Bound}}\right) - \frac{1}{2}\cos\text{ine}\left[\pi\left(\frac{\text{Measured Performance}}{\text{Logical Upper Bound}}\right)\right],$$

where $\pi = 3.1416$,

and cosine values may be looked up in a trigonometric table (function expressed in terms of radians) or computed on an engineering slide rule.

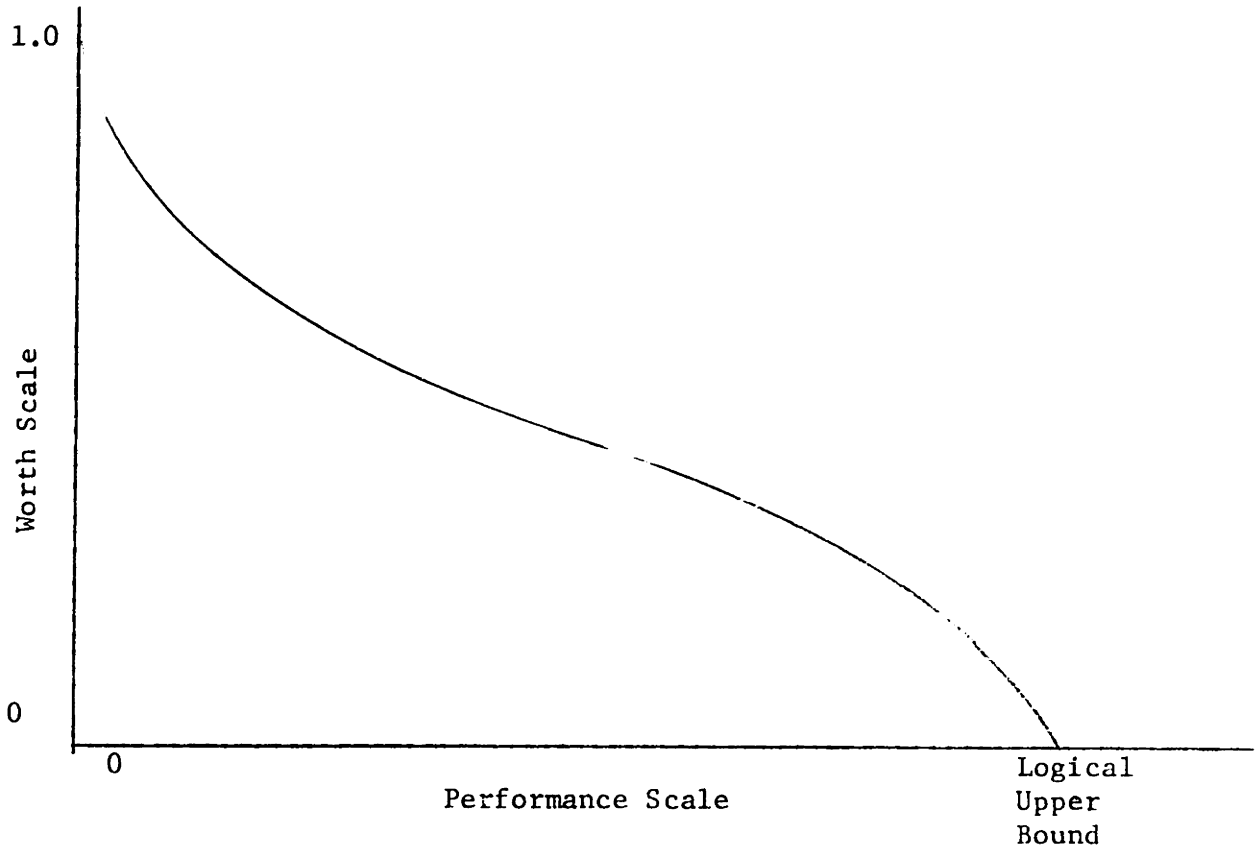


FIGURE 10

To determine whether or not this looks like an appropriate scoring function, it is suggested that a sheet of standard graph paper be procured and that the above equation be plotted thereupon. Five or six representative points should be sufficient to grasp the exact shape of the function and to decide whether or not it seems appropriate. If yes, this completes the procedure. If no, proceed to scoring procedure 20.

APPENDIX XVI

SCORING PROCEDURE 16

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. no logical upper bound;
4. direct preference relationship;
5. worth score zero assigned to zero performance;
6. worth score one assigned to infinite performance;
7. uniformly decelerating rate of change of worth with increases in performance.

A graphical picture of this general shape of scoring function appears below in Figure 11.

Step 1. There is no simple, standardized equation to fit all situations of this type. Although the general shape of this scoring function is given by the equation

$$\text{worth score} = 1 - \exp\left[(-k)\quad(\text{measured performance})\right] \quad ,$$

where exp is the exponential function with basis e = 2.718,

and k is a positive fitting constant,

still, the exact value of the fitting constant connot be determined in a standard way for all performance measures. Consequently, proceed to scoring procedure 20.

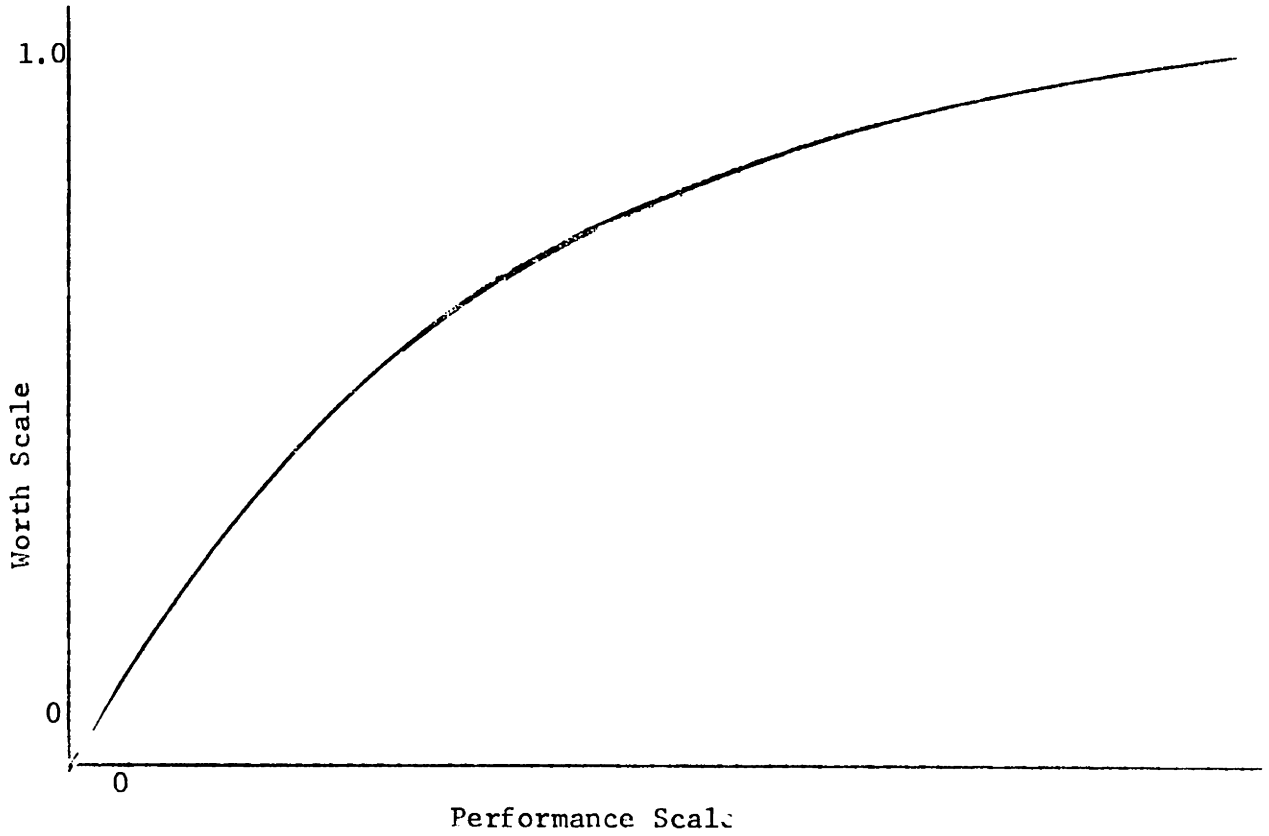


FIGURE 11

APPENDIX XVII

SCORING PROCEDURE 17

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. no logical upper bound;
4. direct preference relationship;
5. worth score zero assigned to zero performance;
6. worth score one assigned to infinite performance;
7. first accelerating, then decelerating rate of change of worth with increases in performance.

A graphical picture of this general shape of scoring function appears below in Figure 12.

Step 1. There is no simple, standardized equation to fit all situations of this type. Although the general shape of this scoring function is given by the equation

$$\text{worth score} = \exp\left[(-a)\ (\text{measured performance})\right]^{-b},$$

where exp is the exponential function with basis e = 2.718,

and both a and b are positive fitting constants ($b \geq 1$),

still, the exact values of the fitting constants cannot be determined in a standard way for all performance measures. Consequently, proceed to scoring procedure 20.

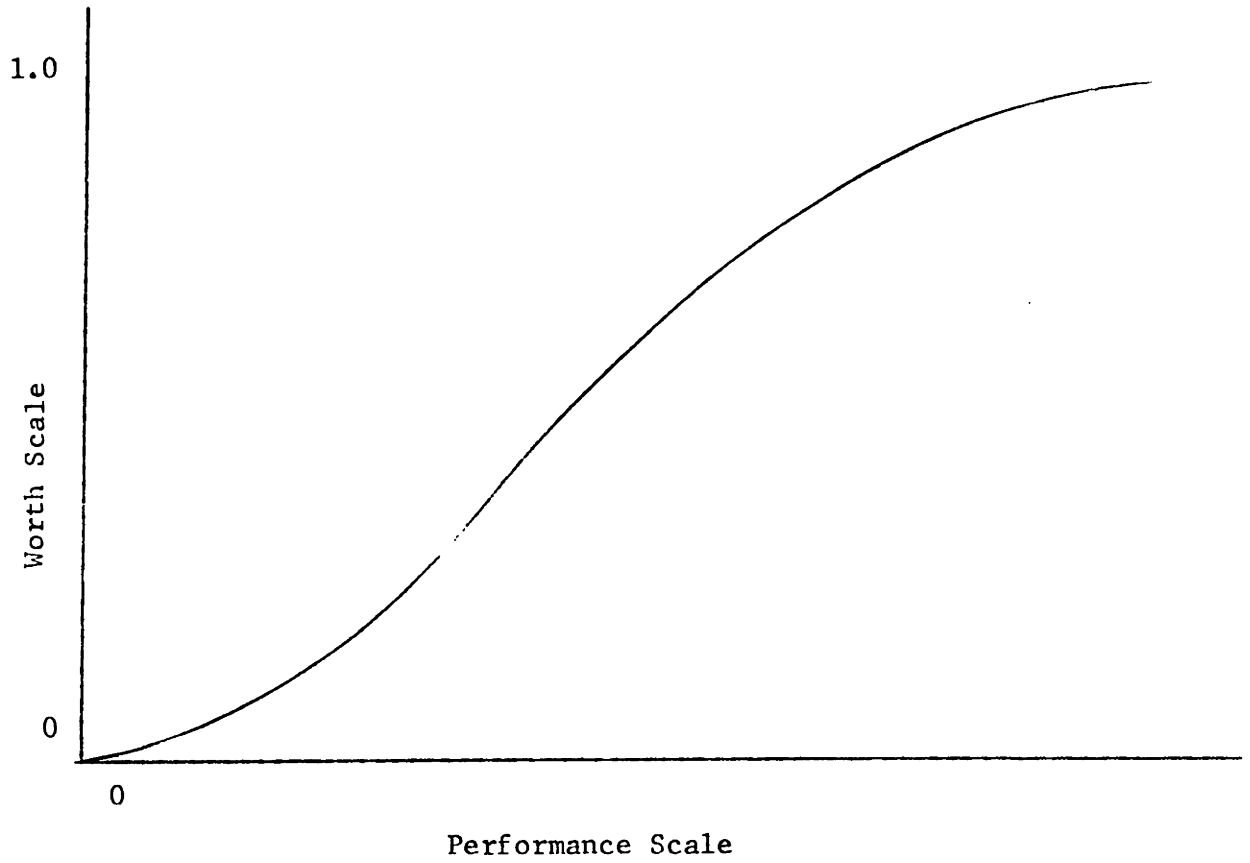


FIGURE 12

APPENDIX XVIII

SCORING PROCEDURE 18

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. no logical upper bound;
4. reverse preference relationship;
5. worth score zero assigned to infinite performance;
6. worth score one assigned to zero performance;
7. uniformly decelerating rate of change of worth with increases in performance.

A graphical picture of this general shape of scoring function appears below in Figure 13.

Step 1. There is no simple, standardized equation to fit all situations of this type. Although the general shape of this scoring function is given by the equation

$$\text{worth score} = \exp\left[(-k)\quad(\text{measured performance})\right] \quad ,$$

where exp is the expontial function with basis e = 2.718,

and k is a positive fitting constant,

still, the exact value of the fitting constant connot be determined in a standard way for all performance measures. Consequently, proceed to scoring procedure 20.

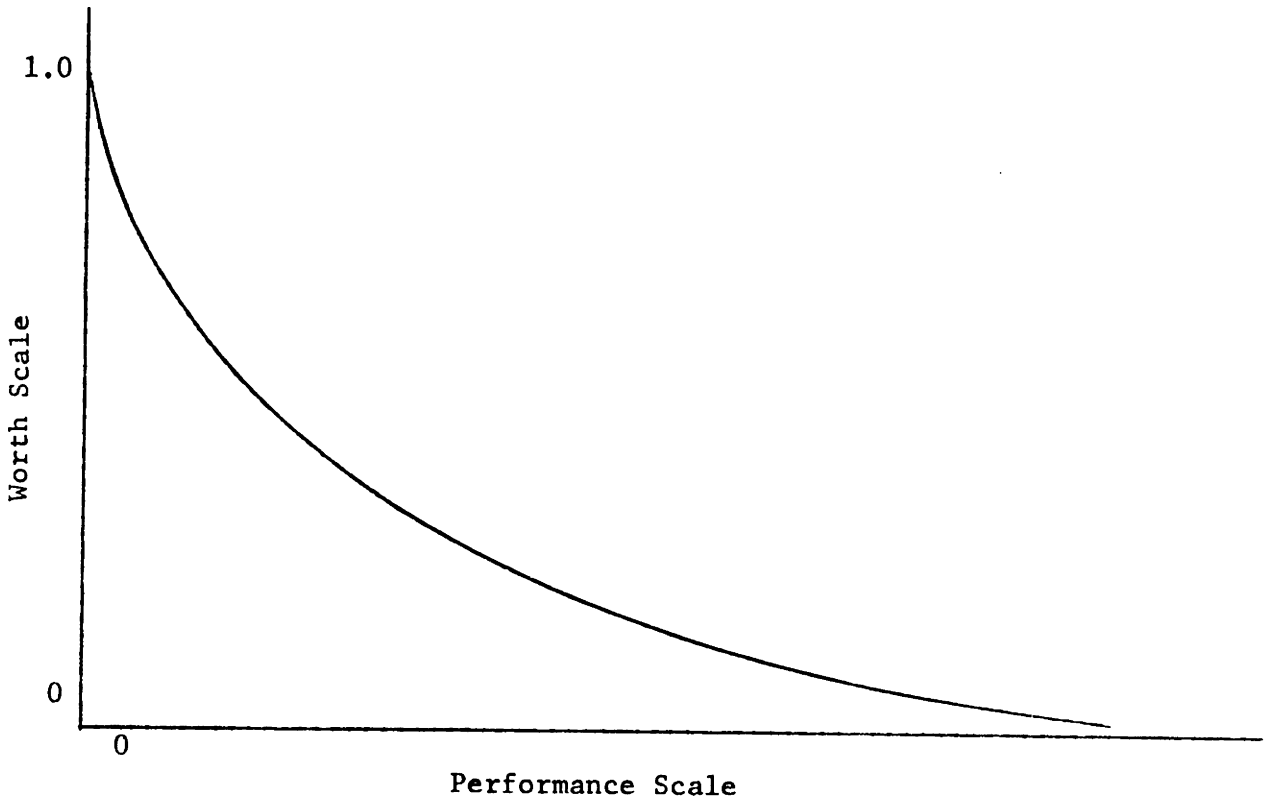


FIGURE 13

APPENDIX XIX

SCORING PROCEDURE 19

References in text: Section 2.2.5

The performance measure under scrutiny has been determined to have the following characteristics:

1. continuous scale;
2. bounded from below by zero;
3. no logical upper bound;
4. reverse preference relationship;
5. worth score zero assigned to infinite performance;
6. worth score one assigned to zero performance;
7. first accelerating, then decelerating rate of change of worth with increases in performance.

A graphical picture of this general shape of scoring function appears below in Figure 14.

Step 1. There is no simple, standardized equation to fit all situations of this type. Although the general shape of this scoring function is given by the equation

$$\text{worth score} = 1 - \exp\left[(-a)\ (\text{measured performance})\right]^{-b},$$

where exp is the exponential function with basis e = 2.718, and both a and b are positive fitting constants ($b \geq 1$),

still, the exact values of the fitting constants cannot be determined in a standard way for all performance measures. Consequently, proceed to scoring procedure 20.

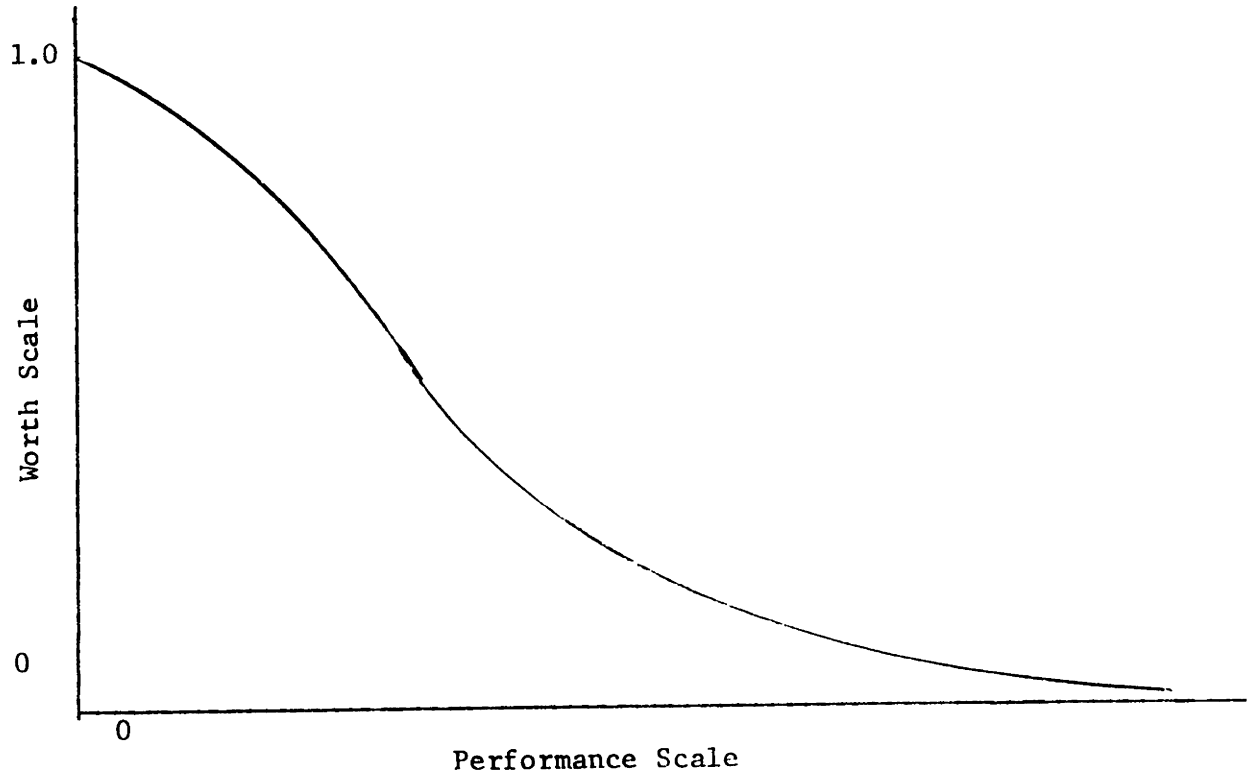


FIGURE 14

APPENDIX XX

SCORING PROCEDURE 20

References in text: Section 2.2.5 and Appendices VII through XIX

This procedure constitutes a continuation of each of the previous procedures listed below:

1. procedure 7
2. procedure 8
3. procedure 9
4. procedure 10
5. procedure 12
6. procedure 13
7. procedure 14
8. procedure 15
9. porcedure 16
10. procedure 17
11. procedure 18
12. procedure 19.

The general shape of the scoring function to be formulated has already been determined and inspected visually in one of these previous procedures. The purpose of this procedure is to select a specific curve of the general shape already determined.

Step 1. Assemble one or more decision makers.

Step 2. Prepare a standard sheet of graph paper for each decision maker layed out and marked off in the following manner.

1. Lay the worth scale along the vertical axis of a cartesian coordinate plane.
2. Mark off zero worth points at the origin of the graph and one worth point on the vertical axis near the top of the graph.
3. Mark off tenths of a point at equally-spaced intervals along the vertical axis between zero and one worth point.
4. Lay the performance scale along the horizontal axis.
5. Mark off zero performance at the origin of the graph and either the logical upper bound (if one exists) or some amount of performance substantially in excess of (say 50 percent greater than) the anticipated maximum proposed performance on the horizontal axis near the right-hand edge of the graph.
6. Establish convenient, equally-spaced performance subdivisions along the horizontal axis, and mark these off.

Step 3. Each decision maker will then ask himself the following question. "What level of performance, if promised by an alternative, should be considered ten percent successful in satisfying the related lowest level performance criterion?" Indicate this level of performance by placing an "x" in the interior of the graph at the position corresponding to that estimated level of performance along the horizontal performance scale and the ten percent or one-tenth worth point level along the vertical worth scale.

Step 4. Repeat Step 3 for the twenty percent, thirty percent, forty percent, fifty percent, sixty percent, seventy percent, eighty percent, and ninety percent worth point levels, respectively.

Step 5. Each decision maker should now have on his sheet of graph paper nine "x" marks. If the performance measure for which a scoring function is being formulated possesses a logical upper bound, two additional "x" marks may be placed on the graph - one at zero performance, and the other at the logical upper bound. If the performance measure possesses no logical upper bound, only one additional "x" mark may be placed on the graph corresponding to zero performance. Place the additional "x" mark(s) on the graph.

Step 6. Collect the graphs from each separate decision maker. Compute the (possibly weighted) arithmetic mean (averaged over separate decision makers) for each of the nine percentage levels along the worth scale.

Step 7. Prepare a new sheet of graph paper identical to the sheets prepared in Step 2.

Step 8. Plot the nine average points computed in Step 6 on this new sheet prepared in Step 7.

Step 9. With the aid of a French curve, draw a smooth curve of the predetermined general shape through the average points plotted in Step 8. The result is a scoring function in graphical form.

Step 10. To use this graphical scoring function, note the actual amount of performance promised by an alternative, and read the corresponding point score directly off the graph.

This completes the procedure.

APPENDIX XXI

THE PRE-EXPERIMENTAL QUESTIONNAIRE

STATEMENT OF PURPOSE

The purpose of this questionnaire is to gather information about your experience, education, and attitudes regarding formal techniques of quantitative analysis. Would you please perform the following six tasks.

1

Please write your name in the top left-hand corner of this page.

2

Although there is neither a minimum nor a maximum time limit involved, we are interested for research purposes in determining how much time you spend answering the next three questions. Therefore, please enter the current time of day in the space provided below.

START TIME_____

3

In the course of your professional duties, how many years of practical experience have you had using numerical techniques of cost analysis, performance analysis, effectiveness analysis, and/or cost effectiveness analysis. Enter this number in the space provided below. If no experience, enter a zero.

NUMBER OF YEARS OF EXPERIENCE_____

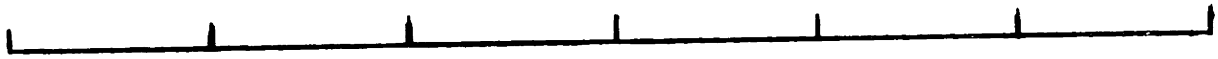
4

How many semester-length courses have you taken dealing with numerical techniques in any one or more of the above types of analysis? Enter this number in the space provided below.

5

Please indicate your present attitude toward such numerical techniques of analysis by circling one of the points on the scale displayed below.

EXTREMELY UN- FAVORABLE	MODERATELY UN- FAVORABLE	SLIGHTLY UN- FAVORABLE	NEUTRAL OR IN- DIFFERENT	SLIGHTLY FAVORABLE	MODERATELY FAVORABLE	EXTREMELY FAVORABLE



6

Except for entering your finish time, this completes the questionnaire. Thank you for your cooperation. Please do not discuss any of the answers you gave on this questionnaire with fellow participants in the course. Now enter the current time of day in the space provided below.

FINISH TIME_____

APPENDIX XXII

THE PRELIMINARY PREFERENCE QUESTIONNAIRE

STATEMENT OF PURPOSE

The purpose of this questionnaire is to permit you to indicate preliminary preferences for team partners in the program management simulation exercise to be conducted in the final week of this course. To accomplish this, would you please perform the following six tasks.

1

Inspect the name written at the top left-hand corner of this page. It should be your name. If it is your name, proceed to the second task. If it is not, please call this discrepancy immediately to Mr. Miller's attention.

2

Now turn to the third page of this questionnaire. There you will find a typewritten list of ten names of participants in this course. One of these names should be yours, and it should be circled. If your name appears circled on the third page, proceed to the third task. If not, please call this discrepancy to Mr. Miller's immediate attention.

3

Although there is neither a minimum nor a maximum time limit involved, we are interested for research purposes in determining the amount of time you spend on the next few tasks. Therefore, please enter the current time of day in the space provided below.

START TIME_____

4

Now return to the third page of this questionnaire. From the list of other participants (excluding yourself) choose exactly six (6) candidates for team partners in the simulation exercise. It is quite possible that you have not yet had an opportunity to acquaint yourself with all the participants on this list. However, please do the best job you can in choosing six (6) candidates on the basis of whatever preliminary impressions you have already gathered.

Place circles around the names of your six (6) preferred candiates. Do not place circles or any other kind of marks around the remaining names on the list.

5

There should now be exactly seven (7) circled names on the third page of this questionnaire - your name and the names of your six (6) preferred candidates. Please check to verify that this is the case. Then proceed to the sixth task.

6

Except for entering your finish time, this completes the questionnaire. Thank you for your cooperation. Please do not discuss any of the answers you gave on this questionnaire with fellow participants in the course. Now enter the current time of day in the space provided below.

FINISH TIME_____

(NOTES TO THE READER: Each subject's name was written in the top left-hand corner of the first page. The third page contained a list of that subject's ten-men sub-group, with his name circled in red.)

APPENDIX XXIII

A SAMPLE SHEET OF ALTERNATIVES

WHANGDOODLE(1101 3)

---	SNODGRASS	DOYSTER	ROYSTER	ZILCH
---	SNOOPDYKE	ZILCH	DOYSTER	SNODGRASS
---	ROYSTER	SNODGRASS	DOYSTER	SNOOPDYKE
---	ROYSTER	DALRYMPLE	SNOOPDYKE	DOYSTER
---	DALRYMPLE	SNOOPDYKE	ZILCH	DOYSTER
---	DALRYMPLE	DOYSTER	ZILCH	SNODGRASS
---	DALRYMPLE	ROYSTER	SNOOPDYKE	SNODGRASS
---	ROYSTER	ZILCH	DOYSTER	DALRYMPLE
---	ROYSTER	SNOOPDYKE	ZILCH	DOYSTER
---	ROYSTER	SNODGRASS	ZILCH	SNOOPDYKE
---	DOYSTER	ROYSTER	SNODGRASS	DALRYMPLE
---	DALRYMPLE	SNOOPDYKE	SNODGRASS	ZILCH
---	ROYSTER	ZILCH	DALRYMPLE	SNOOPDYKE
---	DOYSTER	SNODGRASS	DALRYMPLE	SNOOPDYKE
---	ROYSTER	SNODGRASS	ZILCH	DALRYMPLE

APPENDIX XXIV

THE STANDARD RANKING QUESTIONNAIRE

STATEMENT OF PURPOSE

The purpose of this questionnaire is to permit you to indicate your current preferences for the fifteen (15) groups of team partners in the program management simulation exercise to be conducted in the final week of this course. To accomplish this, would you please perform the following four tasks.

1

Inspect the name written at the top left-hand corner of this page. It should be your name. If it is your name, proceed to the second task. If it is not, please call this discrepancy immediately to Mr. Miller's attention.

2

Although there is neither a minimum nor a maximum time limit involved, we are interested for research purposes in determining the amount of time you spend ranking the fifteen (15) alternatives. Therefore, please enter the current time of day in the space provided below.

START TIME_____

3

On the second page of this questionnaire you will find a list of (15) alternative groups of team partners. An alternative consists of four (4) other participants who, along with yourself, could comprise a five-man team. Please rank the fifteen (15) alternatives from most preferred to least preferred. Place a "1" in front of the most preferred alternative, a "2" in front of the second most preferred alternative, and so on down the line. If you feel no preference between two or more alternatives on the list - that is, if you feel genuinely indifferent about choosing between them - assign the same number to each. Then assign the next higher number to the next most preferred alternative.

Except for entering your finish time, this completes the questionnaire. Thank you for your cooperation. Please do not discuss any of the answers you gave on this questionnaire with fellow participants in the course. Now enter the current time of day in the space provided below.

FINISH TIME_____

APPENDIX XXV

THE FIVE-ITEM QUESTIONNAIRE REFERRING TO NO INFORMATION, NO GUIDANCE

STATEMENT OF PURPOSE

Some time ago, you were asked to indicate your preferences for the fifteen alternative groups of team partners by assigning rank numbers to each group. Today, you were asked to perform the same task again. The purpose of this questionnaire is to assess the intervening passage of time upon your preferences for groups of team partners.

1

Please write your name at the top left-hand corner of this page.

2

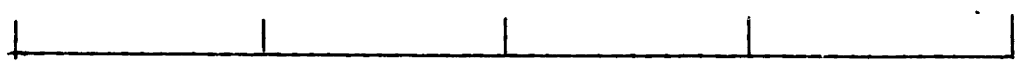
Although there is neither a minimum nor a maximum time limit involved, we are interested for research purposes in determining the amount of time you spend answering the next five questions. Therefore, please enter the current time of day in the space provided below.

START TIME_____

3

During the time which has elapsed since you were first asked to rank alternative groups of team partners, you may have had an opportunity to clarify your preferences for the various groups. Please indicate on the scale shown below whether or not your preferences have been clarified and, if so, to what extent. Circle the appropriate point on the scale.

NOT AT ALL CLARIFIED	SLIGHTLY CLARIFIED	MODERATELY CLARIFIED	SUBSTANTIALLY CLARIFIED	COMPLETELY CLARIFIED

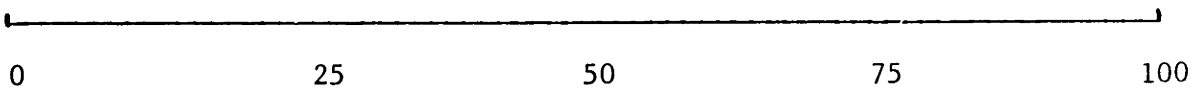


4

During this same time period, you may have reversed some of your preferences for various groups of team partners. That is, you may now prefer one group to another, whereas you may have indicated just the reverse preference for those two groups on the first ranking.

Considering only those pair-wise comparisons on the first ranking in which you indicated a definite preference (i.e., in which you assigned <u>different</u> rank numbers to the two groups being assessed), in about what percentage of those cases do you think you may have indicated a <u>reverse</u> preference on today's ranking? Please estimate the percentage of reversals by placing an "x" at the appropriate place along the scale shown below.

NO REVERSALS WHATSOEVER — ALL COMPARISONS COMPLETELY REVERSED

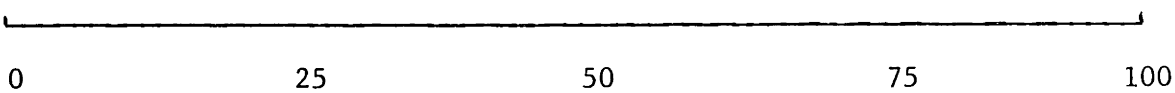


PERCENTAGE REVERSAL SCALE

5

Please indicate how <u>confident</u> you are that the rank numbers you just assigned to the fifteen (15) groups of team partners <u>accurately</u> represent your current preferences for them. Do this by marking an "X" at the appropriate place along the percentage confidence scale below.

NO CONFIDENCE WHATSOEVER — COMPLETE 100% CONFIDENCE

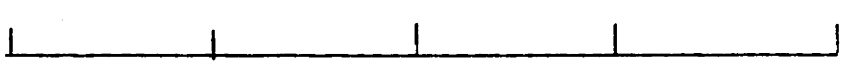


PERCENTAGE CONFIDENCE SCALE

6

Did you find the passage of time since the first ranking at all helpful in improving your ability to make a better choice of team partners? Please indicate whether or not and to what extent this time interval has helped you by circling the appropriate point on the scale shown below.

NOT AT ALL HELPFUL — SLIGHTLY HELPFUL — MODERATELY HELPFUL — SUBSTANTIALLY HELPFUL — EXTREMELY HELPFUL



Do you feel that any improvements made in your ability to make a better choice of team partners during the time interval since the first ranking where worth whatever time and effort you may have devoted to this problem? Please indicate whether or not and to what extent you think your time and effort were well spent by checking the appropriate statement below.

_____ The improvements gained were worth substantially less than the time and effort expended.
_____ The improvements gained were worth moderately less than the time and effort expended.
_____ The improvements gained were worth slightly less than the time and effort expended.
_____ The improvements gained were worth about the same amount as the time and effort expended.
_____ The improvements gained were worth slightly more than the time and effort expended.
_____ The improvements gained were worth moderately more than the time and effort expended.
_____ The improvements gained were worth substantially more than the time and effort expended.

8

Except for entering your finish time, this completes the questionnaire. Thank you for your cooperation. Please do not discuss any of the answers you gave on this questionnaire with fellow participants in the course. Now enter the current time of day in the space provided below.

FINISH TIME_____

APPENDIX XXVI

THE RANKING CHOICE QUESTIONNAIRE

STATEMENT OF PURPOSE

In this questionnaire you will be asked to review and compare two of your previous rankings of the fifteen groups of team partners. The purpose of this comparison is to let you select which of the two rankings better reflects your current preferences for the various groups.

1

Check the name written at the top left-hand corner of this page. Then check the names written at the top right-hand corner of each of the two ranking sheets. All three of these should be your name. If any of them are not, please call this discrepancy immediately to Mr. Miller's attention.

2

Although there is neither a minimum nor a maximum time limit involved, we are interested for research purposes in determining the time you spend answering the next question. Therefore, please enter the current time of day in the space provided below.

START TIME_____

3

Now turn to the two ranking sheets. Please indicate which of these more accurately reflects your current preferences for the fifteen alternative groups of team partners by checking the appropriate statement below.

_____ Sheet 1 is a more accurate representation of my current preferences.

_____ Sheet 2 is a more accurate representation of my current preferences.

_____ Both sheet 1 and sheet 2 reflect my current preferences equally accurately.

Except for entering your finish time, this completes the questionnaire. Thank you for your cooperation. Please do not discuss any of the answers you gave on this questionnaire with fellow participants in the course. Now enter the current time of day in the space provided below.

FINISH TIME_____

APPENDIX XXVII

A SAMPLE SHEET OF BIOGRAPHICAL INFORMATION

Name: James (Jim) B. Carraway

Branch of Service or Company Affiliation: Air Force

Rank or GS Rating (if applicable): Colonel

Present Assignment: System Support Manager, Operations Branch, Weapons Control SSM Division, Det 16, OCAMA, L.G. Hanscom Field, Massachusetts, (AFLC).

Program/Project to Which Assigned: 407L, 412L, 418L, and 682L

Current Home Address: 53 Offutt Road, Bedford, Massachusetts

Local Address and Telephone Number: VOQ 825, Room 255, 766-1

Educational Background:

Degree Awarded: BA

Institutions Attended: Furman University and Duke University

Work Experience: Chief, Electronics Division, Dir of Proc and Pdn, Hq WRAMA. Chief, WRAMA Det ASD SPO's. Chief, Logistics Plans Br. MAP Div, MAAG Vietnam. Logistics Staff Officer, Plans and Policy Br, MACV.

Planned Assignment After Graduation: Continue in present assignment.

Number of Years Experience with Formal Analytical Techniques: 2

Number of Semester-length Courses in Formal Analytical Techniques: 5

Self-Assessed Attitude toward Formal Analytical Techniques: Moderately Favorable

APPENDIX XXVIII

THE FIVE-ITEM QUESTIONNAIRE REFERRING TO RAW INFORMATION WITHOUT GUIDANCE

STATEMENT OF PURPOSE

Yesterday, you were given some printed biographical information about each of the six men whom you previously selected as candidates for team partners in the simulation exercise scheduled for the end of this course. Today, you were asked to indicate your current preferences for the fifteen alternative groups of team partners by assigning rank numbers to each group. The purpose of this questionnaire is to assess the impact of yesterday's biographical information upon your current preferences for the fifteen groups of team partners.

1

Please write your name at the top left-hand corner of this page.

2

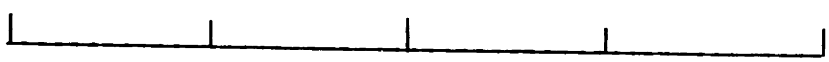
Although there is neither a minimum nor a maximum time limit involved, we are interested for research purposes in determining the amount of time you spend answering the next five questions. Therefore, please enter the current time of day in the space provided below.

START TIME_____

3

On the basis of yesterday's biographical information, you may have had an opportunity to clarify your preferences for the various groups. Please indicate on the scale shown below whether or not your preference have been clarified and, if so, to what extent. CIRCLE the appropriate point on the scale.

NOT AT ALL CLARIFIED	SLIGHTLY CLARIFIED	MODERATELY CLARIFIED	SUBSTANTIALLY CLARIFIED	COMPLETELY CLARIFIED



4

As a result of yesterday's biographical information, you may have reversed some of your preferences for various groups of team partners. That is, you may now prefer one group to another, whereas you may have indicated just the reverse preference for those two groups on the last ranking. Considering only those pair-wise comparisons on the last ranking in which you indicated a definite preference (i.e., in which you assigned <u>different</u> rank numbers to the two groups being assessed), in about what percentage of those cases do you think you may have indicated a <u>reverse</u> preferences on today's ranking? Please estimate the percentage of reversals by PLACING AN "X" at the appropriate place along the scale shown below.

NO REVERSALS WHATSOEVER — ALL COMPARISONS COMPLETELY REVERSED

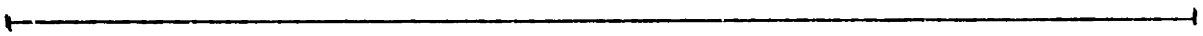
0 25 50 75 100

<u>PERCENTAGE REVERSAL SCALE</u>

5

Please indicate how <u>confident</u> you are that the rank numbers you just assigned to the fifteen (15) groups of team partners <u>accurately</u> represent your current preferences for them. Do this by <u>PLACING AN "X"</u> at the appropriate place along the percentage confidence scale shown below.

NO CONFIDENCE WHATSOEVER — COMPLETE 100% CONFIDENCE



0 25 50 75 100

<u>PERCENTAGE CONFIDENCE SCALE</u>

6

Did you find yesterday's biographical information at all <u>helpful</u> in improving your ability to make a better choice of team partners? Please indicate whether or not and to what extent this information has <u>helped</u> you by <u>CIRCLING</u> the appropriate point on the scale shown below.

NOT AL ALL HELPFUL	SLIGHTLY HELPFUL	MODERATELY HELPFUL	SUBSTANTIALLY HELPFUL	EXTREMELY HELPFUL

7

Do you feel that, on the basis of yesterday's information, any improvements made in your ability to make a better choice of team partners were worth whatever time and effort you may have devoted to this problem? Please indicate whether or not and to what extent you think your time and effort were well spent by CHECKING the appropriate statement below.

_____ The improvements gained were worth substantially less than the time and effort expended.
_____ The improvements gained were worth moderately less than the time and effort expended.
_____ The improvements gained were worth slightly less than the time and effort expended.
_____ The improvements gained were worth about the same amount as the time and effort expended.
_____ The improvements gained were worth slightly more than the time and effort expended.
_____ The improvements gained were worth moderately more than the time and effort expended.
_____ The improvements gained were worth substantially more than the time and effort expended.

8

Except for entering your finish time, this completes the questionnaire. Thank you for your cooperation. Please do not discuss any of the answers you gave on this questionnaire with fellow participants in the course. Now enter the current time of day in the space provided below.

FINISH TIME_____

APPENDIX XXIX

INSTRUCTION PAGE OF THE FIRST NEUTRAL TASK

STATEMENT OF PURPOSE

The purpose of this questionnaire is to permit you to estimate the ratios of lengths of various pairs of printed lines. The questionnaire is divided into three (3) separate parts, each of which will have its own set of printed instructions.

1

Please write your name at the top left-hand corner of this page.

2

Although there is neither a minimum nor a maximum time limit involved, we are interested for research purposes in determining the amount of time you spend estimating line length ratios on this questionnaire. Therefore, please enter the current time of day in the space provided below.

START TIME_____

3

Please turn the page and read the instructions for Part 1 of this questionnaire. After reading the instructions, you may begin immediately and proceed through the entire questionnaire. However, please do NOT go back to earlier parts of the questionnaire after having completed later parts.

APPENDIX XXX

THE FIVE-ITEM QUESTIONNAIRE REFERRING TO BOTH ORDINAL AND CARDINAL GUIDANCE

STATEMENT OF PURPOSE

You have just been asked to indicate your preferences for the fifteen (15) alternative groups of team partners by assigning rank numbers to each group. Prior to this, you were asked to discuss your criteria of preference with Mr. Miller during an interview. The purpose of this questionnaire is to assess the impact of the interview upon your preferences for groups of team partners.

1

Please write your name at the top left-hand corner of this page.

2

Although there is neither a minimum nor a maximum time limit involved, we are interested for research purposes in determining the amount of time you spend answering the next five questions. Therefore, please enter the current time of day in the space provided below.

START TIME_____

3

On the basis of the interview, you may have had an opportunity to clarify your preferences for the various groups. Please indicate on the scale below whether or not your preferences have been clarified by virtue of participating in the interview and, if so, to what extent. CIRCLE the appropriate point on the scale below.

NOT AT ALL CLARIFIED	SLIGHTLY CLARIFIED	MODERATELY CLARIFIED	SUBSTANTIALLY CLARIFIED	COMPLETELY CLARIFIED

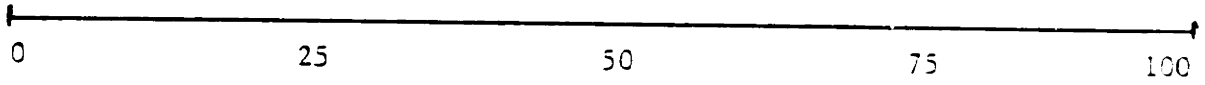
4

Since the last ranking, you may have reversed some of your preferences for the various groups of team partners. That is, you may now prefer one group to another, whereas you may have indicated just the reverse preference for these two groups on the last ranking. Considering only those pair-wise comparisions on the last ranking in which you assigned

different rank numbers to the two groups being assessed (i.e., in which you indicated a definite preference for one over the other), in about what percentage of those cases do you think you may have indicated a reverse preference on today's ranking? Please estimate the percentage of reversals by PLACING AN "X" at the appropriate place along the scale shown below.

NO REVERSALS WHATSOEVER

ALL COMPARISONS COMPLETELY REVERSED

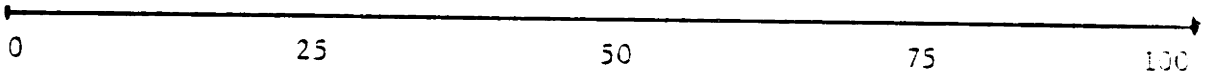


5

Please indicate how confident you are that the rank numbers you just assigned to the fifteen (15) groups of team partners accurately represent your current preferences for them. Do this by PLACING AN "X" at the appropriate place along the percentage confidence scale below.

NO CONFIDENCE WHATSOEVER

COMPLETE 100% CONFIDENCE

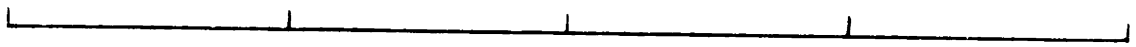


PERCENTAGE CONFIDENCE SCALE

6

Did you find the interview process at all helpful in improving your ability to make a better choice of team partners? Please indicate whether or not and to what extent the interview helped you by CIRCLING the appropriate point on the scale shown below.

NOT AT ALL HELPFUL	SLIGHTLY HELPFUL	MODERATELY HELPFUL	SUBSTANTIALLY HELPFUL	EXTREMELY HELPFUL



7

Do you feel that any improvements made in your ability to make a better choice of team partners as a result of the interview process were worth whatever time and effort you devoted to that task? Please indicate whether or not and to what extent you think your time and effort were well spent by CHECKING the appropriate statement below.

_____ The improvements gained were worth substantially less than the time and effort expended.
_____ The improvements gained were worth moderately less than the time and effort expended.
_____ The improvements gained were worth slightly less than the time and effort expended.
_____ The improvements gained were worth about the same amount as the time and effort expended.
_____ The improvements gained were worth slightly more than the time and effort expended.
_____ The improvements gained were worth moderately more than the time and effort expended.
_____ The improvements gained were worth substantially more than the time and effort expended.

8

Except for entering your finish time, this completes the questionnaire. Thank you for your cooperation. Please do not discuss any of the answers you gave on this questionnaire with fellow participants in the course. Now enter the current time of day in the space provided below.

FINISH TIME_____

APPENDIX XXXI

THE FIVE-ITEM QUESTIONNAIRE REFERRING TO RATIO ESTIMATION

STATEMENT OF PURPOSE

You have just been asked to indicate your preferences for the fifteen (15) alternative groups of team partners by assigning rank numbers to each group. Prior to this, you were asked to estimate various line lengths. The purpose of this questionnaire is to assess the impact of estimating line lengths upon your preferences for groups of team partners.

1

Please write your name at the top left-hand corner of this page.

2

Although there is neither a minimum nor a maximum time limit involved, we are interested for research purposes in determining the amount of time you spend answering the next five questions. Therefore, please enter the current time of day in the space provided below.

START TIME_____

3

On the basis of estimating line lengths, you may have had the opportunity to clarify your preferences for the various groups. Please indicate on the scale shown below whether or not your preferences have been clarified by virtue of participating in the estimation exercise and, if so, to what extent. CIRCLE the appropriate point on the scale.

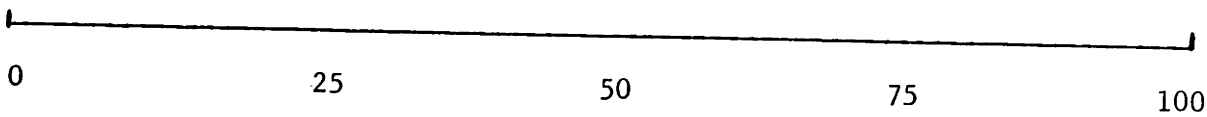
NOT AT ALL CLARIFIED	SLIGHTLY CLARIFIED	MODERATELY CLARIFIED	SUBSTANTIALLY CLARIFIED	COMPLETELY CLARIFIED

4

Since the last ranking, you may have reversed some of your preferences for the various groups of team partners. That is, you may now prefer one group to another, whereas you may have indicated just the reverse preference for these two groups on the last ranking. Considering only those pair-wise comparisons on the last ranking in which you indicated a definite preference (i.e., in which you assigned different rank numbers to the two groups being assessed), in about what percentage of those cases do you think you may have indicated a reverse preference on today's ranking? Please estimate the percentage of reversals by PLACING AN "X" at the appropriate place along the scale shown below.

NO REVERSALS WHATSOEVER

ALL COMPARISONS COMPLETELY REVERSED



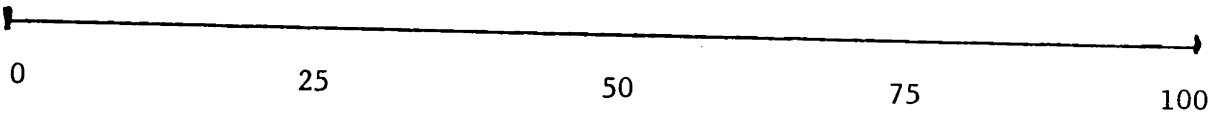
PERCENTAGE REVERSAL SCALE

5

Please indicate how confident you are that the rank numbers you just assigned to the fifteen (15) groups of team partners accurately represent your current preferences for them. Do this by PLACING AN "X" at the appropriate place along the percentage confidence scale below.

NO CONFIDENCE WHATSOEVER

COMPLETE 100% CONFIDENCE



PERCENTAGE CONFIDENCE SCALE

6

Did you find the estimation exercise at all helpful in improving your ability to make better choices among groups of team partners? Please indicate whether or not and to what extent the estimation exercise helped you by CIRCLING the appropriate point on the scale shown below.

NOT AT ALL HELPFUL	SLIGHTLY HELPFUL	MODERATELY HELPFUL	SUBSTANTIALLY HELPFUL	EXTREMELY HELPFUL

7

Do you feel that any improvements made in your ability to make a better choice of team partners as a result of the estimation exercise were worth the time and effort devoted to this task? Please indicate whether or not and to what extent you think your time and effort were well spent by CHECKING the appropriate statement below.

_____ The improvements gained were worth substantially less than the time and effort expended.
_____ The improvements gained were worth moderately less than the time and effort expended.
_____ The improvements gained were worth slightly less than the time and effort expended.
_____ The improvements gained were worth about the same amount as the time and effort expended.
_____ The improvements gained were worth slightly more than the time and effort expended.
_____ The improvements gained were worth moderately more than the time and effort expended.
_____ The improvements gained were worth substantially more than the time and effort expended.

8

Except for entering your finish time, this completes the questionnaire. Thank you for your cooperation. Please do not discuss any of the answers you gave on this questionnaire with fellow participants in the course. Now enter the current time of day in the space provided below.

FINISH TIME_____

APPENDIX XXXII

THE FIVE-ITEM QUESTIONNAIRE REFERRING TO THE REMOTE ASSOCIATES TEST

STATEMENT OF PURPOSE

You have just been asked to indicate your preferences for the fifteen (15) alternative groups of team partners by assigning rank numbers to each group. Prior to this, you were asked to take the Remote Associates Test by filling in a fourth word associated, respectively, with thirty sets of three words each. The purpose of this questionnaire is to assess the impact of your having taken the Remote Associates Test upon your preferences for groups of team partners.

1

Please write your name at the top left-hand corner of this page.

2

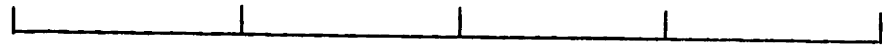
Although there is neither a minimum nor a maximum time limit involved, we are interested for research purposes in determining the amount of time you spend answering the next five questions. Therefore, please enter the current time of day in the space provided below.

START TIME_____

3

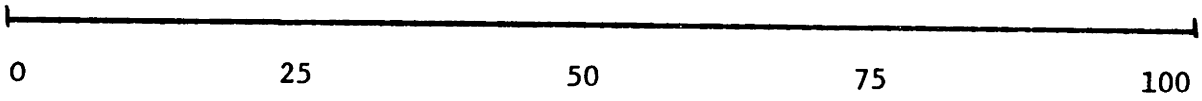
On the basis of taking the Remote Associates Test, you may have had an opportunity to clarify your preferences for the various groups. Please indicate on the scale below whether or not your preferences have been clarified by virtue of taking the Remote Associates Test and, if so, it what extent. CIRCLE the appropriate point on the scale below.

NOT AT ALL CLARIFIED	SLIGHTLY CLARIFIED	MODERATELY CLARIFIED	SUBSTANTIALLY CLARIFIED	COMPLETELY CLARIFIED



Since the last ranking, you may have reversed some of your preferences for the various groups of team partners. That is, you may now prefer one group to another, whereas you may have indicated just the reverse preference for these two groups on the last ranking. Considering only those pair-wise comparisons on the last ranking in which you assigned different rank numbers to the two groups being assessed (i.e., in which you indicated a definite preference for one over the other), in about what percentage of those cases do you think you may have indicated a reverse preference on today's ranking? Please estimate the percentage of reversals by PLACING AN "X" at the appropriate place along the scale shown below.

NO REVERSALS WHATSOEVER — ALL COMPARISONS COMPLETELY REVERSED

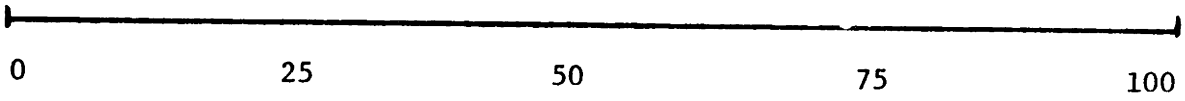


PERCENTAGE REVERSAL SCALE

5

Please indicate how confident you are that the rank numbers you just assigned to the fifteen (15) groups of team partners accurately represent your current preferences for them. Do this by PLACING AN "X" at the appropriate place along the percentage confidence scale below.

NO CONFIDENCE WHATSOEVER — COMPLETE 100% CONFIDENCE



PERCENTAGE CONFIDENCE SCALE

6

Did you find the experience of taking the Remote Association Test at all helpful in improving your ability to make a better choice of team partners? Please indicate whether or not and to what extent taking the test helped you by CIRCLING the appropriate point on the scale shown below.

NOT AT ALL HELPFUL — SLIGHTLY HELPFUL — MODERATELY HELPFUL — SUBSTANTIALLY HELPFUL — EXTREMELY HELPFUL

7

Do you feel that any improvements made in your ability to make a better choice of team partners as a result of taking the Remote Associates Test were worth whatever time and effort you devoted to the task? Please indicate whether or not and to what extent you think your time and effort were <u>well spent</u> by <u>CHECKING</u> the appropriate statement below.

_____ The improvements gained were worth <u>substantially less</u> than the time and effort expended.
_____ The improvements gained were worth <u>moderately less</u> than the time and effort expended.
_____ The improvements gained were worth <u>slightly less</u> than the time and effort expended.
_____ The improvements gained were worth <u>about the same amount</u> as the time and effort expended.
_____ The improvements gained were worth <u>slightly more</u> than the time and effort expended.
_____ The improvements gained were worth <u>moderately more</u> than the time and effort expended.
_____ The improvements gained were worth <u>substantially more</u> than the time and effort expended.

8

Except for entering your finish time, this completes the questionnaire. Thank you for your cooperation. Please do not discuss any of the answers you gave on this questionnaire with fellow participants in the course. Now enter the current time of day in the space provided below.

FINISH TIME_____

APPENDIX XXXIII

THE FINAL CHOICE QUESTIONNAIRE

On the accompanying sheet you will find your fifteen (15) alternative groups of team partners. Please indicate which one of these is your final choice for purposes of participating in the simulation exercise. Indicate your choice by PLACING AN "X" in the space to the left of the selected team. The actual teams will be formed immediately after you indicate your final choices, and the results should be posted by the end of this afternoon.

(NOTE TO THE READER: A standard ranking sheet was appended to this questionnaire.)

APPENDIX XXXIV

THE FIVE-ITEM QUESTIONNAIRE REFERRING THE THE FINAL CHOICE

STATEMENT OF PURPOSE

You have just been asked to indicate your preferences for the fifteen (15) alternative groups of team partners by assigning rank numbers to each group. Prior to this, you were asked to select one of these groups as your team in the simulation exercise. The purpose of this questionnaire is to assess the impact of having made a final choice upon your overall preferences for the fifteen groups of team partners.

1

Although there is neither a minimum nor a maximum time limit involved we are interested for research purposes in determining the amount of time you spend answering the next five questions. Therefore, please enter the current time of day in the space provided below.

START TIME _____

2

The act of making a final choice may have clarified your preferences for the various groups. Please indicate on the scale below whether or not your preferences have been clarified by virtue of making a final choice and, if so, to what extent. CIRCLE the appropriate point on the scale below.

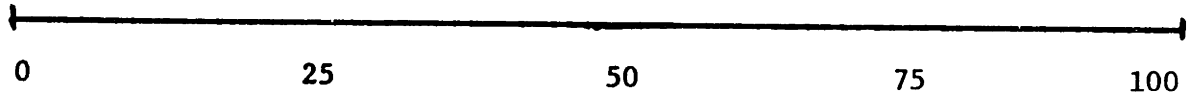
NOT AT ALL CLARIFIED	SLIGHTLY CLARIFIED	MODERATELY CLARIFIED	SUBSTANTIALLY CLARIFIED	COMPLETELY CLARIFIED

4

Since the last ranking, you may have reversed some of your preferences for the various groups of team partners. That is, you may now prefer one group to another, whereas you may have indicated just the reverse preference for these two groups on the last ranking. Considering only those pair-wise comparisons on the last ranking in which you assigned different rank numbers to the two groups being assessed (i.e., in which you indicated a definite preference for one over the other), in about what percentage of those cases do you think you may have indicated a reverse

preference on today's ranking? Please estimate the percentage of reversals by PLACING AN "X" at the appropriate place along the scale shown below.

NO REVERSALS WHATSOEVER — ALL COMPARISONS COMPLETELY REVERSED

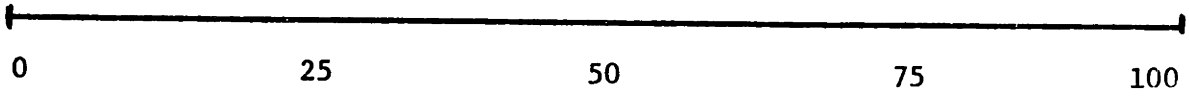


PERCENTAGE REVERSAL SCALE

5

Please indicate how confident you are that the rank numbers you just assigned to the fifteen (15) groups of team partners accurately represent you current preferences for them. Do this by PLACING AN "X" at the appropriate place along the percentage confidence scale below.

NO CONFIDENCE WHATSOEVER — COMPLETE 100% CONFIDENCE



PERCENTAGE CONFIDENCE SCALE

6

Did you find the requirement to make a final choice at all helpful in improving you ability to make a better choice of team partners? Please indicate whether or not and to what extent the requirement to make a final choice helped you by CIRCLING the appropriate point on the scale shown below.

NOT AT ALL HELPFUL | SLIGHTLY HELPFUL | MODERATELY HELPFUL | SUBSTANTIALLY HELPFUL | EXTREMELY HELPFUL

7

Do you feel that any improvements made in your ability to make a better choice of team partners as a result of being required to make a final choice were worth whatever time and effort you devoted to that task? Please indicate whether or not and to what extent you think your time and effort were well spent by CHECKING the appropriate statement below.

_____ The improvements gained were worth substantially less than the time and effort expended.
_____ The improvements gained were worth moderately less than the time and effort expended.
_____ The improvements gained were worth slightly less than the time effort expended.
_____ The improvements gained were worth about the same amount as the time and effort expended.
_____ The improvements gained were worth slightly more than the time and effort expended.
_____ The improvements gained were worth moderately more than the time and effort expended.
_____ The improvements gained were worth substantially more than the time and effort expended.

8

Except for entering your finish time, this completes the questionnaire. Thank you for your cooperation. Please do not discuss any of the answers you gave on this questionnaire with fellow participants in the course. Now enter the current time of day in the space provided below.

FINISH TIME_____

APPENDIX XXXV

THE SOCIOMETRIC PREFERENCE QUESTIONNAIRE

STATEMENT OF PURPOSE

This questionnaire marks the end of our research project. Its purpose is to clarify several issues which have been raised during the last ten weeks. Such clarification should facilitate interpretation of the final results.

1

Please write your name in the space provided at the top left-hand corner of this page.

2

Although you have already made your final choice of a TEAM, we are interested in knowing how you feel about the various individuals from whom you selected a team. Listed below is your name and the names of the nine (9) other individuals you were originally given to choose among during the first week of this course. How do you now feel about having each of these nine (9) other individuals on the same team with you for purposes of participating in the simulation exercise. Please indicate the extent to which you feel either a preference, an aversion, or indifference to including each of them individually on your team by PLACING AN "X" in the appropriate column beside each individual's name. DO NOT bother to place a mark beside your own name.

	SUBSTANTIAL AVERSION	MODERATE AVERSION	SLIGHT AVERSION	INDIF-FERENCE	SLIGHT PREFERENCE	MODERATE PREFERENCE	SUBSTANTIAL PREFERENCE
NAME 1							
NAME 2							
NAME 3							
NAME 4							
NAME 5							
NAME 6							
NAME 7							
NAME 8							
NAME 9							
NAME 10							

3

Thank you again for your extraordinary cooperation in this research project. It has been a great pleasure to work with you.

APPENDIX XXXVI

THE POST-EXPERIMENTAL QUESTIONNAIRE

STATEMENT OF PURPOSE

The purpose of this questionnaire is to obtain information in two areas which relate to your recent participation in the simulation exercise.

1

Please write your name in the top left-hand corner of this page.

2

During the course of our experiment, you were asked to rank alternative teams, to inspect biographical information concerning individual team members, to articulate your criteria for choosing among alternative teams, and, in some cases, to quantify your preferences for each team. Please indicate the extent to which you utilized any of these same evaluation and decision making methods in the course of making decisions during the simulation exercise. Do this by CIRCLING the appropriate point on the scale shown below.

NO UTILIZATION	SLIGHT UTILIZATION	MODERATE UTILIZATION	SUBSTANTIAL UTILIZATION	COMPLETE UTILIZATION

3

Please indicate your current attitude toward numerical and quantitative methods of evaluation as a useful tool in decision-making by CIRCLING the appropriate point on the scale shown below.

EXTREMELY UN-FAVORABLE	MODERATELY UN-FAVORABLE	SLIGHTLY UN-FAVORABLE	NEUTRAL OR IN-DIFFERENT	SLIGHTLY FAVORABLE	MODERATELY FAVORABLE	EXTREMELY FAVORABLE

BIOGRAPHICAL NOTE

James R. Miller III was born in Philadelphia, Pennsylvania, on 21 December 1937. He attended Chestnut Hill Academy, Philadelphia, Pennsylvania, and St. Paul's School, Concord, New Hampshire, graduating in 1955. He received a Bachelor's Degree from Princeton University in 1959 and a Master's Degree in Business Administration from the Harvard Business School in 1962. He has been working at Massachusetts Institute of Technology as a graduate student and staff member since 1963.