

**A Comparative Study  
of Some  
Multiple-Criteria Methods**

# The Economic Research Institute

at the Stockholm School of Economics  
Sveavägen 65,  
Stockholm

## Basic Orientation

The Economic Research Institute, (EFI), at the Stockholm School of Economics is devoted to the scientific study problems in management science and economics. It provides research facilities for scholars belonging to these disciplines, training advanced students in scientific research. The studies to be carried out by the Institute are chosen on the basis of their expected scientific value and their relevance to the research programs outlined by the different sections of the Institute. The work of the Institute is thus in no way influenced by any political or economic interest group.

### Research Divisions:

- A Management and Personnel Administration
- B Finance and Accounting
- C Managerial Economics
- D Marketing Management—Strategy and  
Economic Structure
- F Public Administration
- G Economic Geography
- I Applied Information Processing
- P Economic Psychology—Social and Individual  
Behavior
- S Macro Economics

### Program Director:

- Professor G. Westerlund
- Professor S.-E. Johansson
- Professor B. Näslund
- Professor F. Kristensson
  
- Professor T. Thorburn
- Professor C.-F. Claeson
- Associate Professor S. Persson
- Professor K.-E. Wärnebyd
  
- Professor K. Jungenfelt

### Independent Research Programme:

- |  |                               |
|--|-------------------------------|
| Programme for Studies of Participation and<br>Organizational Development | Associate Professor B. Stymne |
|--|-------------------------------|

Additional information about research in progress and published reports is described in our project catalogue. The catalogue can be ordered directly from The Economic Research Institute, Box 65 01, S- 113 83 Stockholm, Sweden.

BERTIL TELL

A Comparative Study  
of Some  
Multiple-Criteria Methods

**EFI**

THE ECONOMIC RESEARCH INSTITUTE  
at the Stockholm School of Economics  
Stockholm 1976

© EFI

UDK 65.01:159.9

65.012.1

ISBN 91-7528-053-4

LiberTryck, Stockholm 1976

## PREFACE

This report will shortly be submitted as a doctor's thesis at the Stockholm School of Economics. The research has been carried out at the Economic Research Institute at the Stockholm School of Economics, but the author has been entirely free to conduct his research in his own way as an expression of his own ideas.

The Institute is grateful for the financial support which has made this research possible.

Stockholm, March 1976

THE ECONOMIC RESEARCH INSTITUTE  
at the Stockholm School of Economics

Karl-Erik Wärneryd  
Director of the Institute

Bertil Näslund  
Program Director  
Managerial Economics



## FOREWORD

Some knowledge of the background to this study may perhaps increase the reader's understanding of the following report. When I was doing my compulsory military service with the Swedish Air Staff in 1969, I was given the task of constructing a model for evaluating the capability of the combat units of the Swedish Air Force and of developing a method for estimating the necessary components of such a model.

The problem of evaluating a combat unit is one of multiple criteria, as it is not possible to measure the capability of a unit using a single figure. Several criteria have to be applied. For example, a strike squadron may be described by the status of its pilots and its materials, using criteria such as the pilots' ability to navigate, to land, to use bombs or missiles, etc. The need for a model of this kind arose after the introduction of the Planning, Programming, and Budgeting System (PPBS) and, at the time, few methods for handling multi-criteria evaluation problems were available.

The criterion most generally used in selecting a method is the accuracy or precision of the method concerned. But when I was developing a method to solve this particular problem, I realized that decision-makers actually consider many other aspects when they select a method. For instance, they may consider the ease of understanding, the number and complexity of the necessary computations, the user's confidence in the result, etc. As these aspects have generally been neglected, I felt that it could be interesting to examine some multiple-criteria methods, paying particular attention to attributes of this kind. And this in fact became one of the major aims of the present study.

The study was suggested by Professor Paulsson Frenckner as a natural way of following up my earlier thesis for the degree of civilekonom. I spent a year at the University of California at Berkeley, where I received valuable advice from Professor Nils Hakansson.

When I returned from the US, Professor Bertil Näslund became my advisor. He actively influenced the aims and the design of the

project. His good advice and constructive criticism have been very valuable and I am extremely grateful for his support.

Dr Ingolf Ståhl deserves my warm thanks for his many comments and for the informative seminars which he has held and at which my reports have been discussed. I would also like to thank Messrs Nils-Göran Olve and Sven Faugert, with whom I have had many long and fruitful discussions in the course of the project. And I would like to thank all my colleagues at the research division of the Economic Research Institute at the Stockholm School of Economics, for the time they have generously spent reading and discussing various parts of this study.

I further want to thank Professor Erik Ruist and Mr Leif Collén, with whom I have had many rewarding discussions about statistical methods.

I have had the advantage of discussing this project with members of the faculty of the European Institute for Advanced Studies in Management, and I would like to thank most sincerely everybody there who helped me, in particular Professor Stanley Zions.

I would also like to thank all the members of the Swedish Air Force who have made this research possible. In particular I would like to mention Lieutenant-colonel Ingemar Folke, head of the Budget Department at the Air Staff, who introduced me to the PPBS and encouraged my studies of multiple-criteria methods; also Colonel Nils Hansson, and Colonel Bertil Bjäre, and Major Boo Valter Eriks-son with whom I had long discussions when I was constructing my method at the Air Staff and carrying out my experiments.

I am very grateful to Mrs. Gunilla Weidenmark, not only for typing the many versions of this report, but also for helping me to simplify my presentation; also to Mrs. Nancy Adler for her valuable assistance in transforming the final version into idiomatic English.



I am also glad to acknowledge the grants that I have received from Svenska Handelsbanken's Foundation for Social Research and from the Bank Research Institute, which together have helped to finance the project.

Finally, I would like to express my deepest gratitude to my wife, Karin, and our children, David and Helena. To them I dedicate this book.

Bromma, March 1976

Bertil Tell



## CONTENTS

	<u>Page</u>
CHAPTER 1. INTRODUCTION	1
1.1 Multiplicity of criteria	1
1.2 Decision-making	4
1.2.1 Ex-post evaluations	6
1.2.2 Differences between pre-choice and ex-post evaluations	7
1.3 Advantages and disadvantages of evaluation models	9
1.4 Plan of the book	15
CHAPTER 2. A REVIEW OF APPROACHES TO THE MULTIPLE-CRITERIA EVALUATION PROBLEM	17
2.1 Multiple-criteria evaluation models	17
2.1.1 A presentation of some models	17
2.1.2 Comparative studies	22
2.2 Estimation methods	25
2.2.1 Indirect estimation techniques	26
2.2.2 Direct estimation techniques	29
2.2.3 Other estimation techniques	30
2.3 Purpose of this study	33
CHAPTER 3. THE METHODS AND MODELS TESTED	37
3.1 Selection of methods and models	37
3.2 Regression analysis	39
3.3 Keeney's method	41
3.3.1 Model and assumptions	41
3.3.2 Estimation of the compound utility function	43
3.3.2.1 Verifying the assumptions	43
3.3.2.2 Estimating the unidimensional utility functions	45
3.3.2.3 Estimating the parameters	46
3.3.3 Some comments	47
3.4 Miller's method	47
3.4.1 Model and assumptions	47
3.4.2 Estimation of the compound utility function	49
3.4.2.1 Verifying the assumptions	49

	<u>Page</u>
3.4.2.2 Estimating the unidimensional utility functions	49
3.4.2.3 Estimating the parameters	50
3.4.3 Some comments	51
3.5 Tell's method	53
3.5.1 Model and assumptions	53
3.5.2 Estimating of the compound utility function	53
3.5.2.1 Verifying the assumptions	53
3.5.2.2 Estimating the unidimensional utility functions	54
3.5.2.3 Estimating the parameters	54
3.5.2.4 Simplifying the estimation of the parameters	59
3.5.3 Some comments	62
3.6 The consultants' method	62
3.6.1 Model and assumptions	63
3.6.2 Estimation of the compound utility function	63
3.6.2.1 Verifying the assumptions	63
3.6.2.2 Estimating the unidimensional utility functions	63
3.6.2.3 Estimating the parameters	64
3.6.3 Some comments	66
3.7 Hypotheses and expected results	67
3.7.1 Precision of a model	68
3.7.2 Other attributes	70
 CHAPTER 4. THE EXPERIMENT	 73
4.1 A brief presentation of the PPBS	73
4.2 Selection of problem	75
4.3 Ways of measuring the capability of combat units	77
4.4 Structuring the evaluation problem	79
4.5 Participants	83
4.6 Design of the study	84
4.7 Prior acquaintance with the methods	87

	<u>Page</u>
CHAPTER 5. THE INTUITIVE EVALUATION	89
5.1 Presentation of the experiment	89
5.2 Results	94
5.2.1 Consistency over time	94
5.2.2 Consistency between subjects	98
5.2.3 Conclusion	99
5.3 Measurement scales	99
CHAPTER 6. THE PRECISION OF THE MODELS - DIRECT ESTIMATION	103
6.1 Presentation of the experiment	103
6.2 Estimation of the parameters	103
6.2.1 Consistency check of the consultants' method	103
6.2.2 A comparison of Keeney's, Miller's and Tell's methods	105
6.3 Shape of the utility functions	107
6.3.1 Consistency check of Keeney's method	107
6.3.2 Comparing the utility functions of the four methods	108
6.4 The precision of the models	110
6.4.1 Correlations	112
6.4.2 Bias	115
6.4.3 Variance	116
6.4.4 Mean square error	118
6.4.5 The precision of the models and the number of criteria	120
6.5 Conclusions	121
CHAPTER 7. THE PRECISION OF THE MODELS - INDIRECT ESTIMATION	123
7.1 Presentation of the experiment	123
7.2 Analyzing the linear models	125
7.3 Correlation	125
7.4 Bias	128
7.5 Variance	129
7.6 Mean square error	133
7.7 Conclusions	136

	<u>Page</u>
CHAPTER 8. OTHER ASPECTS OF THE MODELS	139
8.1 Presentation of the experiments	139
8.2 Criteria regarding the construction of multiple-criteria models	140
8.2.1 Qualitative aspects	140
8.2.2 Quantitative aspects	143
8.2.3 Some comments made by the participants	144
8.2.4 Conclusions	145
8.3 Criteria regarding the use of multiple-criteria models	146
8.3.1 Qualitative aspects	146
8.3.2 Quantitative aspects	146
8.3.3 Some comments made by the participants	149
8.3.4 Conclusions	150
CHAPTER 9. THE EFFECT OF UNCERTAINTY ON THE SELECTION OF A UTILITY MODEL USING DIRECT ESTIMATION	151
9.1 The indirect method of estimating utility models	151
9.2 The estimation process	152
9.3 Studies of the effect of estimation errors on model performance	154
9.4 The effect of estimation errors on the selection of a utility model using direct estimation	157
9.4.1 The problem	158
9.4.2 Decision-rule	159
9.4.3 An analytical approach	161
9.4.4 A simulation approach	162
9.4.5 Method	166
9.4.6 Results	168
9.4.7 Which model will be chosen?	176
9.4.7.1 Example 1	176
9.4.7.2 Example 2	179
9.4.7.3 Conclusions	180
9.5 Summary and conclusion	181

	<u>Page</u>
CHAPTER 10. SUMMARY AND FUTURE RESEARCH	183
10.1 Summary	183
10.2 Future research	188
REFERENCES	191





## CHAPTER 1

### INTRODUCTION

In this chapter I will describe the background and motives of the following study of multiple-criteria models for decision-making. The existence of multiple criteria in business firms and non-profit organizations is noted in Section 1.1. In Section 1.2, I explain why the present study has been restricted to a single type of decision problem - what are known as ex-post evaluations; in Section 1.3, I investigate some of the arguments for and against the use of models in solving this kind of problem. Finally, in Section 1.4, I will describe the plan of the report.

#### 1.1 MULTIPLICITY OF CRITERIA

Discussion about the number of goals or criteria that are or should be considered by managers has been going on for a long time. The debate was at its most intense during the 1950's and at the beginning of the 1960's, although much was also said both before and after this period. The discussion has embraced business firms as well as non-profit organizations and agencies; it has concerned all levels of organizations and all types of problems; it has evolved from descriptive and normative theories, or from empirical studies. In a study such as this it is not possible - nor is it my purpose - to review the vast literature on this subject. My aim is briefly to review some works in order to indicate the general nature of the multiple-criteria problem. The reader is referred to works such as White (1960), McGuire (1964), and Johnsen (1968) for a more complete survey of the discussion.

Neoclassical economic theory assumed that the business firm has only one goal, profit, and that the firm tries to maximize this goal. This assumption about the goal of the firm has been opposed by many writers on various counts. From the theoretical point of view the assumption has been attacked as not providing an accurate description of managers' behavior, and as providing a poor base for normative theories.

One of the most noted alternatives was put forward by Baumol (1959), and this consisted of a minor extension of the goal of the neo-classical theory. Instead of assuming that business firms maximize their profit, Baumol assumed that they maximize sales subject to a profit constraint. He supported this assumption by his observations of managers who preferred talking about changes in sales or market share to discussing their firm's profits. Another reason for managers wishing to maximize sales rather than the profit could be that attributes such as salary, status, prestige, etc., were associated with sales rather than profit.

Other writers, for example McKean (1958), and Hitch and McKean (1961), have accepted the traditional goal of economic theory. They argue, however, that we are unable to express the outcomes of an action in terms of this ultimate goal, whether it be the satisfaction of an individual, the profits of a firm or the military worth of the military establishment. Instead we have to insert "proximate criteria" to capture the attributes of the actions. Frenchner (1953) has suggested a number of such criteria and Wilhelm (1975) has argued that their use follows from the fact that managers have incomplete information about their own actions. Field (1969) motivates the criteria by applying systems theory, and Miller and Starr (1960 and 1967) by referring to psychological theories about individual decision-making.

Modern portfolio theory provides another example of the explicit consideration of several criteria. In the theories of Markowitz (1952) and Sharpe (1964) it is assumed that investors look not only at the expected profit of the shares and bonds that might be bought, but also at their risks. This assumption has influenced modern investment theory which explicitly uses these two attributes to characterize potential investment alternatives (Hillier, 1963; Hertz, 1964).

The behavioral school represented by Simon (1957), March and Simon (1958), Cyert and March (1963) and others, has tried to describe how decisions are made. Among other things they state that only individuals can have goals; organizations cannot. Thus, the goals of the

organization will be the goals of its participants, and these goals are shaped in the course of a bargaining process which involves side payments in many forms. This means that there will be several goals, that the goals will change as new members enter the organization or as old members leave it, and that the number and the importance of the goals will change as the problems of the organization change. Starr (1974) discusses how more and more participants have gained "power" in organizations, and have thus come to influence its goals. He mentions this as a reason for managers to extend their list of objectives.

Turning to the empirical evidence we find that several empirical studies show managers taking a number of criteria into consideration when they make their decisions. Ahlberg (1972) studied decision-making in four large Swedish companies and found that they all considered more goals than profit alone.

Other examples can be found. Johnsen (1968), for instance, interviewed the founders and managers of 21 Danish firms and discovered that most of them quoted more than one goal for their firms' activities. Blau (1956) reported that several criteria were used to measure the efficiency of employment offices; Keeney (1973) used six criteria for describing the alternatives considered by the Mexican government when selecting a new airport for Mexico City; and Wallenius (1975) used four criteria to help a Belgian company select an optimal production plan.

Another empirical confirmation of the existence of several goals in business firms is given by Druckner (1954). Druckner, who had several years experience of managing business firms, lists eight areas where the firm has to maintain objectives. These are: market standing, innovation, productivity, physical and financial resources, profitability, manager performance and development, worker performance and attitude, public responsibility. He argues that one reason for not following a single objective is that to do so would make the company bankrupt.

The fact that several criteria are used to evaluate business firms

in socialist economies is quite natural, as profit is not accepted as a measure of effectiveness. Granick (1954) gives a long list of criteria that have been used in interfirm competitions in the Soviet Union. Examples are: production of finished goods in the planned assortment, planned mastery of new types of products, fulfillment of labor productivity tasks, lowering of the unit cost, and the increase in the number of worker suggestions for improvements in work methods and conditions.

In future it will become more important for business firms in the Western economies to pay explicit attention to criteria such as these. Many of them are already used in the evaluation of public agencies (see Caro, 1971, for a summary). With the expansion of the public sector in areas such as social security, health and welfare, employment security, etc. we will come to accept a wider use of multiple-performance measures to control the efficiency of public spending.

Finally, a few words should be said about one type of goal that will not be discussed in this study, namely the managers' personal goals. Empirical studies of managers' behavior have revealed that these people often pay attention to several personal goals as well as or instead of the official goals of the firm. See, for instance, Baumol (1959) and Dean (1956). Williamson (1964) showed that managers investigated the effect that certain proposed alternatives would have on their personal goals such as status, prestige, influence, etc.

We conclude from this brief review of empirical studies of managers' behavior that managers do use several criteria to evaluate proposed alternatives or to evaluate the efficiency of an action, a firm, or an agency. Many theoretical works - descriptive and normative - have also accepted that firms or managers have several goals.

## 1.2 DECISION-MAKING

In the previous section we looked at the goals of the organization and found that many theoretical works assume the manager to have several goals and that this assumption has been verified in empirical research. We will now look at the influence of several goals on

the decision-making process.

We assume in this study that the decision-maker already has a set of measurable criteria and that his problem is only to rank the set of alternatives characterized by these pre-determined criteria. Ideas about how to select criteria can be found in, for example, Miller (1966) and Raiffa (1969). Soelberg (1967) and Hill (1974), among others, point out, however, that the decision-maker is often unable to specify his criteria a priori.

This study has been restricted to individual decision-making. Group decision-making involves many factors that, at our present level of knowledge, are very difficult to study, e.g. the effects of power, threats, bluffs, and double-crosses (Rapoport, 1960). Nor can studies of group decision-making get round the problems of interpersonal utility comparisons. The reader who is interested in differences between individual and group decision-making is referred, for instance, to Marquis (1968) and Rose (1974).

Decision-making consists of the clarifying of goals, the generation of alternatives, the search for information, the evaluation of alternatives, the selection of the optimal alternative, implementation, and control (MacCrimmon, 1968; Green and Wind, 1973). These elements are not clear-cut processes performed in sequence. Instead some processes may be skipped, some may coincide, some are repeated, etc. Information is fed backwards and forwards, making interactions between the processes a natural part of any decision-making activity (MacCrimmon, 1974).

My aim here is to help the decision-maker to evaluate items or alternatives that are characterized by several criteria. These evaluations are made on at least two occasions during the decision-making process. The first evaluation is made prior to the selection of the best alternative or alternatives. Much has been written about these multiple-criteria evaluations. See for instance MacCrimmon (1973). The evaluations made during the implementation and control stages have not aroused much attention, however, except in budgetary literature. MacCrimmon (1968, p.2) writes that "although the process of decision-making centers around the element of choice, it

should include all the activities leading up to the final choice, plus a post-choice stage of implementation and control."

Because of the lack of work on ex-post evaluations, I decided to restrict my study to this kind of problem. These evaluations are characterized, in comparison with pre-choice evaluations, by certainty in the outcomes and a smaller number of alternatives to be evaluated. These characteristics will be discussed later on.

#### 1.2.1 Ex-post evaluations

The use of ex-post evaluations changes as the implementation of an alternative proceeds. Before the completion of the selected alternative the manager is interested in evaluating the alternative being implemented, to see that production follows the plan or that it is corrected in the light of new information (Heiser, 1959; Diggory, 1963). These evaluations are generally conducted at the end of every month, every quarter or at some other interval used in the budgets or production plans.

When the alternative has been completed there is generally a final evaluation. This follow-up evaluation or post-completion audit has four main purposes according to Pflomm (1963, p. 80),

- "1. To verify the resulting savings or profit
2. To reveal reasons for project failure
3. To check on soundness of various managers' proposals
4. To aid in assessing future capital expenditure proposals."

If only one criterion is used in the follow-up evaluation, it is quite simple to compare the outcome with the original plan or budget to see whether or not this has been fulfilled. But if multiple criteria are being applied, then these ex-post evaluations become even more important, since without them it will be impossible to judge whether the different outcomes represent a better (or worse) result than expected in the plan or the budget (Ridgeway, 1956).

Post-choice evaluations not only enable the manager to control the implementation of the selected alternative by changing the production

or the production plan; they also provide him with information to restart the decision-making process. If the implementation leads to unfavorable results, he may want to restudy some rejected alternatives or to generate some new alternatives. This is the feed-back role of the control process. The ex-post evaluation will also generate feed-forward information that will make the generation, evaluation, and selection of alternatives more effective in future.

Thus, even if - unlike pre-choice evaluations - the purpose of the ex-post evaluations is not primarily decision-making, these evaluations will affect future decision-making either by generating new alternatives and/or by changing one or several of the processes that are involved in the decision-making activity.

### 1.2.2 Differences between pre-choice and ex-post evaluations

There are some important differences between the evaluation made before the choice of the best alternative or alternatives and the evaluation made during the implementation and control stages. We will discuss only two differences here - the uncertainty in the outcomes and the number of alternatives to be evaluated.

At the time of the pre-choice evaluation there is uncertainty in the outcomes on most of the criteria used to characterize the alternatives. A model for use in this evaluation will have to consider this uncertainty explicitly. Hardly any field studies have been reported in which the models used formally consider both the multiplicity of the criteria and the uncertainty of the outcomes. Instead the uncertainty is introduced independently into the solution to a multiple-criteria problem solved under assumed certainty (cf. Huber, 1974).

As time passes and the selected alternative is implemented the uncertainty is reduced. Depending on the way the problem is formulated, there comes a time when no uncertainty remains in the outcomes of the implemented alternative, except for measurement errors. We will limit ourselves here to this evaluation problem.

An example may help to clarify this point. A company has approved an investment plan specifying not only what buildings are to be erected and what machinery bought over a three-year period, but also how far these activities should have progressed by the end of each year. The evaluation made at the end of the first year is without uncertainty, as the company can determine exactly how far construction has progressed and what machines have been bought. This is the kind of multiple-criteria evaluation we will be investigating in the present study. Had the company instead been interested in the possibility of fulfilling the three-year plan, there would have been an evaluation problem with uncertainty. But we will not discuss this here.

All pre-choice activities are concerned with finding the optimal alternative from a feasible set. This set may be finite or infinite. Examples of problems with finite sets are an airline's selection of a new aircraft for intercontinental routes (there are about ten possible aircraft to choose from), or a company's choice of a new computer or new type-writers. The infinite problem is more a problem of design than of choice (MacCrimmon, 1974). It may involve finding an optimal product mix or an optimal mix of financial alternatives.

Ex-post evaluations are always finite. Each alternative implemented has to be evaluated and compared with other implemented alternatives, or with some standard or budgetary value. To help the decision-maker perform these evaluations we must form a model of his preferences, such that an index will be associated with every evaluated alternative. The utility function is such a model. It is an assumed preference ordering of all possible combinations of the criteria that are considered by the decision-maker and it is assumed to be an adequate basis for describing any decision-maker's value-structure. We can never know with any certainty the exact process used by the decision-maker and thus we cannot know the correct form of the utility function. All we can do is to compare alternative models to determine which one predicts the decision-maker's evaluations best, i.e. which has the highest precision.



### 1.3 ADVANTAGES AND DISADVANTAGES OF EVALUATION MODELS

Ex-post evaluations of actions characterized by many criteria can be made by a human evaluator or by a machine. In the latter case we will have to build a model of the evaluator's preferences, but in many situations it may be advantageous to develop a model even though the evaluations are performed by a human being. We will now investigate some of the advantages and disadvantages of using such a model.

The first aspect to investigate is the quality of the evaluations: will their accuracy be greater if a model is used rather than a human evaluator?

Miller (1956) analyzed a serie of experiments and found that they indicated severe limitations on man's ability to handle several conceptual units at a time. Other experiments have shown that an increase in the number of criteria to be considered simultaneously generally results in a marked rise in the number of errors (Archer, Bourne, and Brown, 1955) and of decision time (Hayes, 1962). Several other experiments reveal that subjects exposed to objects characterized by several criteria tend to rely too heavily on a few of them only (Meehl, 1954) or that they collapse the criteria into one "good versus bad" criterion (Osgood, Suci, and Tannenbaum, 1957; and DeSoto, 1961). Other studies indicate that people pay attention to different subsets of criteria on different occasions (Shepard, 1964). There even seems to be a tendency among subjects to be aware of the way they limit the information they receive (Hoffman, 1960). Further examples of man's difficulty in handling multiple criteria are given by Yntema and Torgerson (1961), Shepard (1964), Einhorn (1972), Dawes and Corrigan (1974), Nystedt (1975) and others.

It seems evident that man finds great difficulty in handling multiple-criteria data. The problem shows similarities with Ackoff's critique of Management Information Systems (MIS), which he accuses of producing too much data and thus of making it difficult for the manager to find the relevant information (Ackoff, 1967).

Meehl (1954) and Sawyer (1966) have examined a great many attempts

to model decision-makers' preferences. They found an apparent superiority for mechanical models over the decision-makers' evaluations. In most of these comparisons man and model have received the same information. As it has to be quantitative, we can only conclude that models have been shown to be superior to man when the data is codable (Dawes and Corrigan, 1974).

The psychological experiments discussed here indicate that man's ability to handle multiple-criteria information is limited. Hence the introduction of evaluation models should improve the quality of the evaluations.

An evaluation model is seldom intended to be a substitute for humans (Keeney and Raiffa, 1972). Rather, the models should help us to make better and faster evaluations (Yntema and Torgerson, 1961; Einhorn, 1972). However, there are many ways in which models can be a complement to man. One is to let the manager evaluate the object first, and then to let the model do the same. If they come up with different values for an object, the reasons for the difference should be analyzed. If they agree, then the manager's result has been confirmed. Näslund (1975) used this approach when applying formal evaluation methods to a strategic planning problem. The same approach has been used by the Swedish Air Force, as will be illustrated later.

Apart from accuracy, another important aspect to examine is the cost. Before we can turn over the evaluations to a machine, we will have to construct a model of the evaluator's preferences. The construction costs will depend on the complexity of the chosen model and the effectiveness of the estimation method used for determining the parameters and functions of the model.

When we come to use the model we will face another cost - the cost per evaluation. It will generally be much lower for the model than for the evaluator (Yntema and Torgerson, 1961). This implies that if the model is used for one evaluation only, e.g. the evaluation of sites for a new airport, it will be more important to consider

the construction costs than to estimate the costs of using the model. In such situations the cost-advantage of a model will not be so great. If, on the other hand, many evaluations are to be made - many objects on one occasion or a few objects on repeated occasions - it will be much more important to consider the cost-advantages of the model. The introduction of an evaluation model will reduce costs, as it will then be possible for the evaluations to be made by less experienced personnel or by a machine, and the manager will be free to perform other more important tasks.

The use of a model over a long time period makes it important to investigate another cost-element besides the two discussed above, namely the cost of maintenance. Here we include the costs for adding new and dropping old criteria, re-estimating parameters and functions due to changes in the preferences of the evaluator, etc. The cost of maintenance will reduce the advantage of using models, particularly for problems in turbulent environments.

One of the main reasons for using models is, of course, the increase in speed that they allow. This is obvious, if we have many objects to be evaluated on a single occasion. If the objects are to be evaluated on repeated occasions, the use of a model will not only increase the speed; it will also become possible to include an evaluation model in a computerized system, e.g. a Management Information System (MIS), or a budgetary system. The more frequent use of computers and electronic calculators at all levels in organizations makes time-saving even easier to achieve (cf. Yntema and Torgerson, 1961).

It is often a great advantage just to carry out the structuring of the problem that is necessary for the construction of a model. This work will not only clarify the problem for the evaluator, but will also help him to understand better which criteria are important to the evaluation and how important they are. This benefit from the use of an evaluation model is discussed by Miller (1970), Keeney and Raiffa (1972), and Keeney (1975) and has been experienced by the present author during the construction of multiple-criteria evaluation models for the Swedish Air Force. The officers taking part in the modeling reported that the work had made them consider

aspects they had previously ignored, or had made them change the values they assigned to some of the criteria.

Similar effects have been noted elsewhere (e.g. Robichek and Myers, 1965). Ackoff, for example, goes so far as to say that "the principal value of planning does not lie in the plans that it produces but in the process of producing them" (Ackoff, 1970, p. 15). Eilon (1972) criticizes Ackoff's view of planning and says that there are other perhaps even more effective ways in which managers can acquire the knowledge that comes to them in the process of producing plans. But the conclusion is clear. The construction of a multiple-criteria model produces positive "spill-over" effects through the learning of the participants. These effects should not be neglected, particularly as there are often more people taking part in the modeling than there are people making the evaluations.

Hoffman (1968) and Keeney (1975) indicate that other learning effects may be generated by an evaluation model. They mention the value of making the criteria and weights explicitly known. The less experienced members of the organization may then learn from the model which criteria are important and how important they are. This may be valuable to them in other similar situations, where the model is perhaps not available or applicable. But Ridgeway (1956) is afraid that the behavioral and motivational consequence of multiple-criteria evaluation models are not properly understood, and he has called for more research in this field.

Another type of effect produced by an explicit model is the ability to bring its assumptions and preferences into an open discussion. In Sweden the Swedish Motor Vehicle Inspection Authority (SMVIA), which makes a yearly inspection of all cars that are more than two years old, and the Police, who makes random inspections of cars on the roads, use two different models (Vi Bilägare, 1975). This means that a car that has passed the SMVIA inspection may not pass an inspection made by the Police. But as two explicit models exist, it should be easy for these agencies to come together and agree on a new model.

The use of a multiple-criteria model will make the preferences of the evaluator explicitly known, and this will make it possible to delegate the evaluating job to lower levels in the organization, to less experienced people, or even to a machine. This will not only reduce costs, as we discussed above, but will also increase the possibility of decentralizing the decision-making in the organization. The advantages and disadvantages of a decentralized organization have been widely discussed in organizational literature, but certain aspects will be mentioned here.

The multiple-criteria model reveals to the divisions the preferences of the headquarters.<sup>1)</sup> Several studies have shown that different levels or groups in an organization often have different preferences (Grayson, 1960; Swalm, 1966; Hammond, 1967). The introduction of an evaluation model will help the divisions to evaluate objects or actions in greater conformity with headquarters by removing the divisions' subjective judgement (Turban and Metersky, 1971). Thus the model will motivate the divisions to achieve a better performance by making it possible for them to decide which criteria to improve (Ridgeway, 1956).

The delegation of authority to the divisions will reduce the amount of information that has to be transmitted between the divisions and the headquarters. The divisions can make the evaluations and decisions and control the implementation of the selected action. Only extreme values have to be sent to the headquarters. This is part of the idea of Management by Exception. This approach has been used in the Swedish Air Force, where the squadron leaders report between five and nine utility indices to the Air Staff. The Air Staff asks for more information, i.e. the input data to the models, only if extremely poor or extremely good utility indices are reported. This is in line with Ackoff's notion of the limited value of too much information.

Unfortunately we lose a lot of the detailed information and knowledge that exists in the divisions as a result of using this model.

---

1) Here we talk about "headquarters" and "divisions", but the two terms represent all superior-subordinate relationships.

However, it is very easy to arrange for some feed-back of this knowledge. In the Swedish Air Force, where the combat units are evaluated by a multiple-criteria model, the following procedure has been implemented to preclude any loss of information. The squadron leaders, who make the evaluation of their squadrons, are ordered to send to the Air Staff not only the result of the evaluation made by the model but also their subjective evaluations. Any big difference between the two evaluations will be analyzed by the Air Staff as it might indicate that the model should be revised. Further indications of necessary changes in the model are obtained by the suggestions that the squadron leaders are supposed to report on these occasions.

A multiple-criteria model can also be used for budgeting or planning simulations. The decision-maker determines the outcome of the criteria for alternative resource allocations. The evaluation model is then used to determine a utility index associated with each alternative.

On the occasion of a sudden one-year cut in the resources allocated to the Swedish Air Force, the Air Staff used their multiple-criteria models to determine the best allocation of the money. The alternatives were two. Either a reduction should be made in the allocation to the strike squadrons only; or it should apply to all squadrons. It was easy to determine how these reductions would affect the pilots' ability to use their aircraft. These values were then inserted into the evaluation models to give the utility indices of the squadrons, and thus the utility values of the two budget alternatives.

We have discussed various aspects of the problem of whether man or model should perform the evaluations. In many situations a number of these effects - and even combination effects - will appear. The examples from the Swedish Air Force show that a great many different advantages often follow the introduction of a multiple-criteria model.

#### 1.4 PLAN OF THE BOOK

After this short introduction to multiple-criteria decision-making, I will review in Chapter 2 some of the models that have been suggested for solving multiple-criteria evaluation problems and some of the methods that have been used for estimating the parameters and functions necessary to these models. We will see that more information is needed about several of these methods and models, particularly about aspects other than accuracy, and we will make it our aim to investigate them.

In Chapter 3 the estimation methods to be examined are selected and presented in greater detail. Chapter 4 is devoted to the real-life problem on which these methods will be tested, and in Chapters 5 to 8 I present the results. In Chapter 5 I analyze the consistency of the participants' evaluations, and in Chapters 6 and 7 the precision of the models, using one direct and one indirect technique. I discuss some other aspects of the models in Chapter 8, for example ease of use, the decision-maker's confidence in the results, and time requirements.

The effect of errors in the weights and utility functions on the selection of a utility model is studied in Chapter 9. In the last chapter I summarize the study and discuss future research.





## CHAPTER 2

A REVIEW OF APPROACHES TO THE MULTIPLE-CRITERIA EVALUATION PROBLEM

In Section 1.1 we saw that multiple-criteria problems have been recognized for quite some time. But although interest in formal methods for handling these problems is relatively recent, there is already an extensive amount of literature on the subject. Reviews focusing on different aspects have been made by e.g. Johnsen (1968), MacCrimmon (1968 and 1973), Roy (1971), Slovic and Lichtenstein (1971), Easton (1973), Green and Wind (1973), and Wallenius (1975).

In this study we will concentrate on multiple-criteria methods suitable to our evaluation problem, i.e. explicit models of the decision-maker's preferences and techniques for estimating the parameters, functions etc. used in these models. We will first look at some models that have been used in other studies (Section 2.1) and then turn to some suggested estimation methods (Section 2.2). At the end of the chapter I will present the purpose of the present study (Section 2.3).

2.1 MULTIPLE-CRITERIA EVALUATION MODELS2.1.1 A presentation of some models

Many types of multiple-criteria evaluation models have been presented in management literature. It is natural to distinguish between models that simply group the evaluated objects, and models that associate a utility index with every evaluated object. As this second group of models is the most common we will review it in more detail. However, we can start by looking at some classifying models.

In one sense the classifying models are the simplest models. MacCrimmon (1973) discussed the use of such models for decision-making. For ex-post evaluations only the conjunctive and disjunctive models will be useful.

The classifying models require the evaluator to set one standard for every criterion. The multiple-criteria object is then compared, criterion by criterion, with the specified standards. The conjunctive model

classifies the object as acceptable only if it is above standard according to all criteria, otherwise it is unacceptable. The conjunctive model represents Simon's notion of "satisficing" (Simon, 1957). Mathematically we can write it

$$U(x_1, x_2, x_3, \dots, x_n) = \begin{cases} 1 & \text{if } x_i > \bar{x}_i \text{ for all } i, \\ & i = 1, 2, 3, \dots, n \\ 0 & \text{otherwise} \end{cases} \quad (2-1)$$

where  $U$  is the utility associated with an object characterized by the criterion-values  $x_1, x_2, x_3, \dots, x_n$ . The symbol  $\bar{x}_i$  is used for the standard value of the  $i$ -th criterion.

The disjunctive model is mathematically isomorphic with the conjunctive (Coombs, 1964). It labels an object as acceptable if only one of the criteria exceeds the standard. The mathematical form of the model is

$$U(x_1, x_2, x_3, \dots, x_n) = \begin{cases} 1 & \text{if } x_i > \bar{x}_i \text{ for at least one } i, \\ & i = 1, 2, 3, \dots, n \\ 0 & \text{otherwise} \end{cases} \quad (2-2)$$

These methods are very easy to use. The only problem is to determine the standards  $\bar{x}_i$ .

The conjunctive model is used in many different contexts. The Swedish Air Force uses the model to classify strike squadrons into A and B squadrons and the Swedish Motor Vehicle Inspection Authority uses a conjunctive model to classify cars at the annual inspections.

The models which we will look at next will associate a utility index with every object that is evaluated.

Probably the most common model is the linear model

$$U(x_1, x_2, x_3, \dots, x_n) = \sum_{i=1}^n a_i x_i \quad (2-3)$$

where  $x_i$  is a measure of the extent to which the evaluated object possesses the  $i$ -th criterion, and  $a_i$  is a parameter expressing the relative importance of this criterion.

To give some examples of the use of the linear model we can mention Nadler, Huber, and Sahney (1967), who evaluated the quality of the patient care provided in various medical wards, Dawes (1971), who predicted the success of the candidates in a Ph.D. program, and Terry (1963), who suggested using the linear model for the selection of new products. Moore and Baker (1969), and Souder (1972) used it for the selection of R&D projects.

The linear model (2-3) has been criticized on various grounds. MacCrimmon (1968) points out the many risks of a misunderstanding of the implications of the compensations implied by this model. Hoffman (1968) criticizes the model because it does not provide a good description of the evaluator's judgemental process, and he suggests some methods for analyzing the data to discover nonlinear, configural components. Green (1968) claims that the linearity is a function of the analysis rather than an inherent property of the data, while Dawes and Corrigan (1973) declare that the descriptive success of the additive model is largely an accidental property of the model.

Hoffman (1968) believes that additive models or additive models with interaction terms will be more useful than the linear model, as they represent configural components of the data better. The additive model is written

$$U(x_1, x_2, x_3, \dots, x_n) = \sum_{i=1}^n a_i u_i(x_i) \quad (2-4)$$

and the additive model with interaction terms

$$\begin{aligned} U(x_1, x_2, \dots, x_n) = & \sum_{i=1}^n a_i u_i(x_i) + \sum_{i=1}^n \sum_{j>i}^n a_{ij} u_i(x_i) u_j(x_j) + \\ & + \dots + a_{123\dots n} u_1(x_1) u_2(x_2) \dots u_n(x_n) \end{aligned} \quad (2-5)$$

The  $u_i$  is some algebraic function of  $x_i$  that is generally interpreted as the unidimensional utility function of the criterion  $i$ . These models have been used by Keeney to solve many problems including a blood-bank inventory control problem (1972) and the location of the new airport for Mexico City (Keeney, 1973, and

de Neufville and Keeney, 1972), and to operationalize the goals of a firm (1975). The additive model has also been used by Miller (1970). Kort (1968) applied the additive model with interaction terms to a juridical problem, but found very few significant interaction terms. In these problems continuous  $u_i$ -functions have been used. Step-wise linear utility functions have been used by the present author in evaluation models for the Swedish Air Force.

Many other models have been suggested as providing a better reflection of the cognitive process of the decision-maker. Einhorn (1970) took the conjunctive and the disjunctive models of Coombs and Kao (1955) and translated them into a continuous mathematical form.

The conjunctive model is often referred to as the minimum model, since the object must show a certain minimum level on all criteria. The value of an object is thus mainly determined by the criterion with the lowest score, and no compensation is possible as it is in the additive model.

Coombs and Kao (1955) take an example from Bennett to illustrate a use of the model. A student is taking a history test in French. He must know enough French to understand the questions. But he also has to know some history, because even great ability in French will not help him to answer the history questions. If, on the other hand, the student has considerable knowledge of history, this will not help him to understand the questions if he does not know enough French. To know some French and some history is better than having specialized knowledge in one of the subjects.

The continuous conjunctive model is written

$$U(x_1, x_2, \dots, x_n) = \prod_{i=1}^n u_i(x_i)^{a_i} \quad (2-6)$$

Moore and Baker (1969) have tried out this conjunctive or multiplicative model on the selection of R&D projects. They compared this model with a linear model (2-3) and with some other models. Goldberg (1971) tested the multiplicative model together with several

other models on representing the judgemental process made by clinical psychologists, and Einhorn (1971) tested this model on two problems. Mertz and Doherty (1974) have used the conjunctive model to predict college success. In all these studies the  $u_i$ -functions have been assumed to be linear. Huber (1968), however, used some nonlinear functions in studying the cost-effectiveness of a transportation system for moving military cargo.

The minimum model

$$U(x_1, x_2, x_3, \dots, x_n) = \min_i u_i(x_i) \quad (2-7)$$

can be considered a special case of the conjunctive model. It has been used by the present author in subsystems of models evaluating the capability of combat units.

Using a minimum model in subsystems of a greater model is a similar although more general way of handling the multiple-criteria problem, compared with maximizing the utility of one criterion subject to constraints on the other criteria.

Einhorn (1970) has called the disjunctive model a maximum evaluation model, since an object is evaluated on its best criterion more or less regardless of the values of the other criteria. The standard example of this model is the selection of football players for a team whose coach wants someone who can kick or run or pass with considerable skill.

Einhorn formulates this model as

$$U(x_1, x_2, \dots, x_n) = \prod_{i=1}^n \left( \frac{1}{c_i - x_i} \right)^{a_i} \quad (2-8)$$

where  $c_i > \max(x_i)$ . This model has been tested in some comparative studies by Einhorn (1971), Goldberg (1971), and Mertz and Doherty (1974).

The conjunctive and disjunctive models pay great attention to the

criterion with the lowest or the highest value respectively. Thus these models assign a higher utility index to an object with small differences in its criteria than to an object with large differences. Distance models belong to a similar class. They use the deviation from an ideal value instead of the criterion value. These models also favor objects with small differences between the deviations over objects with large differences.

In the process of selecting an advertizing agency for a company, Easton (1965 and 1966) used six criteria to describe the agencies. He then used a multiple-criteria model as

$$U(x_1, x_2, x_3, \dots, x_n) = \sqrt{\sum_{i=1}^n a_i^2 (c_i - x_i)^2} \quad (2-9)$$

to evaluate the firms. In this example the ideal point,  $c_i$ , was equal to the maximum value of each criterion, that is 1.

These are the most common models, but two more models will be presented here, namely the exponential model

$$U(x_1, x_2, x_3, \dots, x_n) = \prod_{i=1}^n e^{a_i x_i} \quad (2-10)$$

and the logarithmic model

$$U(x_1, x_2, x_3, \dots, x_n) = \sum_{i=1}^n a_i \log x_i \quad (2-11)$$

Goldberg (1971) has tested these two together with some other models, and Thurstone (1931), motivated by some psychological postulates, used the logarithmic model in an attempt to map indifference curves between bundles of commodities.

### 2.1.2 Comparative studies

During the last few years some comparative studies of multiple-criteria methods have been made. They have all used regression analysis to estimate the models, and in all cases the  $u_i$ -functions

have been linear.

The linear model (2-3) has proved an excellent model in many situations. Even when there have been interactions or nonlinear relationships the model has produced very accurate evaluations.<sup>1)</sup> This implies that the interactions account for a very small part of the total variance, and that most of the variance may be explained by the main effects (Yntema and Torgerson, 1961).

Ashton (1974), using six dichotomized criteria to describe the strength of internal payroll control, found that the percentage of the variance explained by the six main effects was 80.2 % on an average, while the 15 two-factor interaction terms accounted on an average for only 6.4 % of the variance. Ashton found that, on an average, the least important main effect explained six times as much variance as the most important interaction term. Other studies have come to similar conclusions regarding the relative importance of the main effect and the interaction terms. Thus, for example, Hoffman, Slovic, and Rorer (1968) report that about 90 % of the variance is attributable to the main effects, and in studies by Slovic (1969) and Slovic, Fleissner, and Bauman (1972) even smaller interaction effects than Ashton's have been reported.

Einhorn (1971) tested the disjunctive, the conjunctive, and the linear models (2-8, 2-6, and 2-3) on two different problems, one concerning job preference and the other concerning the selection among applicants to a graduate school of psychology.

Einhorn did not choose his data at random; he chose it to show differences between the tested models if such differences existed. In the job preference study he found that the conjunctive model was superior to the linear model for a majority of the judges, while the disjunctive model had very low predictive power.

---

1) The precision of a model is generally defined as the conformity with either the subjects' evaluations or real data. Precision of models will be discussed further in Section 3.7.1.

In the graduate selection study the disjunctive model represented the data much more satisfactorily. But many judges were still represented better by a non-linear model than by a linear model, although the difference was not as striking as in the job-preference study.

Einhorn also varied the number of criteria in an attempt to study the effect of model-selection as a function of the number of criteria. He found that the precision of all models decreased, but not that the preferences for any models changed significantly. But as he did not carry out any test-retest studies, the reduction in precision could equally well be due to a lower degree of precision on the part of the judges.

Goldberg (1971) compared five models, namely the linear (2-3), the conjunctive (2-6), the disjunctive (2-8), the exponential (2-10), and the logarithmic (2-11). He found that for 25 out of 29 judges evaluating 861 MMPI-profiles, the linear model was the best.<sup>1)</sup> For the other four judges the logarithmic model proved slightly better. The other three models performed inadequately. Nystedt and Magnusson (1975) tested some linear models, the conjunctive model, and the disjunctive model on three different problems and found the linear models outperforming the other models very distinctly. Schwartz, Vertinsky, Ziemba, and Bernstein (1975) also found the linear model outperforming the conjunctive and the disjunctive models.

Mertz and Doherty (1974) were interested in studying what effect variations in the correlation between the criteria had on the precision of some models. They used two criteria only, and their subjects were nine experienced high-school guidance counsellors. Among the models tested were the linear model, the linear model with interaction terms, the conjunctive model, and the disjunctive model. They found that although the linear model explained a major part of the variance for all subjects within all conditions, seven of the nine subjects were markedly nonlinear when the criteria were highly correlated. But neither the conjunctive nor the disjunctive models provided a good fit with the data. Mertz and Doherty also found that

---

1) MMPI = Minnesota Multiphasic Personality Inventory.



although the coefficient of the interaction term was high, the term had hardly any effect on the variance. For example, for one subject the interaction model accounted for less variance than the linear model, although the subject had a relatively hyperbolic response pattern.

The conclusion of the experiments seems to be that the linear model will generally fit the data very well, even if the data shows interactions or nonlinear relationships. The conjunctive and the disjunctive models appear to be inferior to the linear model, but this superiority seems to be less when the criteria are highly correlated.

The additive model and the additive model with interaction terms have not been tested in any comparative study, but from a theoretical standpoint they seem promising.

## 2.2 ESTIMATION METHODS

There are a great many different estimation techniques. Huber (1974) makes a distinction between multiple-criteria models that are estimated with the help of direct estimation techniques and those for which indirect techniques are used. A similar distinction is made by MacCrimmon (1973). Other approaches that may be used are interactive programming methods, multidimensional scaling, various mapping techniques, etc.

An indirect method derives the decision-maker's preferences from his past choices or, if no such data is available, from evaluated objects. When a direct method is being used, the decision-maker is asked to express his preferences for the criteria.

Combinations of the two approaches are possible, e.g. a direct technique can be used to establish the unidimensional utility functions and an indirect to find the weights, or vice versa. Not being able to study all methods we decided to examine the two contrasting estimation methods - direct and indirect - and thus leave these hybrid methods for others to investigate. (For an example of a hybrid model see Huber, Daneshgar, and Ford, 1971.)

We are interested here in the indirect and direct methods, and these will be reviewed in Sections 2.2.1 and 2.2.2. In Section 2.2.3 I will briefly present some other possible techniques for estimating the evaluator's compound utility function.

### 2.2.1 Indirect estimation techniques

Indirect methods for estimating multiple-criteria models involve the use of techniques such as regression analysis, discriminant analysis, or the analysis of variance (ANOVA). Here we will describe briefly the principles of regression analysis. The reader who is interested in further details is referred to Draper and Smith (1966), for example. As regards discriminant analysis the interested reader is recommended to study Cooley and Lohnes (1971), for instance, while Scheffé (1959), among others, provides a description of ANOVA.

The distinctive feature of the indirect methods is that they require data about the criteria and the utility index (the dependent variable), and that they infer the decision-maker's preferences from this data. The necessary data is obtained from past evaluations (choices) made by the decision-maker. If no data, or too little, is available, the necessary material may be produced by letting the decision-maker evaluate some hypothetical objects.

A simple linear regression model of the decision-maker's preferences may be written

$$U(x) = b'x + \epsilon = \hat{U}(x) + \epsilon \quad (2-12)$$

where  $x$  is a (column) vector of measurable criteria and  $b$  is a (column) vector of the parameters expressing the decision-maker's preferences for the various criteria.  $U(x)$  is the utility index assigned by the decision-maker to an object characterized by the criterion vector  $x$ .  $\epsilon$  is an error term accounting for the difference between the assigned value  $U(x)$  and the utility index produced by the regression model,  $\hat{U}(x)$ .

The regression model (2-12) is linear. Without any difficulty the regression approach can be applied to models involving nonlinear but

continuous transformations of the criteria, e.g.  $u_1(x_1) = x_1^2$ , although these transformations are not very usual. If, however, these transformations involve any parameters or if the decision-maker requires any non-continuous function of the criteria, he will have to estimate this function by a direct method.

If any qualitative criterion is used it will have to be quantified before we can use regression analysis to estimate our model. For this purpose a direct method has to be used to estimate the necessary transformation (utility) function of each qualitative criterion (Fishburn, 1967; MacCrimmon, 1968).

The most common way of determining the coefficients of the model is to use the method of least-squares. According to this method the coefficients are determined in such a way as to minimize the sum of the squares of the error terms. We then require the expected value of the error terms to be zero and the error terms to be independent of each other. We also demand that the variance of the error terms is independent of the values of the criteria and that it is constant.

The great disadvantage of this technique is its demand for data. This generates two problems - have we any historical data and have we enough data?

If historical data is available, the indirect technique is very easy to use. But this is not generally the case. At best we have data on the criteria, but lack information about the value of the utility index, i.e. the dependent variable (Meehl, 1954; Einhorn, 1971; Goldberg, 1971; Ashton, 1974). In most situations, however, we lack historical data altogether. This was the case when I was constructing evaluation models for the Swedish Air Force, although the squadrons had been operating for several years. Building models for new problems, for evaluating a new airport, for example, in fact renders the collection of historical data impossible.

If there is no available data on the criteria, we will have to construct a set of values instead. Besides being a very time-consuming job, the creation of hypothetical data always involves a risk. How can we prove that the data (and combinations of data) would have been

generated by the problem we are studying?

In the absence of historical data for the dependent variable, we have to present the decision-maker with a great many objects to evaluate. These objects may be described in terms of historical or hypothetical data. The number of objects to be evaluated will depend on the precision required in the estimates, but the work will generally be very time-consuming.

The second problem in using the indirect approach, apart from the possible lack of historical data, is the quantity of data - whether historical or fictitious.

Here it should be noted that there is a permanent conflict between the complexity of the model, i.e. the number of criteria and interaction terms, and the precision of the estimates. The more complex the model, the more observations will be needed in order to keep the precision at a specified level. As the number of observations is generally fixed, perhaps because we have a limited amount of historical data or a maximum amount of time to be devoted to the construction of data, we have to accept a drop in precision if we increase the complexity of the model.

But if we have historical data, there is the problem of its age. The more data we use, the greater the precision of the estimates. But more data also means (some) older data. We then run the risk that the preferences, or the constellations in the data, will have changed over this longer period of time implying that our increase in precision may be spurious.

However, the indirect approach has been widely used in medicine and clinical psychology, for example. Einhorn (1971), Goldberg (1968, 1971), and others have used regression analysis in constructing models to diagnose illnesses with the help of several symptoms. Einhorn (1971) used regression analysis in estimating a conjunctive, a disjunctive, and a linear model (2-6, 2-8, and 2-3), in two situations. One concerned a job preference problem and the other concerned the choice among applicants to a graduate school of

psychology. Goldberg (1971) estimated these three models and a logarithmic and an exponential model (2-11 and 2-10) with the help of regression analysis. The model was used to diagnose psychiatric illnesses.

Ashton (1974) estimated a linear model with interaction terms, using ANOVA. The problem was to determine the need for an external auditing of the payroll system of a firm. ANOVA has also been used by Slovic, Fleissner, and Bauman (1972), for example, in modeling stockbrokers' decision-making. Boggess (1967) has used discriminant analysis to estimate a model that would classify consumer credit applications into those with a high and those with a low risk of default, respectively, based on several personal characteristics of the borrowers.

For an extensive review of the indirect approach, see Slovic and Lichtenstein (1971).

### 2.2.2 Direct estimation techniques

In Section 2.2.1 we found that the indirect method of making evaluation models is often very time-consuming, since we generally lack information about the dependent variable (also see Hoepfl and Huber, 1970). This is one reason for the growing interest in management science for direct estimation methods.

In the case of direct techniques, an axiomatic system is taken as a base (Humphreys, 1975), and the specified model is produced if and only if the assumptions are fulfilled. Thus an important part of these methods consists of validating the assumptions, and this is done by putting some questions about them directly to the decision-maker.

The main feature of this approach, however, is the assessment of the parameters of the compound utility function. This is also done by means of direct questions regarding the decision-maker's preferences for the various criteria. Here there is no need for historical data.

The decision-maker will generally also have to assess some function for each criterion. This function may be regarded as a unidimensional

utility function or simply as a transformation.

A review of methods for assessing parameters and utility functions can be found in Fishburn (1967). Estimation techniques based on the direct approach have been developed by Keeney (1969), Miller (1970), Stanley (1974), the present author and others. Some of these will be presented in greater detail in the next chapter. Several applications have been reported by Keeney. They have been concerned with various problems, e.g. the selection of a location for the new airport of Mexico City (1973), and blood bank inventory control (1972). I have applied a direct method to the construction of models that will enable the Swedish Air Force to evaluate its squadrons.

### 2.2.3 Other estimation techniques

The methods presented in this section may be used for determining compound utility functions. But the methods are very restricted in their utilization and will produce the compound utility functions only as a secondary product. For the sake of completeness, however, they should be mentioned briefly.

One of the oldest techniques for determining a decision-maker's utility function has been mapping. Thurstone used this approach as early as 1931, when he studied indifference curves between some everyday commodities. An indifference curve is the locus of all commodity combinations that gives the consumer a certain utility, i.e. just a special way of illustrating a compound utility function. The problems to which mapping techniques have been applied all concern consumer goods, but the method can, of course, be used to establish indifference maps of any criteria.

Thurstone began by deriving five psychological postulates, of which Fechner's logarithmic law was the most important.<sup>1)</sup> The subject was

---

1) Wallis and Friedman (1942) state that Thurstone did not in fact use Fechner's law as a postulate but as a conclusion from his experimental data; he tried many different psychological postulates but found none that fitted the data as well as Fechner's.

then asked to express her preference between a bundle of hats and shoes and a reference bundle that consisted of some hats and shoes. The question was repeated for other bundles, so that it was possible to derive a region where the reference bundle was preferred and another region where the other bundles consisting of various combinations of hats and shoes were preferred. With the help of his postulates and some statistical technique, Thurstone fitted an indifference curve between the two regions and through the reference point. By repeating this procedure for other reference bundles he derived four different indifference curves.

Rousseas and Hart (1951) used a slightly different method to determine indifference curves for eggs and bacon. They asked their subjects to rank three different combinations of eggs and strips of bacon. As Rousseas and Hart thought it would put too much strain on one subject to rank several triplets, they asked each of 67 subjects to make one such ranking. Assuming homogeneity of tastes and a saturation point at 2.5 eggs and 2.5 strips of bacon, they derived slope vectors with the help of the second and third choices and saturation vectors by comparing the first choice with the second and third choices. These vectors were then used to draw the indifference curves.

The last of the mapping techniques that we will discuss here is the one used by MacCrimmon and Toda (1969). They also took everyday commodities such as ballpoint pens and money, and French pastries and money in two experiments. Their approach is very similar to the revealed-preference technique suggested by Samuelsson (1947).

MacCrimmon and Toda assume that for the two commodities considered, more is preferred to less.<sup>1)</sup> If the subject is then given one commodity bundle,  $P_0$ , we know that all bundles to the right of and above this combination are preferred to  $P_0$  and can be ruled out, i.e. the indifference curve will not pass through this region (cf. Figure 2.1).

---

1) MacCrimmon and Toda use this assumption to reduce the space of combinations. A slightly different approach is used if the assumption holds only for one commodity.

Next the subject is given a bundle,  $P_1$ , in the space that is not ruled out. If the subject prefers  $P_1$  to  $P_0$  we know that all commodity bundles to the right of and above  $P_1$  are preferred to  $P_0$  and can be ruled out. If, on the other hand, the subject prefers  $P_0$  to the other bundle, say  $P_2$ , we can rule out all combinations to the left of and below this bundle  $P_2$ .

Choosing more bundles and comparing them with the reference bundle  $P_0$ , the admissible region of the indifference curve  $P_0$  will be successively narrowed until the curve can be drawn.

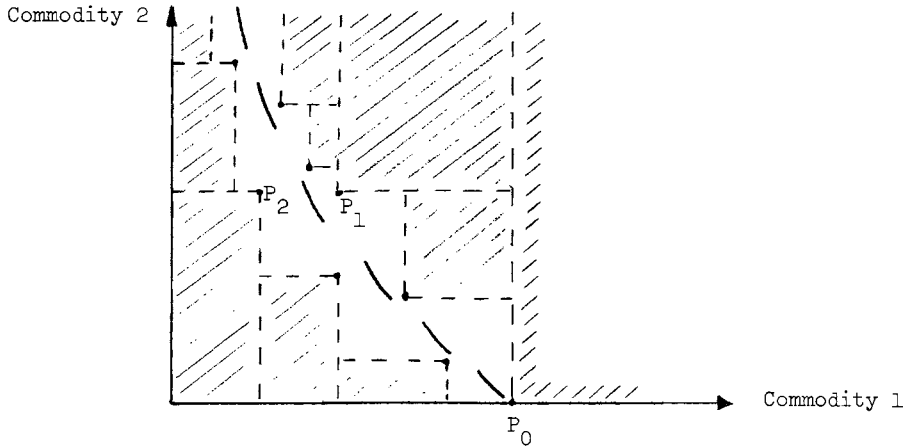


Figure 2.1. Illustration of MacCrimmon and Toda's mapping approach.

Although Thurstone studied three commodities - hats, shoes, and overcoats - he derived indifference curves for only two commodities (criteria) at a time. His technique is not suitable to problems involving more than two criteria, because the number of comparisons would be overwhelming. Even in the two-criteria case the subject had to make 256 comparisons for each indifference curve.

Rousseas and Hart's and MacCrimmon and Toda's approaches are more restrictive as they use graphical displays of the indifference curves. Consequently their techniques are not to be recommended when there



are more than two criteria. MacCrimmon and Siu (1974) have, however, built the MacCrimmon and Toda (1969) method into an interactive computer program that is not too difficult to use when there are more than two criteria. Tests on several two-criteria problems indicate that the MacCrimmon-Siu method gives indifference curves that predict the decision-maker's choices better than those resulting from the MacCrimmon-Toda method.

Interactive techniques are a group of methods that have attracted a good deal of attention during the last few years. Their advocates point out that it may not always be possible to determine the decision-maker's utility function in isolation, or that it may be unnecessarily complicated for the problem at hand. These methods have been developed chiefly for multiple-criteria design problems, but most of them can give the marginal rates of substitution between the criteria at the optimum. These marginal rates of substitution can be converted to weights that may be used in a linear model.

For reviews of interactive methods, see Näslund (1974) and Wallenius (1975).

Multidimensional scaling (see Green and Wind, 1973, or Green and Carmone, 1970, for example) is a technique for generating a multidimensional representation of alternatives based on the similarities or dissimilarities between them as perceived by the decision-maker. The object as well as the decision-maker's ideal are represented by points in this space and the distance (measured by a Euclidean or other measure) from the ideal point can be used to rank the alternatives.

This technique has been frequently applied in marketing (Green and Wind, 1973) but there are some examples of other applications. Klahr (1969), for example, compares this technique with two additive models for a graduate admission problem.

### 2.3 PURPOSE OF THIS STUDY

We are primarily interested here in two aspects of the multiple-criteria evaluation problem. The first of these concerns the models

used to help the decision-maker solve such problems and the second concerns methods of estimating the parameters and functions that are a necessary part of the models.

In recent years the direct estimation approach has been the subject of considerable attention and some new techniques have been presented in management literature. Not many applications or implementations have been reported nor, as far as I know, has any evaluation been made of these methods or the models they produce. I therefore felt it would be valuable to examine these models and methods a little more closely.<sup>1)</sup>

The most usual criterion to apply when evaluating models has been their precision or accuracy. Our discussion in Section 1.3 showed that several other aspects of models and methods also call for consideration (see also Ackoff, 1962, and Sevón, 1974). In a few studies these aspects have received some attention, for example from Dyer, Farrell, and Bradley (1973) in choosing an estimation method for a school problem, and from Moore and Baker (1969) and Souder (1972) who used several criteria in the evaluation of some multiple-criteria models for R&D project selection. Thus we will not only examine the precision of the direct models and methods investigated but will also study some other attributes such as convenience in use, believed precision, time requirements, etc.

In this study I will also analyze the accuracy of some indirect models. Comparative studies of the latter type of models have been made in the case of medical, psychological, and behavioral problems (see Section 2.1.2), but not so often of economic problems. It will therefore be interesting to see whether the results of the earlier studies will hold in such cases as well.

---

1) From now on I use the terms direct or indirect model for models estimated by a direct or an indirect method. Similarly, I will refer to a model estimated by A's method as A's model.

But my study of indirect models has in fact another purpose as well: to provide material for a general comparison of indirect and direct models. As far as I know, no study of this kind has yet been made, and such a study will of course be able to show us differences not only in models but also in estimation techniques.<sup>1)</sup>

The multiple-criteria methods and models will be examined and compared in a complex real-life evaluation problem. The problem involves several sub-problems in which varying numbers of criteria have to be considered (see Chapter 4). Thus we will be able to see how our conclusions about precision will depend upon the complexity of the problem. This is a question that has been discussed in connection with indirect models (Einhorn, 1971; Goldberg, 1971), but we will be studying it in connection with both direct and indirect models.

Without otherwise anticipating the question of how we selected our methods (Section 3.1.1), I can mention that two of the direct methods we tested were taken from the literature and two were developed in the organization in which they were to be tested. This gave me a chance to examine another aspect of multiple-criteria models, namely whether there are any differences between methods that have been useful in developing theory and those that have been more oriented towards practice.

In Chapter 6 we will analyze various aspects of the precision of the direct models and methods and in Chapter 7 we will examine the

---

1) After we had made our experiment we learned about two other studies where comparisons had been made between direct and indirect models (Summers, Taliaferro, and Fletcher, 1970; Nystedt and Magnusson, 1975). This does not make our purpose uninteresting as these studies have utilized more primitive direct methods than we will use and in any case they came up with contradictory results. Summers et al. (1970) found the indirect models superior, while Nystedt and Magnusson (1975) found the direct models to have a precision as good as or better than the indirect models.

accuracy of our selected indirect models and compare them with the direct models. In Chapter 8 we will analyze many other and important aspects of the direct methods and models. The relation between precision and the complexity of the problem is the focus of interest in Chapters 6 and 7, while differences between "theoretical" and "practical" methods will be examined in all these three chapters.

## CHAPTER 3

THE METHODS AND MODELS TESTED

In Chapter 2 I briefly presented some multiple-criteria evaluation methods and models. In this chapter I will first select the five methods to be used in our test (Section 3.1) and present them in greater detail in the following sections (3.2-3.6). I will conclude this chapter by presenting the hypotheses to be tested in our experiment and the results we can expect (Section 3.7).

To illustrate the methods chosen it seemed more fruitful to use the real-life problem on which they would be tested than to take a fictitious example. This problem concerned the evaluation of the capability of squadrons. A squadron is characterized by a number of measurable criteria. These criteria express such attributes as the pilots' ability to use bombs (measured, for example, by interception-technique-test 1, ITT1), missiles (ITT2), and guns (ITT3). The number of pilots in a squadron to fulfill the requirements of an ITT varies between the maximum number of pilots in a squadron, which is the same for all squadrons of a specified type, and zero. The problem is to determine the capability of a squadron, given only the outcomes of the interception-technique-tests, that is the values  $x_1$ ,  $x_2$ , and  $x_3$  in our example. In Chapter 4 I will present this problem in detail.

3.1 SELECTION OF METHODS AND MODELS

I decided to examine four direct-estimation techniques only. I considered this the maximum number of methods that a subject could test before tiring of the experiment, or getting the methods mixed up in the comparisons. The Air Staff, who made this whole experiment possible, wanted me to test two methods developed in the defense sector. These methods had been developed to meet an urgent need for multiple-criteria models on the introduction of the new Planning, Programming and Budgeting System (PPBS). They were geared to existing organizational demands and the problems of the use situation. On the other hand, the theoretical basis of these methods was more

restricted than that of the two direct methods I was to select from the literature.

One of the methods from the defense sector estimated purely additive models and the other estimated additive models with interaction terms. I wanted to select two direct methods from the literature that would contrast with these two "practical" methods, so that it would be possible to compare not only "practical" and "theoretical" methods but also additive models and additive models with interaction terms. I also wanted the methods to have been used, since this implied that they were in fact possible to use in practice. One method developed by Keeney (1969) and another by Miller (published as a Ph.D. dissertation in 1966 and as a book in 1970) fulfilled my criteria and were selected as the two "theoretical" methods.

One of my goals, as I explained in Section 2.3, was to use one indirect technique so that I could contrast it with the direct methods. I chose regression analysis as being the indirect technique most similar, in terms of output data etc, to the direct methods I am investigating, as well as being one of the methods most frequently used.

It is more difficult to select which models to estimate with the help of regression analysis. In the most extensive comparative study of multiple-criteria models estimated by regression analysis that has yet been made, Goldberg tested five models: the linear, the conjunctive, the disjunctive, the logarithmic, and the exponential (Goldberg, 1971). The first three of these have also been tested by Einhorn (1971). But as the conjunctive model transforms both the criteria and the utility index, Goldberg wanted to investigate the effect of transforming either only the criteria (logarithmic) or only the utility index (exponential).<sup>1)</sup> He did not carry out the same check on the disjunctive model, as it had shown very poor results in Einhorn's test. I decided to test all five of these models.

---

1) My use of the term criterion as a synonym for independent variable should not be confused with the fairly common use of the terms criterion and predictor variables for the dependent and independent variables in regression analysis.

As two of the models estimated by the direct technique were additive models with interaction terms, I decided to add a linear model with interaction terms to the models to be estimated by regression analysis. This became the sixth model in our test.

### 3.2 REGRESSION ANALYSIS

Regression analysis is one of the methods most frequently used for estimating the decision-makers' preferences. It was described in Section 2.2.1.

The six models to be estimated by regression analysis were selected in Section 3.1 and are presented in Table 3.1, where I list the conceptual formulas used when presenting the models in Chapter 2. In Table 3.1 I also present the computing formulas. These differ from the former for two reasons: first, because some models are easier to estimate after transformation of the variables, e.g. taking the logarithms of the variables of the conjunctive model; secondly, because of my real-life problem some restrictions on the parameters have to be added to the conceptual formulas. These restrictions, which will be discussed below, are incorporated in the computing formulas.

We know that all parameters of this problem, except those of the interaction terms, must be non-negative, since in the case of all our criteria more is preferred to less. Further, we require the utility index to assume a zero value when all criteria are at their minimum level (zero), and to assume a value of one when all criteria are at their maximum level (one).<sup>1)</sup> Taking these restrictions into account, we can now transform the conceptual models into the computational models presented in Table 3.1. For example, in the case of the linear models the restrictions imply that there will be no intercept and that the sum of the parameters will be one, whereas the restrictions will leave the conjunctive model unaffected.

- 
- 1) Before carrying out the regressions, we have transformed the criteria from a scale from zero to (approximately) 15 to a scale from zero to one. This has been done to facilitate comparisons of the parameters in models estimated by the direct and models estimated by the indirect techniques.

Eq. No.)	Conceptual formula	Computational formula <sup>†</sup>
Linear (3-1)	$U(x) = \sum_{i=1}^n b_i x_i$	$[U(x) - x_n] = \sum_{i=1}^{n-1} b_i (x_i - x_n)$
Conjunctive (3-2)	$U(x) = \prod_{i=1}^n x_i^{b_i}$	$\log U(x) = \sum_{i=1}^n b_i \log x_i$
Disjunctive (3-3)	$U(x) = \prod_{i=1}^n \left( \frac{1}{c_i - x_i} \right)^{b_i}$	$\log U(x) = \sum_{i=1}^n b_i \left[ \frac{1}{1.01 - x_i} - 0.990099 \right]$
Logarithmic (3-4)	$U(x) = \sum_{i=1}^n b_i \log x_i$	$[U(x) - 1] = \sum_{i=1}^n b_i \log x_i$
Exponential (3-5)	$U(x) = \prod_{i=1}^n e^{b_i x_i}$	$[\log (U(x)+1) - x_n] = \sum_{i=1}^{n-1} b_i (x_i - x_n)$
Linear with interaction terms (3-6)	$U(x) = \sum_{i=1}^n b_i x_i + \sum_{i=2}^n b_{n+i-1} x_1 x_i +$ $+ \sum_{i=3}^n b_{2n+i-3} x_2 x_i + \dots + b_{n^2-1} x_1 x_2 x_3 \dots x_n$	$[U(x) - x_n] = \sum_{i=1}^{n-1} b_i (x_i - x_n) + \sum_{i=2}^n b_{n+i-2} (x_1 x_i - x_n) +$ $+ \sum_{i=3}^n b_{2n+i-4} (x_2 x_i - x_n) + \dots + b_{n^2-2} (x_1 x_2 x_3 \dots x_n - x_n)$

† All  $b_i \geq 0$ ,  $i=1,2,\dots,n$

Table 3.1 The six models estimated by regression analysis ( $x$  is a vector of the criteria).



The requirements of our models make it impossible for us to use standard programs for regression analysis. Instead we employ a program for intercept-free regressions developed at the Stockholm School of Economics.

In carrying out the regressions, we will often find that we get negative regression coefficients. If this happens, we exclude the particular criterion from the model of the subject concerned, i.e. we set the parameter of this criterion at zero, and re-estimate the regression coefficients. We will continue to do this until non-negative regression coefficients only are obtained.

### 3.3 KEENEY'S METHOD

As was mentioned in Section 2.2.3, it is very difficult to derive a person's utility function by varying more than two criteria at a time. As it is relatively simple to assess unidimensional utility functions, Keeney (1969) derives the assumptions necessary for re-writing the compound utility functions as

$$U(x_1, x_2, x_3, \dots, x_n) = f[u_1(x_1), u_2(x_2), u_3(x_3), \dots, u_n(x_n)] \quad (3-7)$$

Now the decision-maker only has to determine the  $n$  unidimensional utility functions  $u_i(x_i)$  and the compound utility function  $f$ . If the assumptions are met, this function will be very simple and the decision-maker will in fact only have to assess scaling factors or weights for the criteria.

#### 3.3.1 Model and assumptions

The two basic assumptions in Keeney's approach are called preference independence and utility independence.<sup>1)</sup> The preference independence concerns the ordinal preferences between the criteria only, while the utility independence concerns the decision-maker's cardinal preferences.

In order to deal with the independence assumptions, we will have to introduce some notations. Let  $X = X_1 \times X_2 \times X_3 \times \dots \times X_n$  be the set

---

1) Raiffa (1969) calls them weak conditional utility independence and strong conditional utility independence, respectively.

of all possible outcomes  $X = (X_1, X_2, X_3, \dots, X_n)$  and let  $X_{ij} = X_i \times X_j$  be a subset of  $X$ . Designate the complement to the subset  $X_{ij}$  by  $\overline{X_{ij}}$ , and let  $x_{ij}$  and  $\overline{x_{ij}}$  be members of these sets.

We can now say that  $X_{ij}$  is preferentially independent of  $\overline{X_{ij}}$  if  $(x_i, x_j, \overline{x_{ij}})$  is preferred to  $(x'_i, x'_j, \overline{x_{ij}})$  for all values of  $x_i, x_j, x'_i, x'_j$ , and  $\overline{x_{ij}}$ . This implies that  $u(x_i, x_j, \overline{x_{ij}}) > u(x'_i, x'_j, \overline{x_{ij}})$  for all values of  $x_i, x_j, x'_i, x'_j$ , and  $\overline{x_{ij}}$ . This also means that the ranking of pairs of values of the criteria  $(x_{ij})$  must be identical for all  $\overline{x_{ij}}$ . Thus it will be meaningful to talk about substitution rates or trade-offs between  $X_i$  and  $X_j$  without discussing the value of  $\overline{x_{ij}}$ .

While preference independence refers to pairs of criteria, the utility independence deals with one criterion at a time. A criterion  $X_i$  is said to be utility independent of the other criteria  $\overline{X_i}$  if the decision-maker's preferences among lotteries involving only  $x_i$ , with  $\overline{x_i}$  remaining fixed, do not depend on the value of  $\overline{x_i}$ .

Note that the independence relationships are directional. Thus, for example, we can have  $X_{ij}$  preferentially independent of  $\overline{X_{ij}}$  while  $\overline{X_{ij}}$  is not preferentially independent of  $X_{ij}$ . And we can have  $X_{ij}$  preferentially independent but not utility independent of  $\overline{X_{ij}}$ , but the converse cannot be true (Raiffa, 1969).

Keeney proves that if  $X_{ij}$  is preferentially independent of  $\overline{X_{ij}}$  for all  $i$  and  $j$ , and if  $X_i$  is utility independent of  $\overline{X_i}$  for all  $i$ , then the compound utility function may be written

$$U(x_1, x_2, x_3, \dots, x_n) = \sum_{i=1}^n a_i u_i(x_i) + K \sum_{i=1}^n \sum_{j>i} a_i a_j u_i(x_i) u_j(x_j) + \dots + K^{n-1} a_1 a_2 a_3 \dots a_n u_1(x_1) u_2(x_2) \dots u_n(x_n) \quad (3-8)$$

where  $u_i$  is a utility function over the  $i$ -th criterion scaled from zero to one,  $a_i$  are weights and  $K$  is a scaling factor deduced from the  $a_i$ 's so as to make the compound utility function assume values in the interval zero to one.

When  $\sum a_i = 1$  then  $K = 0$  and the compound utility function may be simplified to

$$U(x_1, x_2, x_3, \dots, x_n) = \sum_{i=1}^n a_i u_i(x_i) \quad (3-9)$$

i.e. to an additive form.

When, on the other hand,  $\sum a_i \neq 1$  then  $K \neq 0$  and the compound utility function will be additive with interaction terms as in (3-8). This is a special case of the additive model with interaction terms presented in Chapter 2 as it requires only  $n$  parameters to be estimated and not  $2^n - 1$  as the model (2-5). Because of these special properties it is possible to rewrite Keeney's model in a multiplicative form,

$$K \cdot U(x_1, x_2, x_3, \dots, x_n) + 1 = \prod_{i=1}^n [K \cdot a_i u_i(x_i) + 1] \quad (3-10)$$

### 3.3.2 Estimation of the compound utility function

The estimation of the compound utility function consists of three separate parts: (1) the verification of the assumptions, (2) the estimation of the unidimensional utility functions, and (3) the estimation of the parameters or weights by which the utilities of the criteria will be amalgamated into one figure. This partition of the estimation process will be made for all the direct methods.

#### 3.3.2.1 Verifying the assumptions

Before estimating the compound utility function we have to verify the two basic assumptions. As we use the utility function for ex-post control, we do not have to assess any probability distributions over the outcomes of the various criteria.

Let us, as an example, check whether ITT1 ( $x_1$ ) and ITT2 ( $x_2$ ) are preferentially independent of the other attributes, in this case only ITT3 ( $x_3$ ). We will ask the decision-maker to consider two squadrons of which one has two pilots who have succeeded in doing ITT1 and 14 pilots who have managed to do ITT2, while the other squadron has  $x_1$  (that is to be determined) and three pilots, respectively,

who have successfully passed ITT1 and ITT2. The number of pilots who managed ITT3 successfully was the same in both squadrons and remained at a high level, say 13. The decision-maker is asked to determine  $x_1$  such that he is indifferent between the two squadrons. Assume he answers "12 pilots". The question will now be repeated for other levels of  $x_3$ , i.e. the number of pilots able to manage ITT3. If our decision-maker assigns "about" the same value to  $x_1$  for any value of  $x_3$ , we conclude that  $x_1$  and  $x_2$  are preferentially independent of  $x_3$ . This procedure should be followed with all pairwise combinations of the criteria.

The next verification concerns the utility independence assumption. Let us find out whether ITT1 ( $x_1$ ) is utility independent of the other criteria,  $x_2$ , i.e. ITT2 ( $x_2$ ) and ITT3 ( $x_3$ ). Once again we ask the evaluator to consider two hypothetical squadrons. One has  $x_1$  pilots who can use the interception technique tested by ITT1, while in the other squadron there is a 50-50 lottery between 0 or 15 pilots who are able to perform this interception technique. 1), 2) The two squadrons both have 15 pilots who are able to use the interception techniques tested by ITT2 and ITT3. The evaluator answers "nine". The question is repeated for other values of  $x_2$  and  $x_3$ . If the evaluator puts the value nine at  $x_1$  for all combinations of  $x_2$  and  $x_3$ , then we may say that ITT1 is utility independent of the other criteria.

This procedure should be followed for all criteria.

- 
- 1) Assume here and throughout this study that the maximum number of pilots in a squadron is 15. The real number varies with the type of squadron.
  - 2) This type of lottery questions becomes cumbersome to interpret, since skill is not an ability that changes drastically from one day to another. Traveling time, number of accidents, etc. may be due to external events and may vary very much, thus giving some sense to the question.

By using the formulation "being able to" we indicate that the low values may be due to external factors and not to a sudden decrease in the pilots' abilities. These factors may be malfunctions in the aircraft or in the weapons, disabilities among the crews, bad weather, etc.

### 3.3.2.2 Estimating the unidimensional utility functions

If we have found all criteria preferentially as well as utility independent of the others, we may assess the unidimensional utility functions of the various criteria. The technique used by Keeney is the standard lottery technique discussed by Raiffa (1968) and Schlaifer (1969), for example. We will illustrate this technique for our multiple-criteria model.

In our experiment the maximum and minimum values of the criteria are known and need not be determined. They are 15 and 0 for all criteria as they all deal with the pilots' ability to use their planes. Thus we know that, for example,

$$*u_i = u_i(15) = 1 \text{ and } *u_i = u_i(0) = 0 \text{ for all } i, i = 1, 2, \dots, n.$$

We now present our evaluator with a 50-50 lottery as follows. Consider two hypothetical squadrons. One has  $x_1$  pilots who can use the interception technique tested by ITT1, while in the other there is a 50-50 chance of either 0 or 15 pilots being able to apply this technique. How many pilots ( $x_1$ ) do you require in the first squadron, to be indifferent between these two squadrons? Suppose the evaluator answers "nine pilots".<sup>1)</sup>

We can now determine the certainty equivalent as

$$u_1(9) = .50u_1(15) + .50u_1(0) = .50$$

By asking more questions we can determine more values of  $u_1(x_1)$ . If we plot these values in a diagram, it is often possible after five or six questions to draw a smooth curve through the points representing the decision-maker's utility function for this criterion.

Using the same procedure we can determine utility functions for all the other criteria.

---

1) Note that this question is in fact the same as the one used for verifying utility independence except for the omission of the other criteria.

### 3.3.2.3 Estimating the parameters

Finally we will determine the weights. Several methods for doing this have been presented (Keeney, 1969; Raiffa, 1969), but for this experiment we decided to apply the technique used by Keeney (1973) in determining the compound utility function for the location of a new airport for the Mexico City metropolitan area.

A lottery technique is also used here. The evaluator has to choose between two squadrons. All pilots of the first squadron can use the interception techniques tested by ITT1, but none of them can apply the interception techniques tested by ITT2 and ITT3. All the pilots in the second squadron can use all the interception techniques, but the probability of this happening is only  $p_1$ . With a probability of  $1-p_1$ , none of the pilots will be able to apply any of the interception techniques.<sup>1)</sup> The decision-maker is asked to specify the probability  $p_1$ , such that he will be indifferent between the two squadrons.

This question is repeated for the other criteria.

It is easy to show that the parameter  $a_1$  in equation (3-9) or (3-10) must equal  $p_1$  (see e.g. Keeney, 1973).

If  $\sum a_i = 1$ , then we have fully determined our utility function which will be the same as (3-9). However, if  $\sum a_i \neq 1$ , then we need to determine the scaling constant  $K$ . This is a simple mathematical task that does not involve the decision-maker.

Insert the values of the weights  $a_i$  into equation (3-10) and evaluate for  $u_1(x_1) = u_2(x_2) = \dots = u_n(x_n) = 1$  for which  $U(x_1, x_2, x_3, \dots, x_n) \equiv 1$ . We will get

$$K \cdot U(x_1, x_2, \dots, x_n) + 1 = \prod_{i=1}^n [K \cdot a_i u_i(x_i) + 1] \quad (3-11)$$

and

$$K + 1 = \prod_{i=1}^n (K \cdot a_i + 1) \quad (3-12)$$

---

1) In the first squadron there may be malfunctions in the weapons or in parts of the aircraft (e.g. the sight) used for the interception techniques tested by ITT2 or ITT3. The second squadron will be "grounded" because all the aircraft, all the weapons, the runways etc. will be out of order.

which is an equation with only one unknown variable. Keeney (1969) has shown that  $K > 0$  if  $\sum a_i < 1$  and that  $-1 < K < 0$  if  $\sum a_i > 1$ .

### 3.3.3 Some comments

Keeney's method has at least two disadvantages when it comes to implementation. One is the lottery question per se and the other concerns the chosen outcomes of such a lottery.

My experience of using the lottery approach in courses in business administration is that people - in this case business students - find it hard to grasp the lottery question. Similar distrust of the approach has been expressed by Becker and McClintock (1967). When Keeney uses the same type of question for finding the parameters as well as the unidimensional utility functions of multiple-criteria problems, it seems likely that the users will find it even harder to understand and answer the lottery questions. Their belief in the evaluations made by this model will probably be negatively affected by the questions themselves.

The outcomes of the lotteries often involve at least one of the most or the least preferred outcomes. From a mathematical standpoint this is justified, since a line between two points will be estimated with less uncertainty if the two points are far apart than if they are close together. But in practice it will be necessary to consider the psychological aspects of the approach as well. Will these very extreme outcomes have any meaning for the decision-maker? Wouldn't we get better estimates if we chose values in a region where the decision-maker is more familiar with the outcomes? This problem has been touched upon by Schlaifer (1969).

## 3.4 MILLER'S METHOD

### 3.4.1 Model and assumptions

Miller (1970) starts his analysis by stating that the decision-maker wants the alternatives or objects that he evaluates to have high ratings in the positive and low in the negative attributes. These attributes, often very general statements of desirability, are called objectives. This

means that, for example, a fighter squadron may be described by the objectives interception capability at VMC, interception capability at IMC, interception capability in a jammed environment, and ability to accomplish other missions.<sup>1)</sup>

These objectives are very general and therefore a more detailed specification of them is necessary. This is provided by a process of successive subdivision of the objectives into lower-level criteria. Very often even these lower-level criteria have to be decomposed into further lower-level criteria. This process continues until the decision-maker finds that an adequate specification of the objectives has been obtained.

Then the decision-maker determines a physical performance measure at each lowest-level criterion of the pyramidal criterion-structure. He has thereby connected his inner-mind criteria, describing what he wants to have, with the physical world and the measurable attributes of the objects to be evaluated. Thus the overall objectives have been broken down into a hierarchical structure of successively more specific criteria, until measurable performance measures have been reached.

To illustrate the decomposition of the objectives we can look at the objective "interception capability at VMC" in our fighter squadron. It may be broken down into, let us say, high altitude, medium altitude, and low altitude interception technique, since the selection of an interception method depends among other things on the altitude. In this case the Air Force did not find it necessary to specify the criteria further, because there were three tests - ITT1 to ITT3 - that were accepted as adequate measures. Thus the ITT:s provide our physical performance measure.

Miller demands that the criteria be utility (worth) independent and the trade-offs between them to be constant. This implies the use of an additive model written as

---

1) VMC stands for Visual Meteorological Conditions and IMC for Instrument Meteorological Conditions.



$$U(x_1, x_2, x_3, \dots, x_n) = \sum_{i=1}^n a_i u_i(x_i) \quad (3-13)$$

We will now look at Miller's method for estimating the weights and the unidimensional utility functions.

### 3.4.2 Estimation of the compound utility function

#### 3.4.2.1 Verifying the assumptions

Before estimating the compound utility function we will have to check for utility independence. This check is performed pairwise with all combinations of criteria belonging to the same set. For each pair the decision-maker is asked if his willingness to accept reduced satisfaction on either subcriterion in return for increased satisfaction on the other is influenced by the degree of satisfaction already obtained on each. If his willingness is not affected the two criteria are considered utility independent.

#### 3.4.2.2 Estimating the unidimensional utility functions

The purpose of the unidimensional utility function is to specify a relationship between each one of the lowest-level criteria and its associated performance measure. In formulating the unidimensional utility functions, the decision-maker is asked to consider only one criterion at a time.

A utility of zero is assigned to the logical lower boundary and a utility of one to the logical upper boundary. If no logical bounds can be found, a utility of one or zero is given to the indefinite value(s) of the performance measure scale. Miller adds that an unbounded performance scale can often be bounded by redefining the performance measure, e.g. from the number of spectators at the Yankee Stadium to the number of seated spectators at the Yankee Stadium (Miller, 1970, p. 66).

Once the upper and lower value of the criterion and the performance measure have been established, the decision-maker is asked some questions that will reveal the shape of his utility function for this criterion. He is asked to give the level of performance that he considers ten percent successful in satisfying the related lowest-level.

criterion. The question is repeated for 20, 30, 40, 50, 60, 70, 80, and 90 percent fulfillment of the criterion.

These eleven combinations of percentage fulfillment of the criterion and the given levels of the performance measure are plotted in a graph, and a smooth curve is drawn as close to these points as possible. Miller also discusses some ways of assessing mathematical functions for the given data if a graphical form is unsuitable. For our experiment we chose the first of his suggestions, the "approximated scoring function" which is produced by linear interpolation between the points given above.

This procedure is followed for all lowest-level criteria.

#### 3.4.2.3 Estimating the parameters

The weight-setting procedure is divided into two successive stages. In the first the decision-maker is asked to give his weight to each criterion in a set of criteria in the hierarchical structure; in the second he adjusts his weights to compensate for imperfections in the performance measures.

The first step in determining the weights is to rank the criteria from the most important to the least important in the set. Starting at the top of the list, the decision-maker then compares the criteria pairwise down the list. At each comparison the decision-maker assigns a weight of one to the first criterion of the pair and a weight lower than one (equal to one if the criteria are of equal importance) to the other, expressing the importance of the second relative to the first. Thus the decision-maker is asked to reveal his marginal rates of substitution between the criteria down his ranked list.

To find the weights of the criteria, some numerical computations have to be performed. The weight of the most important criterion was set at one. The weight of the second most important criterion is equal to the weight it was given in the process described above. The weight of the third (fourth, fifth, etc.) criterion on the list is equal to

the product of the weight it was given in the process described above and the weight of the criterion immediately above in the ranked list (second, third, fourth, etc.). If we want the weights to add up to one, we simply have to normalize them by dividing by their sum.

The procedure described here is repeated for all branch points in the hierarchical tree.

The first step in determining the weights of the compound utility function was concerned with the determination of the relative importance of the criteria. But our input data is expressed in physical performance measures which are more or less perfect measures of the lowest-level criteria. Miller introduces an explicit mechanism for the handling of these imperfections.

The decision-maker is asked to consider the relationship between each lowest-level criterion and its associated physical performance measure. To this relationship he is to assign a value characterizing the interpretative ability of the physical performance measure. A value of one hundred percent indicates a perfect interpretation of the performance measure, while a zero value means that the performance measure bears no relation at all to its lowest-level criterion. Values in between indicate a less than perfect relationship.

After multiplying each of these quality measures by their corresponding weights and normalizing the weights, we will obtain the final weights for use in our additive model.

### 3.4.3 Some comments

At first glance it might seem valuable to be able to adjust the weights in accordance to the interpretative quality of the performance measures. But a closer analysis reveals some negative effects.

Miller claims that large differences in relative interpretative quality can seriously distort a decision. It cannot be right, he argues, that one important criterion which is heavily weighted and

measured by a very poor performance measure should overwhelm one or several other less important criteria which are easily interpreted by high-quality performance measures.

Two criticisms of Miller's procedure can be made. First, it is likely to be very difficult for a decision-maker to assign weights to broad and vague criteria which are not well specified. Miller must agree with this, as he suggests the decomposition of the broad objectives into better specified criteria. But the lowest-level criteria are still not connected with any performance measures and are, in that sense, still vague. Thus, asking the decision-maker to reveal his weights for these vague criteria means that he will try to find some performance measures associated with these criteria in order to grasp the problem. He will choose some performance measures either implicitly or by using those determined in the process of structuring the problem, depending on the order of the different steps. The decision-maker will use these performance measures instead of the criteria when assigning his weights and so he implicitly incorporates the interpretative quality already when he assesses the weights.

The other difficulty in Miller's approach is that if one of the criteria is easy to measure while the others are measured with less perfect instruments, the adjustment of the weights may convert the multiple-criteria problem into a unicriterion problem. This will happen very often, since the cost criterion is one that is often considered and it is easy to measure. This means that in fact we lose the whole point of using many criteria.

Another object of criticism is Miller's pairwise estimating procedure. The weights generated by pairwise comparisons may be quite different from those resulting from a comparison of all the criteria at once, because the decision-maker may lose the overall perspective. A comparative study by Eckenrode (1965) indicates that there are no differences in the weights generated by paired comparisons and five other estimation techniques while one recent study by Thiriez and Houri (1975) supports our hypothesis of differences in weights due to differences in estimation methods.

The pairwise approach may also lead to incorrect estimates of the parameters, as it is greatly affected by the decision-maker's preferences for "simple" numbers such as 1/3, 1/4, 1/5, etc. Thus the decision-maker in fact uses a very crude measuring device as he chooses between 1/3 and 1/4, for example, without considering intermediate values such as 4/13, 2/7, 3/11, etc. The decimal system produces similar errors, as the decision-maker tends to use "simple" numbers that are multiples of 0.05, 0.1, or 0.25.

### 3.5 TELL'S METHOD

This method was developed at the Swedish Air Force for evaluating Air Force combat units, primarily squadrons. Evaluation models for the main types of squadrons have been constructed with this method, and these models have now been in use for about four years.

#### 3.5.1 Model and assumptions

We assume the evaluator's utility function to be separable into a sum of the utilities of the various criteria. As several studies have suggested (e.g. Yntema and Torgerson, 1961), the main effects very often have accounted for most of the variance. Thus a model that uses main effects only would give us a fairly high precision while being very easy to estimate and use. Emphasizing these latter aspects of multiple-criteria models we decided to use an additive model,

$$U(x_1, x_2, x_3, \dots, x_n) = \sum_{i=1}^n a_i u_i(x_i) \quad (3-14)$$

The model assumes the criteria to be utility independent.

#### 3.5.2 Estimation of the compound utility function

##### 3.5.2.1 Verifying the assumptions

The aim of the unidimensional utility functions is to transform values of the performance scale into utility indices. If the criteria are found to be utility independent, the shape of the unidimensional utility functions can be determined one at a time and quite independently of one other.

The procedure for checking utility independence has been to determine each unidimensional utility function by the procedure presented below while assuming the other criteria to take on some specified values. We have here tried to use a set of values that are very likely so as to facilitate the estimation. Having estimated all the unidimensional utility functions we reestimated them once or twice depending on the time available but used other and less likely sets of values of the criteria. If the estimated utility functions of a criterion were similar we concluded that this criterion was utility independent of the other.

#### 3.5.2.2 Estimating the unidimensional utility functions

The first step in establishing these utility functions is to determine for each criterion one value on the performance scale that is higher than the highest expected performance value and one value that is lower than the lowest expected performance value. These two performance values are given the utility indices one and zero, respectively.

The next task is to determine the shape of the utility function between these two extreme values. This is done by asking the decision-maker to determine a value on the performance scale that gives him a 50 percent utility of the maximum utility value. This question is repeated for utility levels of 90, 75, 60, 40, and 20 percent of the maximum utility value. The distribution of utility indices around 50 percent is not symmetric as very few combat units have values that low.

For easy handling by the computer we define the unidimensional utility function by linear approximation between the points given above.

#### 3.5.2.3 Estimating the parameters

The value of an object depends on the situation in which it is being evaluated. Most objects are consumed in several time periods, or have effects that last for several periods and that should be considered when the object is being evaluated. But as a good deal of the effects, and probably much of the consumption, are going to occur

in the future, we cannot be certain at the time of evaluation of the states of nature that will prevail at these later times. Thus, to be able to assign any weights, the decision-maker will have to consider all possible states of nature, their likelihood, and the values he will assign to the criteria at these different states. The other methods do not explicitly take this into account. In this method we will explicitly discuss and evaluate the states of nature, their likelihood, and the evaluation problem at the various states of nature.

Let us illustrate the method in our fictitious example from the Air Force. Assume that the criteria used to describe a fighter squadron are the pilots' (the squadron's) ability to use guns, IR-missiles or radar-guided missiles. We further assume that there are three reliable tests of the pilots' capability - one for each criterion. We finally assume that the model of the compound utility function is additive and scaled in such a way that if all pilots know how to use their weapons we will give the squadron a utility index of one, and if no pilot knows any of the interception techniques we will give the squadron a utility index of zero. We have determined the unidimensional utility functions using the method presented in Section 3.5.2.2, and our only remaining problem is to determine the weights of our model. When this has been done, we can evaluate the capability of any squadron.

Our first step in determining the weights of the three criteria will be to determine states of nature that are suitable to the multiple-criteria problem. For a fighter squadron it was found suitable to let different types of enemy aircraft represent the states of nature, e.g. strike, bomber, transport, fighter, and other aircraft. These states are exhaustive as well as exclusive. As the states characterize future events, we cannot know which one or which ones that will occur. We can, however, attach probabilities to the various states expressing our attitudes to their occurrence. Factors influencing the probabilities of the various states are, for example, the type of war in which Sweden is expected to be involved and the varying number of aircraft of the different types that the

enemy has at his disposal and may use. This is shown in Table 3.2.

Probability of occurrence	States of nature				
	Strike	Bomber	Transport	Fighter	Other aircraft
	0.5	0.1	0.2	0.1	0.1

Table 3.2. Probability of the different states of nature,  $P(S_j)$ . Fictitious data.

Having specified the states of nature, it is now possible to determine the utility of each of the criteria or interception techniques, given the state of nature. We are interested not in the absolute utility values but only in the relative utility values. Thus it is possible to assign an arbitrary value, let us say one, to one of the elements and to express all other utilities in relation to this value. The utility values are presented in Table 3.3.

Criteria	States of nature				
	Strike	Bomber	Transport	Fighter	Other aircraft
Guns	0.8	1.2	0.6	0.6	0.5
IR-missiles	1.0	1.7	0.4	0.8	0.5
Radar-guided missiles	1.0	2.0	0.4	0.7	0.5

Table 3.3. The utility of the criteria at different states of nature,  $U(C_i, S_j)$ . Fictitious data.

In assigning these utilities we have implicitly taken certain aspects into consideration. In a more complete model, several of these aspects may be made explicit. Here we list some of them.

- (a) The utility of a destroyed enemy aircraft will not be a constant, but will vary with the type of aircraft.
- (b) The probability of success in destroying an aircraft with the various interception techniques.



- (c) The probability of not being shot down by the enemy aircraft while attacking it.
- (d) The time (or place) for destruction of the enemy aircraft.  
(It is better to destroy the aircraft before it has accomplished its mission than after, and even better to shoot it down outside Swedish territory.)
- (e) The opportunity cost of using an expensive weapon (e.g. a missile) against an aircraft that could have been destroyed more cheaply (e.g. by guns).

Had it not been for aspects like (b) - (e), the columns of the utility matrix would have been identical, i.e. there would have been no differences between the criteria, only between the states of nature.

From Table 3.3 we can see which weapon is the most effective weapon against various aircraft. We find in this example that the use of IR-missiles is the best interception technique when the enemy aircraft is a fighter. But there are many reasons why the best interception technique, i.e. the one giving the highest utility, will not always be used against one type of enemy aircraft. One is that the enemy aircraft may have countermeasures at its disposal, or may use tactics that will render the use of the best interception technique impossible. Other reasons might consist of technical restrictions in the aircraft, the equipment, or the weapons; or standard armaments may make the use of certain techniques more likely; or uncertainty at the take-off of the enemy aircraft (the type of aircraft, its countermeasures, its tactics, etc.) may call for a flexible strategy - that is, an interception technique that will work in most states of nature, etc.

This uncertainty in the pilot's selection of an interception technique against specified types of enemy aircraft will be considered explicitly in our model. This will be done by determining the probabilities of the events concerned. Table 3.4 contains the conditional probabilities of using the various criteria, given the different states of nature.

We must assume that each state of nature is decomposable into as many exhaustive and exclusive sets as there are criteria, and that every set is associated with one and only one criterion. This means that the 10-percent strike aircraft, against which we will use guns, will not be intercepted by either of the other two methods if the pilot has not mastered the guns.

Criteria	States of nature				
	Strike	Bomber	Transport	Fighter	Other aircraft
Guns	0.1	0.0	0.8	0.3	0.6
IR-missiles	0.6	0.2	0.1	0.6	0.1
Radar-guided missiles	0.3	0.8	0.1	0.1	0.3

Table 3.4. Probabilities of using the criteria, given the different states of nature,  $P(C_i|S_j)$ . Fictitious data.

Given this information, we can compute the expected utilities of the criteria or the interception techniques as

$$EU(C_i) = \sum_{j=1}^m P(C_i|S_j) \cdot P(S_j) \cdot U(C_i, S_j) \quad (3-15)$$

where

$S_j$  = state of nature  $j$ ,  $j = 1, 2, \dots, m$

$C_i$  = criterion  $i$ ,  $i = 1, 2, \dots, n$

$P(S_j)$  = probability of state of nature  $j$ ; Table 3.2

$P(C_i|S_j)$  = probability of using criterion  $i$ , given state of nature  $j$ ; Table 3.4

$U(C_i, S_j)$  = utility from using criterion  $i$ , given state of nature  $j$ ; Table 3.3

$E[U(C_i)]$  = expected utility of criterion  $i$ ; Table 3.5

Using our illustrative example where we have  $m = 5$  and  $n = 3$ , we obtain the data presented in Table 3.5.

Criteria	States of nature					Expected utility	Expected utility (normalized)
	Strike	Bomber	Transport	Fighter	Other aircraft		
Guns	0.040	0.000	0.096	0.036	0.030	0.202	0.215
IR-missiles	0.300	0.034	0.008	0.048	0.005	0.395	0.422
Radar-guided missiles	0.150	0.160	0.008	0.007	0.015	0.340	0.363

Table 3.5. Conditional and unconditional expected utilities of the criteria. Fictitious data.

Under the assumptions made above, the normalized expected utilities will fulfil the requirements of weights in our additive model. In determining them we have considered all possible future events - their likelihood, and the value of the criteria, given different states of nature.

#### 3.5.2.4 Simplifying the estimation of the parameters

It would be very cumbersome to make all the estimations required by the method in this form. Often, however, it is possible to introduce simplifications to reduce the estimation requirements substantially. In the present case - in connection with the fighter squadron evaluation problem - we will look at two alternatives.

The differences in utilities between different criteria for a given state of nature in Table 3.3 will be accounted for, among other things, by differences in the probabilities of hitting the enemy aircraft and differences in the probabilities of being shot down by the enemy aircraft. If these effects are very small, or if they are contradictory, the rows of Table 3.3 will be equal. This implies that the utility of a destroyed aircraft will be independent of the selected interception technique.

The value of a destroyed enemy aircraft does not depend only on the type of aircraft but also on the specific situation in which it is destroyed. This implies that a destroyed bomber - which can carry

three times the bomb-load of a strike aircraft - is not always three times as valuable as a strike aircraft that has been shot down. If the bomber is heading for a landing strip and the target of the strike aircraft is an important radar station, the latter aircraft may even be rated more valuable than the bomber.

This means that it is not enough to specify the types of enemy aircraft as our states of nature. For every type of aircraft we will also have to consider all possible situations where an aircraft that has been shot down will be assigned a different value than it would have been allotted in the "standard" situation. But as more states of nature have to be defined, the number of extra estimations to be made increases exponentially.

Adding more states of nature often makes it possible to introduce a simplification that greatly reduces the number of estimations. The new states are often characterized by very high utility values, as they would otherwise have been included in the "standard" situation. An example of such a new state may be bombers carrying H-bombs. The probability of their occurrence, on the other hand, is very low. For the great majority of states, however, both the probabilities and the consequences are more modest.

In situations like the one above, it is reasonable to assume the product of the utility and the probability of a state of nature to be equal for all states. This means that

$$P(S_j) \cdot U(C_i, S_j) = k \quad \text{for all } j \text{ and } i \quad (3-16)$$

Inserting this in our formula for the expected utility (3-15) it will give us

$$E[U(C_i)] = k \sum_{j=1}^m P(C_i | S_j) \quad (3-17)$$

where  $m$  is the number of states of nature.

Our estimation problem has now been greatly simplified. We now only have to estimate the conditional probabilities of using the various criteria at all states of nature. In estimating these probabilities we will often gain little by using a detailed description of the

states of nature; thus a reduction in the number of states can take place before the probabilities have been estimated. This will reduce the estimation problem even further.

There are two more justifications for making this simplification. First, it will not be easy to estimate the utility and the probability of the very unlikely events that are explicitly introduced when the number of states is increased (cf. Raiffa, 1968; Schlaifer, 1969). Further, we will probably find that the relative size of the estimation errors increases as the states of nature become more and less distinct. This suggests that in many cases the simplification may provide more than a mere crude approximation.

An alternative way of simplifying the estimations is to say that all targets that we try to destroy are equally valuable. This assumption is motivated by the difficulty in estimating the utility values in any reliable way. The approach bears great similarities to the principle of insufficient reason that is often used when a decision-maker finds it impossible to assign probabilities to states of nature (Fabrycky and Thuesen, 1974).

By this simplification the columns of the utility matrix (Table 3.3) will be made equal. As we had already equalized the rows, the utility matrix will now be an identity matrix, i.e.

$$U(C_i, S_j) = k \quad \text{for all } j \text{ and } i \quad (3-18)$$

Inserting this in our formula for the expected utility (3-15) we now obtain

$$E[U(C_i)] = \sum_{j=1}^m P(C_i | S_j) \cdot P(S_j) \cdot k = P(C_i) \cdot k \quad (3-19)$$

From equation (3-19) we can see that our estimation problem has now been reduced to finding the probabilities of using the criteria or interception techniques. This is the approach that has been used by the Air Staff, as it was considered that the assumptions of this model were fulfilled.

### 3.5.3 Some comments

The essence of this approach is not the finding of the estimates necessary for our multiple-criteria model, but the ability to help the expert to consider various aspects of his problem (cf. Section 1.3). These aspects include not only the utility and probabilities quantified in the model, but also the underlying assumptions about, for example, hitting probabilities, etc. In some cases simplifications can be introduced to give us equation (3-17) or (3-19); the necessary estimates are then very few. In other situations it may be appropriate to use the basic model (3-15) and perhaps even explicitly include aspects such as hitting probabilities etc.

The most restrictive assumption in the application of Tell's method is the one requiring the criteria to be exhaustive and exclusive. Before formulating the problem fully in the next chapter, we can mention already here that this assumption will be fulfilled for goals 1 - 5 for fighter squadrons, but not for the sixth goal.

Another restriction of Tell's method is the many estimates that have to be made if neither of the two simplified versions is used. Further studies will reveal how likely it is that the assumptions of these versions will be fulfilled in other problems, thus showing us how applicable the method is to other non-military problems.

The approach used in Tell's method should be attractive to the decision-maker, as it makes him explicitly discuss and evaluate the states of nature, their likelihood, and the evaluation problem at the various states of nature.

### 3.6 THE CONSULTANTS' METHOD

A group of consultants contracted by the Materiel Administration of the Armed Forces to "develop methods for evaluation of the readiness of the material" (Tre Konsulter AB, 1972, p.3) proposed the fifth of our methods. Their method has only been used to evaluate the readiness of the materials of artillery and anti-aircraft artillery battalions. The consultants argue that their method can also be used for evaluating the capability of combat units.

### 3.6.1 Model and assumptions

The consultants started with an approach somewhat similar to Tell's but, finding it too extensive, they decided to use a simpler method.

The consultants presented the additive, the conjunctive, and the disjunctive models (2-4, 2-6, and 2-8), but found them too restrictive for the problem they were to model. Instead they chose an additive model with interaction terms (2-5).

There was no discussion of independence.

### 3.6.2 Estimation of the compound utility function

#### 3.6.2.1 Verifying the assumptions

The estimation technique developed by the consultants is actually a hybrid method; the unidimensional utility functions are estimated by a direct method and the weights by a special type of indirect technique. For the sake of simplicity we will still refer to it as a direct method.

As the method is not a direct method there are no assumptions to verify.

#### 3.6.2.2 Estimating the unidimensional utility functions

In their papers (Tre Konsulter AB, 1972 a and b; Försvarets Materielverk 1974 a and b) the consultants do not say how the unidimensional utility functions are to be estimated; but they implicitly assume that these functions exist.

In a report on multiple-criteria evaluation models, a committee of the Department of Defense discussed the method developed by the consultants very fully. However, their report does not throw much light on the estimation procedure. An example (PPBG, 1974, p.99) illustrating a two-criteria evaluation problem gives the impression that the utility functions should be linear, but otherwise nothing is said about them.

The consultants claim<sup>1)</sup> that they wrote nothing about the estimation of the utility functions of the criteria; they usually assumed them to be linear, but if there was any doubt they asked the decision-maker for his utility function. The consultants then presented the decision-maker with a value for one criterion and asked him to reveal the utility index of this value. The value was then plotted on a diagram. The question was repeated for several values of the criterion until a smooth curve could be drawn through the points in a diagram.

The consultants do not tell us how to determine the performance values associated with the utility indices 1 and zero. In the evaluation of combat units this is no problem, as these criteria values are given and known.

This is the only one of the four direct methods that asks the decision-maker to reveal his utility index for a value of the criterion. All the other methods ask the decision-maker to indicate a value for the criterion that corresponds to a certain utility index. The reason for the inverse procedure is that the consultants want the decision-maker to answer by utility indices when estimating the parameters of the compound utility function as well as the unidimensional utility functions.

### 3.6.2.3 Estimating the parameters

The method for determining the compound utility function is, as we have mentioned, an indirect technique, as the decision-maker does not give the weights of the model but evaluates some alternatives from which the weights are derived. It differs from the indirect methods presented in Section 2.2.1 as the number of alternatives is the lowest possible for determining the utility function and by not using statistical techniques for computing the parameters.

The alternatives to be evaluated are all the corners of a polygonal spanned by the  $n$  criteria. We will scale the compound utility function

---

1) Telephone-conversation with Mr. Reveman, February 1974.



so that

$$U(*x_1, *x_2, *x_3, \dots, *x_n) = 1 \quad (3-20)$$

and

$$U(*x_1, *x_2, *x_3, \dots, *x_n) = 0 \quad (3-21)$$

where  $*x_i$  and  $*x_i$  are the most and the least preferred values, respectively, of the  $i$ -th criterion. The decision-maker is asked to reveal his preferences, i.e. to give his utility index, for  $(2^n - 2)$  more alternatives. These alternatives consist of all permutations of the criteria when each criterion assumes the values  $*x$  or  $*x$  only. Inserting the utility indices in the compound utility function (2-5) we will get  $n$  equations with  $n$  unknowns. Solving this system of equations we will get the parameters of the model.

Let us illustrate the procedure by a three-criteria object. The compound utility function (2-5) will here look like

$$\begin{aligned} U(x_1, x_2, x_3) = & a_0 + a_1 u_1(x_1) + a_2 u_2(x_2) + a_3 u_3(x_3) + \\ & + a_4 u_1(x_1) u_2(x_2) + a_5 u_1(x_1) u_3(x_3) + \\ & + a_6 u_2(x_2) u_3(x_3) + a_7 u_1(x_1) u_2(x_2) u_3(x_3) \end{aligned} \quad (3-22)$$

We scale the compound utility function so as to fulfil (3-20) and (3-21) and assume that there are unidimensional utility functions. The decision-maker is presented with six alternatives to evaluate. These alternatives are presented as numbers 2 to 7 in Table 3.6. The estimates given by the decision-maker are also shown in this table, as well as the two scaling restrictions (numbers 1 and 8). The column to the far right of the table shows the equations that we get by inserting the values of the criteria and the estimated utility index into formula (3-22). Solving these equations, we get the parameters shown in Table 3.7.

The first parameter,  $a_0$ , will always be zero. The  $n$  next parameters, representing the main effects, will be non-negative. The coefficients of the interaction terms may be positive as well as negative. If all these later coefficients reach zero, the model will be purely additive.

No.	$u_1(x_1)$	$u_2(x_2)$	$u_3(x_3)$	Estimated utilities	Equations to solve the parameters
1	0	0	0		$0.00 = a_0$
2	1	0	0	0.20	$0.20 = a_1$
3	0	1	0	0.20	$0.20 = a_2$
4	0	0	1	0.10	$0.10 = a_3$
5	1	1	0	0.35	$0.35 = a_1 + a_2 + a_4$
6	1	0	1	0.30	$0.30 = a_1 + a_3 + a_5$
7	0	1	1	0.60	$0.60 = a_2 + a_3 + a_6$
8	1	1	1		$1.00 = a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7$

Table 3.6. Example of evaluated alternatives for determining the parameters of the compound utility function.

Coefficient	$a_0$	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$	$a_7$
Value	0.00	0.20	0.20	0.10	-0.05	0.00	0.30	0.25

Table 3.7. Parameters from the example in Table 7.6.

### 3.6.3 Some comments

All the methods make use of extreme values in determining the uni-dimensional utility functions, but only Keeney and the consultants use extreme values in finding the parameters of the compound utility function. This means that we make the same objection to this method as we made in the case of Keeney's. Will the decision-maker be able to make any reliable estimates of objects such as Nos. 2 to 7 in Table 3.6? Could we not present intermediate combinations of performance values? It seems likely that the decision-maker will find it almost as difficult to answer the questions posed by this method, as they

would if Keeney's method were used. We could also expect them to feel little confidence in the precision of this model, but their understanding of the questions asked in the consultants' method would probably be somewhat greater.

The consultants' model requires many estimates. For a problem of  $n$  criteria we will need to determine the value of  $(2^n - 2)$  parameters. To accomplish this the consultants have to ask the decision-maker to evaluate  $(2^n - 2)$  alternatives. It is easy to see that as  $n$  grows the job performed by the decision-maker will grow exponentially. This was one reason why a committee of the Department of Defense, whose task was to examine evaluation models, was doubtful about using the consultants' model when the number of criteria exceeded four (PPBG, 1974).

The consultants' method for estimating the unidimensional utility functions is the only one of our four methods that asks the decision-maker to give utility indices that corresponds to particular values of the criterion. This should be one of the advantages of this method when the criterion can assume only very few values.

### 3.7 HYPOTHESES AND EXPECTED RESULTS

In the various comments following the presentation of the selected methods some hypotheses have been put forward. Some of these will be tested in the course of our experiment, while others will be left to future research into utility methods. In this section I will look at the attributes that will be used in evaluating our multiple-criteria models and our estimation methods. We will also discuss the results that we can expect. The hypotheses and expected results are summarized in Table 3.9.

Two of our four direct methods were developed in the defense sector, while two were chosen from the literature of the subject. The methods can also be classified according to their flexibility. By a flexible method we mean one that can estimate pure additive models as well as additive models with interaction terms, while non-flexible methods are those which can only estimate pure additive models. As

we selected our methods to represent both types of estimation methods, we can classify them as in Table 3.8. We will refer to this classification when we forecast the outcome of our experiment.

		The adaptation of the method to the investigated problem	
		great	small
Flexibility	great	Consultants	Keeney
of the method	small	Tell	Miller

Table 3.8. Classification of the selected methods.

### 3.7.1 Precision of a model

One important aspect to examine - and probably the most common - is the precision or the accuracy of various multiple-criteria models. In this section we will discuss ways of measuring the precision of multiple-criteria models and what result we can expect of our models.

Our real-life problem concerns the determination of the utility functions of the Air Staff for use in evaluating its squadrons. For our experiment we used a group of officers, each of whom estimated this function individually.

The subjects of our experiment were all members of the organization concerned and we assumed, in the spirit of the theory of teams (Marschak and Radner, 1972), that they all contributed to an agreed organizational objective - a strong Air Force.<sup>1)</sup> Thus, differences of opinion will not depend on differences in preferences between the subjects but will be explained by factors such as differences in mood, differences in the aspects or alternatives that come to the mind of the subject when he states his values, and so on (cf. for example Block and Marschak, 1960, and Luce and Suppes, 1965, regarding random utility models and Shepard, 1964, for psychological support for this approach). This does not imply that we will produce only one model. We will determine the

---

1) In Chapter 5 we will examine the validity of this assumption by studying similarities in our subjects' evaluations of identical squadrons.

models individually for each subject and then compare the characteristics of these models by computing averages, variances, and mean square errors over all subjects. This approach shows similarity with the approach used in experimental psychology as well as that used in correlational psychology (Wiggins, 1973), and it also resembles what has been suggested by, for example, Churchman and Ackoff (1954) and Miller (1970) with regard to the determination of the utility function of an organization.

Multiple-criteria models are often validated by comparing the predictions of the model with real data. Meehl (1954) refers to a study by Sarbin, in which some models for predicting academic success were tested. The predictions of the models were compared with actual data. Einhorn (1972) also used real data to compare the ability of some models to forecast the survival time for patients suffering from Hodgkin's disease.

In many cases, however, - for instance in the case of the Air Force problem - it is not possible to measure the dependent variable or the utility index. No real data of the capability of the squadrons are or will be available, and so it becomes necessary to find another way of validating the models.

In Section 1.3, we stated that a model could be either a substitute for or a complement to the evaluator. In both cases it seems natural to consider the model that produces the evaluations that are closest to those made by the evaluator as the best model. Such a criterion has been used by, for instance, Einhorn (1971), Goldberg (1971), and Mertz and Doherty (1974). This means that we need the evaluator to make some intuitive evaluations of some squadrons. We then let the models evaluate the squadrons as well, so that we can see the concordance between the different evaluations. In psychological research this concordance has generally been measured by correlation coefficients (Dawes, 1973), but our approach makes it possible to use other statistical measures as well.

The intuitive evaluations necessary to validate the direct models will also be used to derive multiple-criteria models by the indirect

method. To avoid validating the indirect models on the same data as was used to estimate them, we split the intuitive evaluations into two groups and applied regression analysis to one group in constructing our models. We then determined the precision of these models in the same way as in the case of our direct models, except that we used data from the second group only.

Now that we have discussed how to measure the precision of the models in this study, we can proceed to a forecast of the results.

We can expect the models estimated by the flexible technique to show greater precision than the others, as they can portray a greater variety of utility functions. We will probably also find that the "practical" methods developed for this type of problem are not as good as their "theoretical" counterparts. We expect the precision of the indirect models to be even greater, as other studies have shown that subjects do not use the weights that they say they do (Hoffman, 1960).

### 3.7.2 Other attributes

As we have already pointed out several times, it is seldom satisfactory to select a method or a model on a basis of its precision only. Several other attributes of the methods and models have to be considered as well and the interest in such attributes has increased during recent years (cf. Section 2.3). Aubin and Näslund (1972) and Näslund (1974) have argued for experiments to determine how multiple-criteria methods can be rated according to attributes other than precision. Feinberg (1972), Dyer (1973), and Wallenius (1975) have studied some such attributes in evaluating interactive multiple-criteria methods. Vertinsky and Wong (1975) have done the same with regard to the von Neumann-Morgenstern method (used by Keeney as described in Section 3.3) and the MacCrimmon-Toda method for obtaining unidimensional utility functions.

We will specify here some attributes other than precision which appear to us to be important and which we will examine in our experiment. At the same time we will say something about the results we anticipate.

The users' impression of the methods and models is more difficult to predict than the precision. Before we specify our attributes in greater detail, we can make some general comments on the methods. On one hand, it might be expected that the methods developed in the defense sector would be given higher rankings, as they were developed internally and should be better adapted to the organization and its problem. On the other hand, though, we might expect the flexible methods to be given less favorable rankings, as they will always be more complicated.

We made a distinction between using the methods and using the models, and we will begin by presenting the attributes used to evaluate the methods. The ease of using the methods was assessed according to how easy it was to understand what to do, and how easy it was then to do it. We expected it to be difficult to understand and to answer the lottery questions used in Keeney's approach. Miller's method would probably be easier in both respects, but more complicated to understand than to follow. The questions asked in Tell's method should be very easy to grasp but not so easy to answer, since it might be difficult to estimate the probabilities. The consultants' method would be very easy to understand but more difficult to follow, because of the extreme values assigned to the criteria.

The decision-maker's belief in the precision of the models is a very important aspect. A decision-maker may prefer using a simple model that he understands to a more realistic but also more complicated model. At first we might expect the decision-maker to believe the same as we did above, as none of the models is very complicated. But as the two flexible methods are more difficult to use, we could expect some halo-effects to lower the rankings of these models.

If we distinguish between ease of understanding and ease of following the instructions in the models, then the additive model will be the winner. Keeney's interactive model ought to be easier than the consultants', at least if we are handling a great many criteria.

Finally we will look at the amount of time required by the methods and the models. The time needed to use the model will produce the

same rankings as the ease with which they can be used, because none of the models involves any difficult procedures. Thus we can expect these attributes to overlap considerably. At the use of the methods I expect the consultants' method to be the most time-consuming technique, because of the large number of parameters to estimate. The two additive methods will be the least time-consuming since they are easy to use, while Keeney's method will lie somewhere between these two groups.

In Table 3.9 we present the attributes we have used to evaluate the direct models and estimation methods, and the expected results.

Attribute	Ranking of methods/models <sup>†</sup>			
Precision	C	K	M	T
Believed precision	T	M	C	K
<u>During construction</u>				
Ease of understanding the questions	(CT)		M	K
Ease of answering the questions	(MT)		(CK)	
Time requirement	(MT)		K	C
<u>In use</u>				
Ease of understanding the instructions	(MT)		K	C
Ease of performing the instructions	(MT)		K	C
Time requirement	(MT)		K	C

† The methods/models are ranked with the best to the left. Methods/models in parantheses are expected to be equal. C stands for the consultants', K for Keeney's, M for Miller's, and T for Tell's method/model.

Table 3.9 Expected results of the experiment.



## CHAPTER 4

THE EXPERIMENT

In this chapter I will present the multiple-criteria problem used for the real-life study of some evaluation models and methods and the design of the experiment.

In the previous chapter I mentioned that our problem was taken from the defense sector. Before motivating the selection of this problem - which I will do in Section 4.2 - I will make a brief presentation of the Planning, Programming, and Budgeting System (PPBS) in the Swedish defense sector (Section 4.1). In Sections 4.3 and 4.4 I will give a more detailed presentation of the chosen multiple-criteria problem. Finally I will describe the design of the experiment in Sections 4.5 to 4.7.

4.1 A BRIEF PRESENTATION OF THE PPBS

In 1968 and 1969 two special groups in the Swedish Department of Defense published several reports concerning the design of a new Planning, Programming, and Budgeting System (PPBS) for the Swedish defense sector (SOU 1968:1, 1969:24 and 1969:25). Parts of the suggested system had been tested before publication of the reports and the tests continued up to July 1, 1972, when the PPBS was put into general operation. At that time some parts of the system - mainly the methods for evaluating the production and the status of the combat units - still needed to be tested and analyzed.

The essence of the PPBS is that costs should be related to results, at all levels of the organization and for all types of decisions. Thus the PPBS directs attention away from any detailed breakdown according to input categories (various types of materials or salaries, for instance), and towards program goals. In the PPBS there are only a few appropriations per program and about 70 appropriations altogether for the defense sector as a whole compared with about 300 in the old system.

The change in the system of appropriations is one of the alterations that are necessary to allow for comparisons between output and input. Another important change is the stress placed on negotiations between higher and lower agencies<sup>1)</sup> to achieve a balance in tasks and resources at all levels in the defense sector. These negotiations continue until the two agencies agree that the tasks and resources now balance. The contrasts resulting from the negotiations provide the basis on which budgets are designed in the agencies. These budgets are aggregated to form the program budget for the defense sector which is presented to the Riksdag.

To control the implementation of the program budget and the budgets of the agencies, a cost-accounting system has been introduced to follow up the resources agreed in the contracts and approved in the appropriations. But a system is also needed to check the fulfilment of the contracted tasks. This system for measuring tasks or performance must contain information on:

- (a) the state of the total stock of combat units, i.e. their quantity and quality,
- (b) the state of the production facilities, e.g. barracks, simulators, training areas, and runways, and
- (c) the fulfilment of the annual contracts at all levels, e.g. the quantity and quality of material produced, pilots trained, and officers examined.

The performance-measuring system has many uses - first of all, in budgetary control, where the Air Force uses data from the system to determine the capability of all its combat units.<sup>2)</sup>

---

1) The term agency includes here any military unit that draws up its own budget. Thus wings and regiments are included in the term in the present context.

2) The examples are taken from the first of the three purposes of the system.

Another use is in the planning of mobilization for war. The war-time tasks of every combat unit will be changed (replanned) as their actual capability changes, and data from the performance-measuring system will provide the basis for this planning.

But changes in the capability of a combat unit do not necessarily imply a change in war-time duties. It may be considered easier, cheaper, or in some way better to allocate more resources to a unit, thereby increasing its capability to a level necessary for the fulfilment of its original war-time tasks. This means that information from the performance-measuring system will be used for decision-making concerning, for instance, the next year's production, or the products themselves (i.e. the combat units), or about changes in the production facilities.

Although the information from the performance-measuring system concerns historic events or objects, this data will be used for various kinds of decision-making, e.g. which combat units to improve or which goals should have priority, and the information will be used at different levels in the organization. The highest levels will use the data for decisions such as those mentioned above. Lower levels will use the performance-measures in planning the training of their pilots and officers, in planning which materials to improve or which courses to provide.

In obtaining data for all these three uses of the performance-measuring system, we are faced with multiple-criteria problems. It is impossible, for example, to find any single measurable indicator of the quality of a combat unit, a runway, or the officers examined from a course.

#### 4.2 SELECTION OF PROBLEM

My first concern was the type of problem to use for the experiment. Many studies have used hypothetical problems<sup>1)</sup> but I wanted a real-

---

1) As examples we can mention Thurstone (1931) and MacCrimmon and Toda (1969), both described in Section 2.2.3, Dyer (1973) who had students evaluate hypothetical cars, and Wallenius (1975) who had students and managers solve a management problem for a fictitious company.

life problem for several reasons. One is that the subjects are better motivated by a real-life problem and will be likely to give more consideration to their answers. Another reason is the chance of getting general information about the implementation of the methods examined, and a third advantage is the greater generality of the findings.

In the defense sector I found many problems showing several of these advantages. Thus any of these problems would have yielded at least as reliable results as any problems outside this sector. And it would still be possible to generalize my results to other problems outside the defense sector. I will now mention some of the advantages of the particular problem I chose.

For the purpose of the experiment I decided to select one problem from one of the three uses of the performance-measuring system presented in Section 4.1. I chose the problem of evaluating the capability of combat units, as hardly any work had yet been done on the other two types of evaluations that have to be included in a complete performance-measuring system.

At the time of the experiment a particular multiple-criteria evaluation method (i.e. Tell's method, see Section 3.5) had already been in use for constructing models for four types of squadrons and one type of ground-to-air missile company. These models had also been in use for some years. To be able to use a previously structured problem for the experiment was a great advantage. It meant that when the subjects were evaluating the tested estimation methods and multiple-criteria models, they could concentrate on them without being distracted by the presentation of the problem.

Of the combat units for which multiple-criteria evaluation models existed, I decided to concentrate on two types only, namely fighter and reconnaissance squadrons, as these would yield the most subjects.

The approach used by the Air Force in structuring the problem had another advantage as well. The problem was structured hierarchically, and this gave me the opportunity to evaluate the methods and models of several different problems simultaneously. The complexity of the problems also varied providing another dimension of the methods and models that was interesting to examine. One of the greatest advantages of this defense problem was, however, that it gave me more subjects than I would have been able to find in most other problems. Business firms generally only have four or five people with enough knowledge of any single problem to be included in a test like this one. In the Air Force there are at least a hundred officers who know enough about how a combat unit should be evaluated to make them suitable as potential participants in an experiment of this kind.

Having looked at all the advantages of the selected problem we can now consider why this problem is not as unique to the military sphere as may first appear. The chosen problem concerns the evaluation of the capability of combat units, but it could just as well have concerned the capability of a hospital, a computer department, or the R&D-department of a business firm. The capability figures that our models will produce will be used together with cost estimates etc., as was mentioned in the last section, for various types of decision-making. And these decisions are not specific in any way to the military sector. This problem can therefore be considered as one of very many similar multiple-criteria evaluation problems.

We will now present the chosen evaluation problem in greater detail.

#### 4.3 WAYS OF MEASURING THE CAPABILITY OF COMBAT UNITS

When a new type of combat unit is being established, several goals for this type of unit have to be determined. This is done by studying the military threat to Sweden and the resources available for this type of combat unit. These resources may be monetary or non-monetary, e.g. the possibility of using available materials, buildings, etc.

The goals are fairly well specified. For example, the goal for the

interception capability at VMC for a fighter squadron is described by defining the number of missions of various types that a squadron should be able to perform each day. Having established and specified the goals, the necessary resources for achieving them are then determined. The resources may be the number of various kinds of personnel, the quantity and quality of materials, buildings, fortifications, etc. All this information - the well specified tasks and the resources allotted - is presented in the document TOEM (from the Swedish for tactical-organizational-economic goal).

The evaluation problem arises because the combat unit seldom receives all the resources it needs according to the TOEM. There can be many reasons for this: external reasons such as price increases, recruiting problems, training difficulties due to bad weather, etc., or internal reasons such as the inefficient use of the resources allocated to training. As units generally get less resources than planned, they are also capable of doing less than was planned according to the TOEM. The problem is to determine how far these shortcomings affect the various goals and the total capability of the combat unit.

The traditional way of measuring the capability of a combat unit is by inspections. In the Swedish Air Force inspections are made by the Commander-in-Chief of the Air Force and by the different regional commanders. In the case of the combat units the inspections made by the Commander-in-Chief of the Air Force are neither very frequent - every wing is inspected approximately every third or fourth year - nor very thorough.

The regional commanders' inspections of the squadrons and the air defense control systems are more frequent - about once or twice a year. Their inspections of combat units other than squadrons are less frequent and occur on an average every fourth to sixth year. Their inspections are more detailed than those of the Commander-in-Chief of the Air Force.

The PPBS requires evaluations of the combat units at least once a year, if the main purpose of the PPBS - the balancing of resources and tasks - is to be fulfilled. However, it would have been too costly to increase the inspection-capacity to meet the requirements of the PPBS. Another reason for not extending the inspection system is that the combat units evaluated at the inspections are not identical with the units about which the PPBS requires information.

Thus the Department of Defense decided that multiple-criteria evaluation models should be developed instead. These models were to support the performance-measuring system with data. The quality of the data was to be checked by the inspections. The models were also to be used for budget-simulations.

#### 4.4 STRUCTURING THE EVALUATION PROBLEM

We will describe the problem of evaluating a fighter squadron in greater depth, while the problem of the reconnaissance squadron will be treated very briefly at the end of this section, as these two problems are very similar.

As mentioned above, the TOEM-document contained information on the goals of a fighter squadron. Four of the more important goals were used in the experiment, and these are listed in Table 4.1. The Air Staff is interested in knowing the status of all its squadrons expressed in terms of these goals. This means that one multiple-criteria model has to be made for each goal, and each model uses a set of measurable indicators as inputs.

Interception capability at VMC day
Interception capability at VMC night and/or IMC
Interception capability when countermeasures are used
Other tactical assignments (e.g. fighter sweep, fighter attack)

As indicators the Air Staff decided to use existing tests of the pilots' abilities. Let us call these ITT, i.e. interception-technique tests. One test may be the interception of a fast high-flying aircraft, while another may be directed towards a fast and very low-flying strike aircraft against which only IR-missiles can be used, etc. The tests are standardized so that the problems and situations facing the pilots should be as similar as possible from one occasion to another. The Air Staff requires all pilots to undergo the tests at least once a year. There are also tests for other aspects, e.g. to check instrument flying ability.

Some measurable indicators or tests, however, will concern all four goals, for example tests of the pilots' ability to take-off, land, navigate, etc. The Air Staff found it better to determine first the value of each of the four goals, without taking this group of indicators into consideration. There were many reasons for this: for instance that the pilots generally ranked very high on these common indicators which were basic elements in all flying and the subject of frequent training, and that a separate report on some goals would make it easier to find any weaknesses in them.

The Department of Defense and the Riksdag, however, were not interested in details; they wanted one overall measure describing the capability of each squadron. For this purpose we needed another multiple-criteria model, using the goals of the TOEM as its attributes. This means that our evaluation problem assumes a hierarchical structure, as can be seen in Figure 4.1.

To the right in Figure 4.1,  $x_i$  indicates the measurable indicators or criteria. The six multiple-criteria problems or goals to be evaluated are marked  $U_i(x)$ ,  $i = 1, 2, 3, 4, 5, 6$ .  $x$  is a vector of the measurable criteria used for evaluating this goal. Thus our six evaluation problems for the fighter squadron may be written as in Table 4.2.



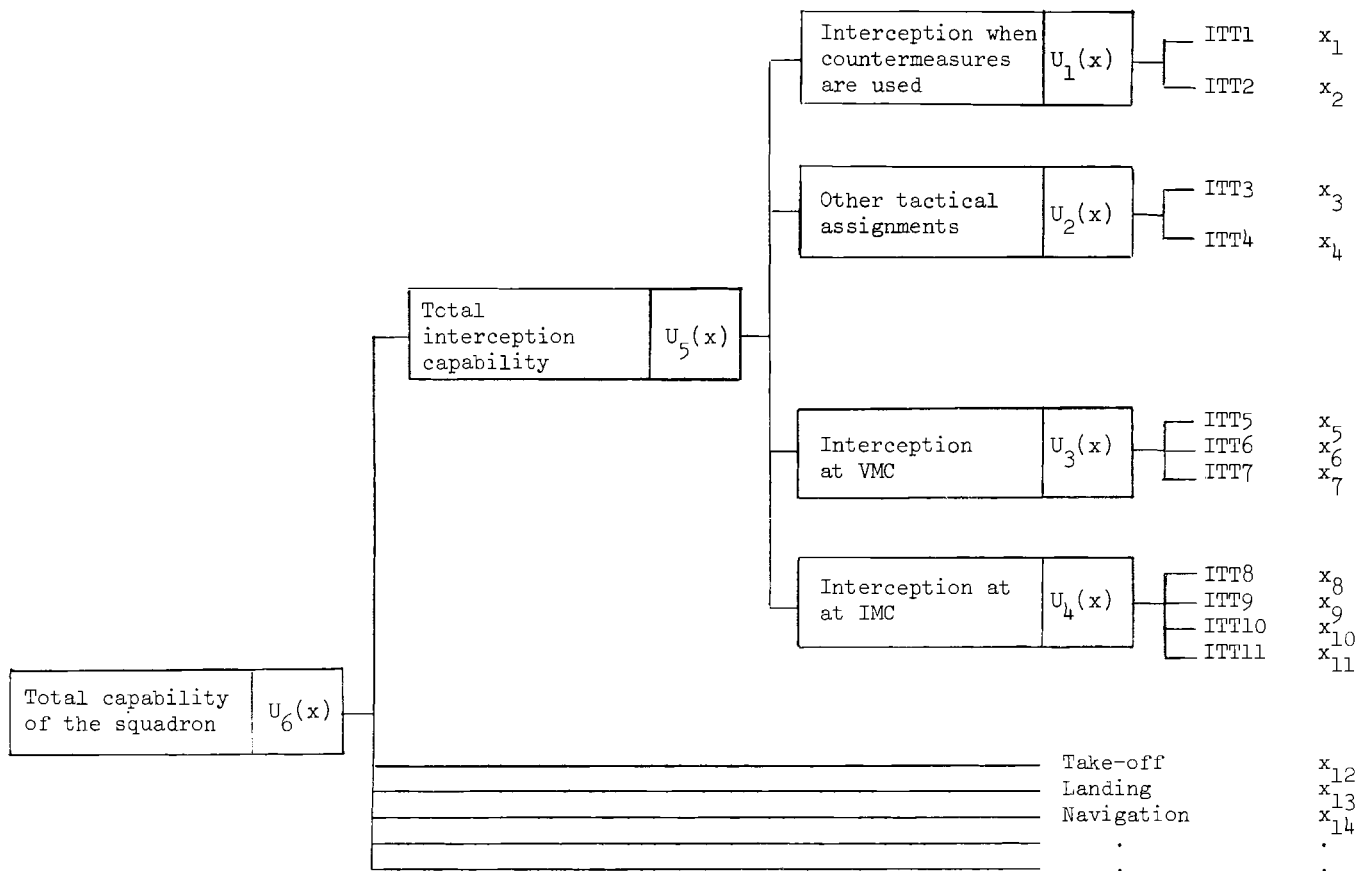


Figure 4.1 Evaluation structure of fighter squadron. Branch-tips to the right are measurable performance criteria.

$$\begin{aligned}
 U_1(x) &= U_1(x_1, x_2) \\
 U_2(x) &= U_2(x_3, x_4) \\
 U_3(x) &= U_3(x_5, x_6, x_7) \\
 U_4(x) &= U_4(x_8, x_9, x_{10}, x_{11}) \\
 U_5(x) &= U_5(U_1(x), U_2(x), U_3(x), U_4(x)) \\
 U_6(x) &= U_6(U_5(x), x_{12}, x_{13}, x_{14}, \dots)
 \end{aligned}$$

Table 4.2 The six multiple-criteria problems of a fighter squadron used in the experiment.

The multiple-criteria evaluation problems are ordered in increasing order of complexity. Complexity is defined here as the number of measurable performance criteria included in the evaluation model.

As we mentioned above, the Air Staff is primarily interested in the utility indices  $U_1(x)$  to  $U_5(x)$ , while the Department of Defense is more interested in  $U_6(x)$ .

The evaluation problem for the reconnaissance squadron was structured very similarly. As there were fewer tests (reconnaissance-technique test, RTT) and goals, the hierarchical structure was simpler. See Figure 4.2.

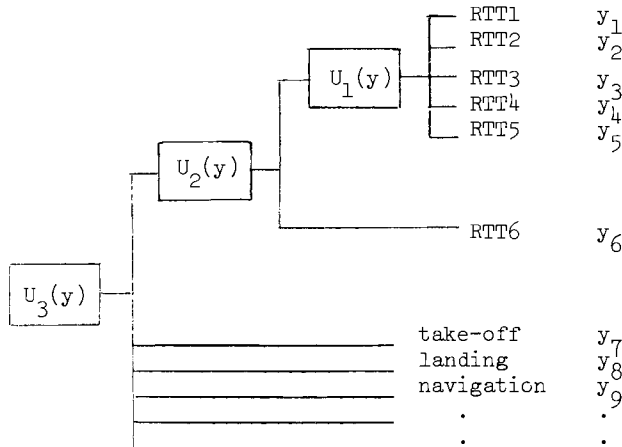


Figure 4.2 Evaluation structure of reconnaissance squadron. Branch-tips to the right are measurable performance criteria.

#### 4.5 PARTICIPANTS

Having decided the real-life problem on which to test my methods and models, my next task concerned the selection of participants. At first I considered using senior officers, as they would presumably have the greatest experience and the best overall view of the goals of the combat units of the Air Force. But the disadvantage would have been that these officers are few and are probably too busy to take part in an experiment. It would also have restricted the experiment solely to interviews, and this I did not want. I then considered the experienced squadron leaders who were presently studying at the Royal Staff College of the Armed Forces. They had all been squadron leaders for several years at least and held the rank of major. As squadron leaders they should be fully aware of the goals of a squadron and of their importance. This meant that in one respect at least they were more experienced than the senior officers, although they lacked something of the latter's overall view.

Other experiments have shown that experience does not greatly affect the way information is used. Hoffman (1968) found no difference between the use made by trainees and the use made by experienced clinicians of information about the patients. And there is less difference in experience between the senior officers and the squadron leaders than there was in Hoffman's experiment.

Thus, as my test subjects I selected students and teachers at the Royal Staff College of the Armed Forces. For the preliminary test conducted a few weeks before the main test, I used some officers from the Air Staff and some teachers and students at the Air Force Squadron Leader School. These students had a few years less experience than their colleagues at the Royal Staff College of the Armed Forces.

The subjects were motivated by knowing that the results of the experiment would be presented to the Air Staff and the Department of

Defense and would be used by them in choosing the evaluation method to be used in the defense sector. They also fully accepted the structuring of the problem, knowing that it had been made by the Air Staff and approved by all squadrons a short time before the experiment.

#### 4.6 DESIGN OF THE STUDY

The study was divided into three separate experiments conducted at about 14-day intervals. In the first experiment the subjects were asked to make intuitive evaluations of certain squadrons. I used this information to derive some models according to the indirect approach. I also took it as the "true" value with which the precision of all models was compared. In the second experiment, to find their parameters and utility functions, the subjects used the four direct methods that I primarily wanted to test. Using this information I constructed the models. The subjects also answered some questionnaires about their preferences for the methods. In the third experiment the participants used the models to evaluate certain squadrons, and they answered various questionnaires about their preferences for the models.

The Royal Staff College of the Armed Forces and the Air Force Squadron Leader School were both kind enough to reserve five hours of scheduled classes for the experiments. I decided to allocate one hour to the first and two hours to each of the other two experiments.

When methods are described in the literature, it is often stated that evaluation models should be constructed in the course of discussions with experts on the multiple-criteria problem. In a comparative study of methods this approach is unsuitable for several reasons. For one thing, there is a considerable risk of the researcher's preferences for certain methods influencing the subjects and, thus, the results. This approach is also much more time-consuming. The interviews alone would have required about a month in the present case.

I chose to present the methods in written form as well as the questions that the methods posed to reveal the subjects' preferences. In this way it was possible to reduce and control the influence of the researcher.<sup>1)</sup> In any case the recommendation in the literature is directed more towards the process of structuring the problem rather than towards the determination of the parameters and functions of the chosen model. Secondly, this change could be expected to affect all methods equally. Thus, conclusions about differences between the methods and their models should not be affected by the chosen design of the experiment.

As was mentioned in Section 4.2, I decided to study both fighter and reconnaissance squadrons. In many situations the number of people involved in constructing models for reconnaissance squadrons is too small to produce any significant results. But as the structures of the two evaluation problems are very similar in many respects, it is often possible to pool the data and thus increase the number of observations.

Many methods require the criteria to be (preferentially and utility) independent. Having chosen the written form for the experiment I found it unsuitable to let the subjects check these assumptions, as it would have been difficult to decide whether they abandoned a method because the assumptions were unfulfilled, because they found the method difficult, or just because they disliked it. Instead I asked officers at the Air Staff verify the assumptions. The omission of this part will only affect the estimates of the time required, which will be slightly underestimated, as the deleted questions were identical with those remaining.

As I pointed out in Section 2.3, I wanted to study the precision of the models and the time used to estimate these models. As the precision as well as the time depends on the number of questions that are asked the decision-maker, I had to fix that number. In order

---

1) As one of the methods had been developed by the author for the Air Staff, it was considered extremely important to control any influences in favor of the author's method. The written form was considered the best solution.

not to favor any method I decided for each method to pose only the minimum number of questions. Thus I excluded any extra questions that might be used to check the consistency in the data given by the decision-maker or to provide more reliable estimates of his preferences. This requirement of an equal amount of input data to all models that are to be compared has also been used by Moore and Baker (1969) in a study of scoring models for R&D project selections.

Each experiment was conducted in a large classroom. I began every session with a short introduction, presenting the purpose of the experiment, the reasons for the study, its implications for the work of the participants, etc. The participants were then asked to work individually with their material until they had finished it, except in the first experiment when a fixed time limit had to be set. Participants were free to ask any questions that occurred to them.

Some details, especially associated with only one of the three experiments, will be found at the beginning of each of Chapters 5, 6, 7, and 8 where the outcomes of the experiments are presented. Further information about the experiments, including the forms, can be found in Tell (1974).

The number of participants varied somewhat from experiment to experiment, because other assignments (such as flying or staff work) had a higher priority. There were 30 participants in the first experiment (24 representing fighter squadrons and 6 representing reconnaissance squadrons), 27 in the second (21 and 6, respectively), and 19 in the third (14 and 5, respectively). The values given in the first experiment will, in some cases, be matched with the values given by the same subjects in the second experiment. 25 subjects (19 and 6, respectively) participated in both these experiments.

As the test material was only slightly changed after the preliminary

test, I have included the results of the preliminary test in part of the analysis.

#### 4.7 PRIOR ACQUAINTANCE WITH THE METHODS

As one of the methods had been developed by the present author for the Air Staff, it was considered to be extremely important that any influences in favor of the author's (or the Air Staff's) method should be checked. The tendency among many subjects to give a higher rating to things with which they are familiar (in this case the Air Staff's model), or with which they are ego-involved (the Air Force), is known as the error of leniency (Guilford, 1954). To diminish and check this tendency I chose, as mentioned above, written forms instead of interviews. I also decided not to reveal the origin of the methods to the participants. In the forms they received, I used only the code-names A, B, C, and D. I also asked the subjects about their prior acquaintance with the methods or the models.

It is very unlikely that any one of the participants had studied the methods by Keeney or Miller. None of the participants from the Royal Staff College of the Armed Forces or from the Air Force Squadron Leader School had used any of the methods developed in the defense sector. Of the two subjects from the Air Staff taking part in the preliminary test, one had read articles about the two tested methods developed in the defense sector. The other subject had used Tell's method, but in quite different a setting. In subsequent interviews, none of them claimed to have recognized the two methods from the defense sector. Once again it must be stressed that nobody who took part in the experiment knew which methods were being tested.

As regards the use of multiple-criteria models, a few participants might have used the Air Force's (i.e. Tell's) model in evaluating their squadrons. Their experience from these evaluations depended on when they left their squadrons to attend the Staff College or the Squadron Leader School. Nobody would have used any of the other models.

To control any leniency effects at the third experiment, I did three things. First, I did not tell the participants which models were to be used. Secondly, I used code-names as in the second experiment. Thirdly, I made sure that a quite different presentation of the evaluation models was made and used completely different forms than the subjects might have seen in the Air Force.

To check their acquaintance with the Air Force's (Tell's) model, I asked the participants to mention any occasion on which they had used the Air Force's model before. I also asked them to compare the Air Force's model with the models used in the experiment.

Several subjects left this question unanswered, probably because they had not used the Air Force's model. (They had nothing with which to compare the tested models.) Of the subjects who did answer the question, the great majority had no experience of the Air Force's model, just as I had expected. Three subjects had had some previous experience, and they all said that the additive model used in the experiment (which is the Air Force's model) was better than the Air Force's model!

Thus the risk that results may have been influenced by any prior knowledge of some method or model can be ignored.



## CHAPTER 5

### THE INTUITIVE EVALUATION

In this chapter we will examine the first of the three experiments. Since we are taking the intuitive utility indices as the values with which the indices of the models should be compared, we will have to check their consistency before going any further with our analysis. The consistency checks will be described in this chapter.

#### 5.1 PRESENTATION OF THE EXPERIMENT

For the intuitive evaluation we needed data on a number of squadrons. During the two years preceding the experiment, all Swedish fighter squadrons had been evaluated twice and all reconnaissance squadrons once by a multiple-criteria evaluation model developed by the Air Staff. This model employed the same criteria that I intended to use in my experiment, plus some further criteria. But I was not able to use this data, as it carried a high security rating. Instead I decided to use it to generate simulated squadrons for the experiment by applying the Monte Carlo technique. (The subjects were told, however, that the data was real.)

Taking the data available regarding the capability of all fighter and reconnaissance squadrons, it was possible to form a distribution function for each criterion. Unfortunately, there was no one-to-one correspondence between the utility index used by the Air Force and the criterion measure. The Air Force utility index<sup>1)</sup> used only five values, 1, 2, 3, 4, and 5, while the criterion, expressing the number of pilots who fulfilled certain requirements, ranged from 0 to 15.<sup>2)</sup> The input values for our experiment were to be expressed not in the AF-utility scale but in criterion measures. The original data

- 
- 1) Henceforth called the AF-utility index, to avoid confusing this scale with others to be used later.
  - 2) The maximum number of pilots varies between the types of squadron, but in this report it will be fixed at 15 (a fictitious value).

(expressed in criterion measures) was not available, so I had to transform the AF-utility scale to the criterion scale by equalizing every AF-utility index with the maximum value of its corresponding interval on the criterion scale. Linear interpolations were used in the intervals (cf. Figure 5.1).

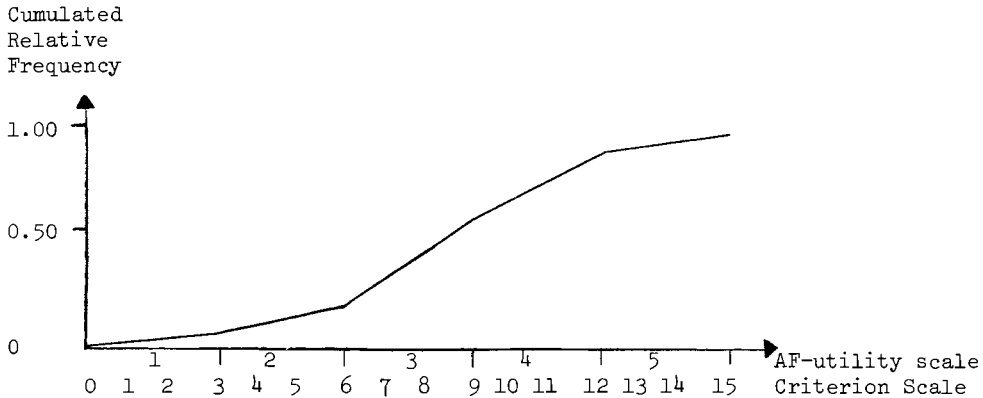


Figure 5.1 Example of a distribution function for one criterion.

In order to construct the squadrons, I had a computer generate random numbers with a rectangular distribution from zero to one, and then I transformed these numbers by the distribution function for every criterion until I had data about all criteria, i.e. I had formed one squadron. I repeated this procedure until I had generated 32 squadrons.

Only one general check on the consistency of the generated data was made. I refrained from making detailed checks on the individual squadrons for two reasons. First, it would have been very difficult to build a simulation model for generating squadrons without at least somebody finding it inconsistent. Secondly, there were some errors in the original data, perhaps due to misinterpretations or a misunderstanding of the instructions for the Air Force evaluation model; thus, as the subjects believed that they were receiving real

historical data and that any errors emanated from this, I saw no reason to make a careful check of the squadrons. I received no comments on the data about the fighter squadrons, and only very few on the reconnaissance squadrons.

The general check on the simulated squadrons consisted of comparing the real historical data and the generated data to see how far they agreed. I computed the distribution functions for all criteria and compared them with the original distribution functions. Using the AF-utility scale I made 64 comparisons for fighter squadrons and 44 for reconnaissance squadrons. Only eight differences for the fighter squadrons and five for the reconnaissance squadrons were greater than 0.10. The maximum difference was 0.25. Thus I concluded that, on an average, the simulated squadrons were not abnormal.

For each squadron the subjects were asked to estimate the capability or the utility of each of the goals. There were six goals for fighter and three for reconnaissance squadrons<sup>1)</sup> (cf. Figures 4.1 and 4.2). An example of the form used is given in Table 5.1.

The subjects were told to give their utility estimate in multiples of 0.05, i.e. as 0, 0.05, 0.10, 0.15, 0.20, etc.<sup>2)</sup> Experiments indicate that subjects rarely use a scale with greater precision than that which I demanded of the participants here (Davidson and Marschak, 1959; Sayeki and Vesper, 1973.) By restricting them to one scale only, I knew that differences between participants would not be due to differences in the length of the intervals used, i.e. that one subject was using multiples of 0.10, perhaps, while another used multiples of 0.01.

- 
- 1) The subjects were in fact asked to estimate the utility of seven and four goals respectively. The extra goals have not been used in the other two experiments or in the analysis except at consistency checks.
  - 2) In fact we used a scale from 0 to 100 in the experiment, thus making the scale more similar to well-known percentages and index figures.

7. FIGHTER SQUADRON J 35 F

Number of crews fulfilling the requirements on

Start _____	15
Landing _____	14
Instrument flying _____	14
Tactical flying at operational performance limits _____	10
Formation flying _____	8
Navigation _____	12
ITT1 _____	11
ITT2 _____	10
ITT3 _____	5
ITT4 _____	15
ITT5 _____	14
ITT6 _____	12
ITT7 _____	9
ITT8 _____	7
ITT9 _____	13
Fighter sweep _____	14

For the following missions consider only the actual fighting, i.e. without paying any attention to the capability of the squadron in starting, landing, navigating, etc.

Fighter defense, VMC day _____	
Fighter defense, VMC night and/or IMC _____	
Fighter defense, jammed environment _____	
Other fighter missions (fighter attack, fighter sweep, etc.) _____	
Fighter missions, total _____	

Finally, consider the total effect of the squadron. At this stage, include the capability in starting, instrument flying, navigating, etc.

Total effect during VMC night and/or IMC _____	
Total effect for all missions _____	

Table 5.1 Form used to describe a squadron to be evaluated in the first experiment. Fictitious data.

In this first experiment I wanted to study the subjects' ability to make the evaluations consistently. For this I used a test-retest procedure which meant that some squadrons would be presented to the participants on two occasions to check whether evaluations would be identical. It is usually recommended that the test and the retest are made on several different occasions to minimize memory effects. However, as I was studying preferences, there was also a risk that if I checked on different occasions the subjects' preferences might have changed between the first and the second test. In any case, for several reasons, the risk of memory effects seemed slight in this experiment.

All the information about the squadrons was expressed in numbers, and quantitative data is more difficult to memorize than qualitative. It is also easier to memorize unlikely and strange combinations of data than to remember the more ordinary values that I was using. Had I taken specially selected alternatives, as Einhorn (1971) does, for example, then the memory effect would have been more significant. It takes time for the subject to translate the given information into utility values.<sup>1)</sup> As at least two other squadrons intervened between the first and the second appearance of any particular squadron, it is highly unlikely that the subject would have recognized the duplicate (which anyway carried a different squadron identification) or that he would have remembered his answers. Finally, I should add that the subjects were told not to look at the squadrons they had already evaluated and I saw no sign of the subjects disobeying this order or even noticing the duplicates.<sup>2)</sup>

The 32 squadrons were divided into two groups with 16 squadrons in each group. Four of the eight duplicates were chosen at random from the first group and four were chosen from the second. Finally, the

- 
- 1) The average time for evaluating all the goals of a fighter squadron was 5.5 minutes and for a reconnaissance squadron 2.7 minutes.
  - 2) At the preliminary test, by mistake, one subject was given a duplicate squadron following directly on the original. His two evaluations did not agree.

order of each group was randomized for every subject, to cancel out possible training and position effects.

The material was split into two parts, as it seemed unlikely that any participant would evaluate all squadrons in the 60 minute period available. As I wanted them to make as many evaluations as possible I purposely gave them too many squadrons. For consistency checks it was necessary that several squadrons should be evaluated twice and that different subjects should have several evaluated squadrons in common. The chances of this happening would increase if the 40 squadrons were divided into two sets.

Squadrons evaluated twice (original and duplicate) were represented by their average for the purpose of the subsequent analysis.

## 5.2 RESULTS

If we want to use the intuitive values as the base with which to compare all the models, we have to check their consistency. I used a two-pronged approach to this question. First I tested the participants' ability to give the same utility index for one squadron at different times, thus examining the consistency of every subject over time. Secondly, I studied the correlation between the utility indices given by different subjects, thus examining the extent to which the participants agreed in their evaluation of the squadrons.

### 5.2.1 Consistency over time

The analysis of intra-subject consistency was based on the duplicated squadrons. Twenty subjects had evaluated at least one squadron twice. Consistency is analyzed in several ways. The data is presented in Tables 5.2 to 5.4.

In Table 5.2 I present four different measures of the agreement between the subjects' evaluations of the duplicated squadrons. Only those subjects who have evaluated at least one squadron twice are included in the table.

Subject No.	Number of squadrons evaluated twice	Average absolute error	Square root of the mean square error	Number of different utility indices with a difference				Relative number of different utility indices with a difference			
				$\geq 0.05$	$\geq 0.10$	$\geq 0.15$	$\geq 0.20$	$\geq 0.05$	$\geq 0.10$	$\geq 0.15$	$\geq 0.20$
Fighter Squadrons											
7	4	0.084	0.110	20	16	8	2	71	58	29	7
8	3	0.050	0.070	13	7	1	0	63	34	5	0
13	1	0.079	0.106	5	4	2	0	72	58	29	0
16	5	0.077	0.109	27	13	6	5	78	38	18	15
21	4	0.095	0.128	23	15	8	3	83	54	29	11
23	5	0.073	0.120	20	13	7	4	58	38	21	22
29	8	0.020	0.040	18	4	1	0	37	8	2	0
30	8	0.038	0.058	28	14	1	0	58	29	2	0
31	8	0.048	0.072	31	16	7	0	65	34	15	0
32	8	0.031	0.094	8	7	5	5	16	14	10	10
33	8	0.029	0.060	20	7	3	2	41	14	6	4
34	8	0.024	0.044	22	3	1	1	46	6	2	2
35	8	0.035	0.056	28	8	2	1	58	17	4	2
36	8	0.029	0.047	25	6	1	0	52	12	2	0
Average		0.051	0.080					57	30	12	5
Reconnaissance Squadrons											
10	2	0.006	0.019	1	0	0	0	13	0	0	0
15	7	0.029	0.043	14	2	0	0	50	7	0	0
17	8	0.055	0.079	20	9	6	0	62	28	19	0
24	3	0.013	0.026	3	0	0	0	25	0	0	0
25	4	0.053	0.076	11	3	3	0	69	19	19	0
28	4	0.000	0.000	0	0	0	0	0	0	0	0
Average		0.026	0.068					37	9	6	0

Table 5.2 Agreement between the subjects' evaluations of the duplicated squadrons. Fighter and reconnaissance squadrons.

We assume here that, on an average, all errors are of the same magnitude and of the same importance to all six goals. This assumption seems reasonable, as we have two effects pulling in different directions: on the one hand it will probably be more difficult to make consistent evaluations of the more complex goals but, on the other, it will probably be easier to make correct judgments of the total effect of a squadron than of the different goals.

I found the average absolute error<sup>1)</sup> for fighter squadrons to be 0.051 and for reconnaissance squadrons 0.026. These errors must be considered very small as the subjects were asked to use multiples of 0.05 only in their evaluations of the combat units. This means that we must accept an error of  $\pm 0.025$  in every evaluation, and a difference between two evaluations of the same object of  $\pm 0.05$ . The mean square error, indicating the dispersion in the estimation errors, is also very low.

Another way of checking the consistency is to count the number of differences of various sizes. This is shown in the columns to the right in Table 5.2. There we can see that "real" errors, i.e. errors greater than  $\pm 0.05$ , occurred in only 30 % and 9 % of the evaluations, respectively. This must be considered a low figure.

One of the most common ways of analyzing agreement uses the correlation between the utility indices given at the first and second evaluation. The averaged correlation coefficients in the present experiment are presented in Tables 5.3 and 5.4, where we see the agreement between the evaluations of each of the six goals of the fighter squadrons and of the three goals of the reconnaissance squadrons. There are fewer observations than in Table 5.2, as some subjects had evaluated too few squadrons.

---

1) The absolute error was used, since both positive and negative differences were regarded as errors. Using the average error would have resulted in these differences compensating each other.



We have used a rank-order coefficient (Spearman) as well as the product-moment coefficient (Pearson) of correlation, but they both produce very similar results. The coefficients are high, as we would expect after the analysis in Table 5.2. The coefficients also tend to get lower as the complexity of the evaluated system increases, although they are still high for the most complex system. The coefficients for the reconnaissance squadron are more uncertain, due to the limited number of observations.

Correlation coefficient	Goal No.					
	1	2	3	4	5	6
Pearson's	0.95	0.91	0.92	0.77	0.75	0.74
Spearman's	0.91	0.86	0.88	0.68	0.67	0.76
No. of observations, P/S	13/13	13/13	13/13	13/13	13/13	12/11

Table 5.3 The average correlation coefficient for the six goals of a fighter squadron.

Correlation coefficient	Goal No.		
	1	2	3
Pearson's	0.86	0.93	0.91
Spearman's	0.83	0.96	0.91
No. of evaluations, P/S	3/3	4/4	5/5

Table 5.4 The average correlation coefficient for the three goals of a reconnaissance squadron.

The conclusion of the analysis of the intra-subject consistency is that our participants can consistently evaluate the capability of squadrons, although the agreement decreases as the system becomes more complex.

### 5.2.2 Consistency between subjects

The concordance between participants provides a very interesting subject for analysis. It has generally been considered impossible to compare the utility functions of individual people but some researchers - e.g. Wallis and Friedman (1942) - have considered it possible provided the background characteristics of the subjects are similar. In our study the subjects had extremely similar backgrounds: they had all been officers and squadron leaders for a long time and had received the same education and training in the Air Force, thus fulfilling Wallis and Friedman's requirement.

There is also another important distinction between classical utility comparisons and the comparisons I have made here. The participants in our study strive for the same goal, i.e. a strong Air Force, and they thus form a team (Marschak, 1955; and Marschak and Radner, 1972). They were asked to estimate the utility function of the team, and the only comparison we make is between their estimates. Thus we do not compare their personal utility functions.

The subjects received more squadrons to evaluate than I expected them to manage in the time available. This circumstance, together with the fact that the order was random, makes a multiple-correlation study impossible without first selecting a subset of either squadrons or participants. Instead I decided to compute the Kendall coefficient of concordance (Conover, 1971) for the thirteen subjects having the greatest number of evaluated fighter squadrons (16) in common. The results are presented in Table 5.5. Increasing the number of subjects and thus decreasing the number of evaluated squadrons in common makes very little difference to the coefficient of concordance.

Kendall's coefficient of concordance	Goal No.					
	1	2	3	4	5	6
	0.82	0.72	0.61	0.59	0.51 <sup>†</sup>	0.39

† Only 12 subjects

Table 5.5 The Kendall coefficient of concordance for 13 subjects who had evaluated 16 fighter squadrons

We note that the coefficients of concordance are high except in the case of the more complex systems, and that they become lower as the complexity grows. The results support our assumption about the participants' ability to express the utility of the Air Force.

Two other ways of analyzing inter-subject concordance, which yield similar results, can be found in Tell (1975).

### 5.2.3 Conclusion

The conclusion of the analysis of the data from the first experiment is that there is a high degree of inter-subject as well as intra-subject agreement in the evaluations. This means that every subject assigns very much the same utility indices to the same squadron on different occasions and that the evaluations of different subjects tend to coincide. In both cases the concordance decreases as the evaluated systems become more complex.

## 5.3 MEASUREMENT SCALES

In the analysis of the data generated by the experiments, the type of measurement scale used will play an important role. It will be important not only in testing hypotheses concerning the participants' attitudes to the methods and models; it will also affect the conclusions we draw about the precision of the models. So far we have been using statistical measures requiring at least ordinal scale data as well as measures demanding, at the least, interval-scale data.

The advantage of the latter type of scale is that it allows us to use more powerful statistical measures and tests, for instance means and tests of differences between means. However, it is impossible to prove that a measurement scale is in fact of a certain type, for instance of the interval type. Thus, by presenting different pieces of evidence, I will try to show that, in the present case, we can

assume the utility scale to be of the interval type without any risk of reaching false conclusions.

Nunnally (1967) opposes the idea of "real" scales. He finds it more appropriate to think of a measurement scale as an agreement among scientists that a particular scaling of an attribute is a "good" one. This means that the use of the scale will be the crucial factor in determining the quality of the scaling. Since neither correlations between scores of individuals on different measures, nor mean differences between differently treated groups<sup>1)</sup>, are much affected by monotonic transformations, the use of a "real" scale or of another scale monotonically related to it, will usually make very little difference to the analysis.

Some statistical studies have indicated that the requirement regarding the use of interval-scale measurements and normally distributed variables when using t-test or F-tests can be violated without affecting the conclusions to any great extent (Abelson and Tukey, 1959, 1963; Boneau, 1960; Baker, Hardyck, and Petrinovich, 1966). This means "that strong statistics such as the t-test are more than adequate to cope with weak measurements" (Baker et al., 1966, p. 303). Other authors, e.g. Siegel (1956), and Stevens (1968), are more critical. McKennel (1970) presents arguments that have been brought forward in discussions in the field of psychology and sociology regarding the use of interval-scale-requiring-measures on ordinal scale data.

If we examine the process of evaluating the capability of squadrons, we will see that both the maximum as well as the minimum capability are unique. The former, resulting in a utility of one, requires all prescribed resources to be available; the latter, giving a utility of zero, demands zero resources. This means, in fact, that the requirements of a ratio scale are fulfilled, i.e. we

---

1) These are both measures which we will use.

have a rational zero and the intervals are expressed as multiples of the maximum-capability squadron.

This utility scale is intuitively attractive and easy to use, since the upper and lower limits of the capability are fixed by definition. It seems very plausible that the procedure which generates data fulfills at least the requirements of producing interval scale data. For example, one 0.70-util squadron should produce the same effect, i.e. the same number of destroyed aircraft, photo-missions, or the same quantity of transported materials, as one 0.30-util and one 0.40-util squadron.

Looking at other studies, we find the assumption about interval scale properties of the utility function to be the dominating one (Meehl, 1954; Dawes, 1971; Einhorn, 1971; Goldberg, 1971; and others).

In some tests it is not only necessary that the subjects use, at the least, interval-scale measurements, it is also necessary that they use the same interval scale. Logically the participants will use the same scale. Because of the nature of the data (cf. the inter-subject consistency check in Section 5.2.2), it is realistic to assume that they do so.

Up to now we have been concerned mainly with the measurement scales used by the participants when they are evaluating squadrons and answering questions about the methods and models used. When it comes to the scales used in the models, the determining factor will be the scale of the utility function of the variables.

The indirect approach that I was testing, i.e. regression analysis, assumes that the utility index given by the decision-maker as well as that given by the model is a measurable criterion, i.e. it contains at the least interval-scale information (Green and Carmone, 1974). The direct methods lead to data on different scales. Keeney, using the von Neumann-Morgenstern approach to finding utility functions,

gets an interval scale (von Neuman and Morgenstern, 1947), Miller (1970) speaks of a ratio scale, while the others say nothing. But as I cannot find any differences between the utility functions generated by the different methods (Section 6.2.2) I conclude that they are all of the interval type. This justifies computing the difference between the intuitive and the model values.

We will assume henceforth that all utility values arise from interval scale measurements, since it seems that the procedures generating the data meet the requirements. Moreover, if this assumption were wrong, the statistical techniques we are employing are very resistant to the type of errors we might introduce. Finally I should mention that the indications of the data are very clear, regardless of the measurement scales and statistical tests used. In some cases I have tried several statistical tests with different assumptions, but they all produce identical tendencies although the levels of significance may vary somewhat. I conclude from this that the results of this study are very invariant to the measurement scale used.

## CHAPTER 6

THE PRECISION OF THE MODELS - DIRECT ESTIMATION

In this chapter I will first present the second experiment, in which the subjects used the direct estimation methods (6.1). The first part of the analysis deals with the methods: their efficiency in estimating the parameters and the unidimensional utility functions (6.2 and 6.3). I have included this analysis of the methods in the present chapter, as any improvement in this respect will also improve the precision of the models. The second part analyzes the precision of the multiple-criteria models (6.4). Other aspects of the models and methods such as the time required, the ease with which they can be used, etc. will be discussed in Chapter 8. Section 6.5 consists of a summary of our findings.

6.1 PRESENTATION OF THE EXPERIMENT

In the second experiment each participant received a booklet divided into four parts - one for each method. The order of the parts was randomized to counteract possible training and position effects. Each method was presented briefly, and followed by the questions necessary for forming the compound utility function, i.e. for finding the unidimensional utility functions of the criteria and the parameters. There were no extra questions for consistency checks, control of independence etc. (cf. Section 4.6). The booklet had been critically examined by several of my colleagues at the Stockholm School of Economics. A preliminary test was carried out and this led to some minor modifications.

6.2 ESTIMATION OF THE PARAMETERS6.2.1 Consistency check of the consultants' method

In the consultants' method it is possible to check for the consistency of the answers given by the subjects without putting any extra questions. This possibility is not discussed by the consultants. The parameters of their model were derived by getting the subjects to evaluate  $2^n - 2$

combinations of utility indices of the  $n$  criteria. For three criteria a form similar to Table 6.1 was presented to the subjects. (The order between the objects is not random in this example.)

Object No.	$u_1(x_1)$	$u_2(x_2)$	$u_3(x_3)$	$U(x_1, x_2, x_3)$
1	0	0	0	0
2	1	0	0	
3	0	1	0	
4	0	0	1	
5	1	1	0	
6	1	0	1	
7	0	1	1	
8	1	1	1	1.0

Table 6.1 Form used in the consultants' method to gather information about the parameters.

As more is preferred to less of all criteria, the compound utility function should increase monotonically. This gives us an opportunity for checking the consistency. In this example object number 5 should have a utility greater than objects 2 and 3 respectively. The utility of object 6 should be greater than that of objects 2 and 4, etc.

As the compound utility is bound to be between zero and one, we have no checking possibilities for two criteria. There are 6 possibilities with three criteria, 36 with four, etc. The number of potential checks grows exponentially as the number of criteria increases.

In this study the utility of each criterion increases monotonically, but as the subjects were asked to give the compound utility indices in multiples of 0.05, we will accept as a consistent answer that, for example, object 5 is equal to or greater than objects 2 and 3 respectively. This means that on account of imperfections in the measurement, we require the compound utility function to be increasing instead of non-decreasing.



The results are presented in Table 6.2, from which we read that the average numbers of inconsistencies are small. They increase as the systems grow more complex and they tend to grow exponentially. We must remember, however, that an inconsistency here is rather a serious error, which means that this technique is less reliable for more complex systems.

	Goal No.								
	Fighter squadron						Reconnaissance squadron		
	1	2	3	4	5	6	1	2	3
Number of criteria	2	2	3	4	4	7	5	2	6
Average number of inconsistencies per subject	-	-	0.05	1.40	1.85	40.50	4.17	-	3.66
Number of subjects	-	-	19	20	20	14	6	-	6

Table 6.2 Inconsistent answers from subjects when using the consultants' method

#### 6.2.2 A comparison of Keeney's, Miller's, and Tell's methods

Keeney's, Miller's, and Tell's methods have an equal number of parameters. The average parameter values cannot be compared in any meaningful way, as they will be used in different compound utility functions. Their variances are comparable, however, since these express the concordance among the subjects about the value of every parameter. The lower the variance, the better the method will be for finding the parameter.

The standard deviations among the subjects for every parameter estimated by the three methods are presented in Table 6.3. In Table 6.4 we can see how many subjects had used the various methods.

Parameters (fighter and reconnaissance)	Standard deviations		
	Keeney's method	Miller's method	Tell's method
1	17.5	12.8	9.3
2	19.8	16.2	14.2
3	18.2	11.6	11.8
4	15.1	9.3	7.6
5	18.9	12.6	10.3
6	13.8	7.5	11.5
7	18.9	14.0	17.4
8	17.0	8.4	11.0
9	18.7	8.4	11.0
10	19.9	21.8	22.6
11	18.9	23.1	25.3
12	21.5	15.1	12.5
13	11.8	9.7	8.0
14	22.2	8.7	10.9
15	13.9	6.2	6.9
16	30.4	6.8	6.6
17	29.2	6.6	5.8
18	27.7	5.5	3.7
19	26.5	5.5	4.6
20	30.3	6.2	3.7
21	29.8	17.7	5.5
22	31.6	15.4	9.8
23	34.2	7.0	5.5
24	5.0	6.0	6.6
25	15.2	20.6	8.2
26	7.4	7.5	5.0
27	8.9	9.5	7.0
28	24.3	9.6	15.4
29	35.1	9.6	15.4
30	7.1	7.6	3.7
31	7.1	7.6	3.7
32	7.9	5.5	4.2
33	6.5	4.6	2.7
34	7.2	5.9	3.7
35	7.2	4.4	1.5
Average	18.4	10.1	8.9

Table 6.3 The standard deviations of the estimates of the parameters - fighter and reconnaissance squadrons. (The order of the parameters is not the same as in Figures 4.1 and 4.2.)

	Keeney's method	Miller's method	Tell's method
Fighter squadron	15, 14 or 10	19, 18, 17, 16, or 15	18, 17, or 16
Reconnaissance squadron	6, 5, or 4	6 or 5	6 or 5

Table 6.4    Number of observations of the parameters.<sup>1)</sup>

We observe that the standard deviation for Keeney's method is much higher than it is for the other two. If we rank the methods according to the variance for every parameter, Keeney's method will receive a sum of its ranks amounting to 94. Comparing this with the maximum rank sum of 105, we see clearly that his method has a higher variance than Miller's and Tell's methods.

If we compare the variances of Miller's and Tell's methods, we find that those of Tell's are lower. The hypothesis of equal variances in the estimates of parameters can be rejected at the 10 percent level, whether we use the sign test or the Wilcoxon sign rank test (Conover, 1971).

The conclusion is that Keeney's method shows very high variances in the estimates of the parameters. Miller's and Tell's methods show significantly lower variances, and of them Tell's method has significantly the lowest.

### 6.3    SHAPE OF THE UTILITY FUNCTIONS

#### 6.3.1    Consistency check of Keeney's method

A printing error gave me a chance to check the consistency of Keeney's method. Two questions (the first and the last) for  $u_1(x_1) = 0.50$  were

---

- 1) In this and other tables the number of observations will be presented in this way. This means that we do not consider it important to show which parameters (in this case) are associated with 18 observations and which are associated with 17 observations. We consider only the magnitude to be important.

included but the  $u_1(x_1) = 0.375$  question was missed out for all criteria. This loss hardly affects the rest of the analysis as very few criteria had any x-value as low as this.

In Table 6.5 we see that a surprisingly large number of subjects found this method difficult, as the number of inconsistencies is large. In almost 40 percent of the utility functions the subjects gave criteria values that differed by more than 10 percent. This must be considered a very high figure.

Difference in x-values expressed as percent of the maximum x-value (note that the x-value is an integer between 0 and appr. 15)	0	0.01-0.10	0.11-0.20	0.21-0.30	0.31-0.40	> 0.40
Number of differences in percent of total number of estimated utility functions by Keeney's method <sup>†</sup>	35.8	25.4	30.3	3.3	5.4	0

† Number of utility functions: 330. Number of subjects: 23.

Table 6.5 Relative number of inconsistent answers, using Keeney's method. Fighter and reconnaissance squadrons.

### 6.3.2 Comparing the utility functions of the four methods

All methods estimate unidimensional utility functions of the criteria. Thus the experiment has given us four utility functions for every criteria and subject.

Miller's utility function fulfills the requirements of a ratio scale. Von Neumann and Morgenstern (1947) have shown that their lottery technique satisfies the requirements for producing an interval scale, and this technique is used in Keeney's method. The other two methods postulate nothing about their scales.<sup>1)</sup> If we cannot find any difference

---

1) As Miller's and Tell's methods for estimating the unidimensional utility functions are almost identical, we will conclude that Tell's method also generates utility functions with ratio scale properties.

between the functions, we can say that the simplest is to be preferred, as the methods generate utility functions of equal quality (at least interval scale). We will examine our hypothesis of equal utility functions by comparing all four of them. We will make a pairwise comparison between the two methods with the weakest scales (the consultants' and Tell's) and the one using a ratio scale. Finally we will simultaneously compare all four methods.

To examine our hypothesis that there are no differences between the utility functions generated by the different methods, we calculated the biggest difference between the utility functions that we had compared. This gave us one difference for every pair/group of utility functions determined by every subject and every criteria. As we are not interested in differences for single subjects or criteria, we will regard the data as a sample from a universe of utility functions determined by these methods.

Maximum difference in utility	Cumulative relative frequencies between methods		
	The consultants' and Miller's	Tell's and Miller's	All four methods
0.00	8.3	9.5	5.5
0.05	16.0	18.6	11.7
0.10	34.4	44.4	16.3
0.15	52.1	71.0	35.8
0.20	79.4	84.4	53.8
0.25	82.4	89.1	63.5
0.30	91.7	92.7	71.7
0.35	95.3	93.5	79.1
0.40	96.5	97.5	87.4
0.45	97.6	97.5	90.3
0.50	100.0	100.0	96.8
0.55			98.5
0.60			99.0
0.65			100.0
0.70			
0.75			
Number of utility functions	169	275	307

Table 6.6 Cumulative relative frequencies of maximum differences between utility functions estimated by various methods over all subjects and criteria. Fighter squadrons.

In Table 6.6 we can see the cumulative relative frequencies of the maximum differences computed over all criteria and all subjects. We can see that there are small differences between Miller's, the consultants', and Tell's methods. In less than ten percent of all utility functions, the maximum difference between the functions exceed 0.30 and the median difference will be less than 0.15. Naturally the simultaneous test yields higher values but the median is still low, and we can conclude that the methods, at least in this experiment, seem to produce identical utility functions.

#### 6.4 THE PRECISION OF THE MODELS

My criterion of a good model was that it should yield the utility indices closest to those given by the evaluator. Thus I calculated the difference between the utility index given by the subject at the intuitive evaluation and the utility index generated by each of the four models. These differences were calculated for each subject, for all the squadrons he had evaluated and for all the goals. The differences were presented as in Figure 6.1<sup>1)</sup>, and they formed the basis for the rest of the analysis.

If we are to be able to compare the utility indices generated in the first experiment with the indices produced by the models constructed at the second experiment, the subjects' preferences have to be stable. This assumption will probably be fulfilled as any military and technical changes likely to affect the officers' preferences would be negligible over a short period of time. There is also evidence from other problems (Grayson, 1960; Goldberg, 1968) that utility functions stay reasonably stable over fairly long periods and here the period was only two weeks.

To analyze the precision I followed two different approaches. The usual way is to calculate the correlation coefficient between the

---

1) From now on the models will often be designated C (= consultants), K (= Keeney), M (= Miller), and T (= Tell). The intuitive evaluation will be marked by I.

Subject No. 3							
Subject No. 2							
Subject No. 1							
Squadron No.	Goal No. 1 (C-I)      (K-I)      (M-I)      (T-I)				Goal No. 2 (C-I)      (K-I)      - - - -		
4	.10	.16	- .10	.14	.33	.22	- - - -
6	- .10	.13	- .13	.03	.37	.08	- - - -
7	- .02	.08	- .12	.02	.17	.19	- - - -
8	- .11	.09	- .15	- .11	.31	.17	- - - -
Average	0.063	0.080	-0.091	0.048	0.175	0.149	- - - -
Variance	0.00363	0.00713	0.00312	0.00117	0.00951	0.01375	- - -
Mean square error	0.00759	0.01357	0.01132	0.00348	0.04010	0.03593	- - -

Figure 6.1 Form used for analyzing the precision of the models.

intuitive utility indices and the utility indices of the model, and to consider the model receiving the highest correlation coefficients as the best (Dawes, 1973).

But choosing a model is in fact similar to finding a good estimator in statistics. In the latter case, criteria such as bias (i.e. the amount of systematic error), variance, and mean square error are common measures for expressing the precision of an estimator; generally the estimator with the lowest mean square error is considered the best. Here we lack the true value used to calculate the variance and the mean square error but, as was indicated in Section 3.7.1, we will use the decision-maker's intuitive value instead. Ackoff (1962) has argued in favor of using these criteria when evaluating models.

One reason for using two methods to analyze the precision of the models is the insensitivity of the correlation coefficient to systematic errors in the models. This is revealed by the "statistical" approach. The correlation coefficient has also been criticized for assigning higher values to incorrect models than to the correct model (Birnbaum, 1973 and 1974).

#### 6.4.1 Correlations

The average product-moment correlation coefficients are presented in Table 6.7. The correlations are calculated on real data from the first and the second experiments. Rounding off the indices of the models to multiples of 0.05, i.e. to the same precision as the intuitive values, had no noticeable effect on the correlations. In the computations I have included all subjects who had evaluated at least three squadrons.

Model	Goal No.					
	1	2	3	4	5	6
Consultants	0.67	0.77	0.64	0.63	0.49	0.25
Keeney	0.73	0.67	0.57	0.62	0.39	-
-----						
Miller	0.72	0.75	0.72	0.54	0.41	0.39
Tell	0.72	0.75	0.69	0.67	0.40	0.23

Table 6.7 Average product-moment correlation coefficients for the four models. Fighter squadron.

One of our hypotheses concerned the difference in precision between the a priori additive models and the flexible models. The latter models may become either additive or additive with interaction terms, depending on the decision-maker's preferences. To simplify the conclusions about this hypothesis I have separated the two types of



models by a dotted line. The flexible models will be found above the line and the additive below it.

Another hypothesis concerned differences between "theoretically" and "practically" developed methods. We will find the two "theoretical" models close to and on either side of the dotted line, while the two "practical" models will be found as the first and the last model in the tables.

In cases where, for some reason, there were too few observations to yield any result, I have inserted a dash.

We see that the average correlation coefficients are fairly high; approximately 0.70 for the first three goals, 0.60 for the fourth, and lower for goals 5 and 6. For more complex goals the coefficients are smaller. All models show very similar correlation coefficients, except perhaps Keeney's where they are somewhat lower.

This method of analyzing the coefficients by averages has been used by, for example, Huber, Sahney, and Ford (1969). As the differences between the correlation coefficients are neither very large nor pointing in any specific direction, and as they are lower in the case of more complex problems, I decided to analyze the correlations by some other methods, to see whether these could shed any more light on the precision of the different models.

Einhorn (1971) and Goldberg (1971) used the maximum correlation coefficient to determine the best model for every subject. Then they used the number of times a model was considered the best as the criterion of a good model. If we apply this to all the subjects who had constructed all four models, we get the results of Table 6.8.

Keeney's model is now showing good results, while the others seem equally good but on a lower level. Most of the success of Keeney's model is acquired at the first goal.

Model	Goal No.						Total
	1	2	3	4	5	6	
Consultants	1	2.5	0	2	2	-	7.5
Keeney	5	2	2	2	1	-	12
-----							
Miller	3	1	3	5	1	-	8.5
Tell	0	1.5	4	3.5	0	-	8

Table 6.8 Number of times a model has had the highest product-moment correlation coefficient for the participants who had constructed all four models. Fighter squadron.

Using the sum of the ranks as a criterion instead of the sum of the times the models has the rank of one, gives us the results shown in Table 6.9. Keeney's model is now found to be as good as the consultants' and Tell's models. They all proved better than Miller's model, which had a slightly higher sum of the ranks.

Model	Goal No.						Total
	1	2	3	4	5	6	
Consultants	24.5	13	25	21	7	-	90.5
Keeney	18	17.5	24.5	18	12	-	90
-----							
Miller	22	18.5	19.5	26.5	11	-	97.5
Tell	25.5	21	21	14.5	10	-	92

Table 6.9 Sums of the ranks of the product-moment correlation coefficients of the four models. Fighter squadron.

The conclusion of this analysis of the correlation coefficients is that we can find no differences in precision between the models. The correlation coefficients are fairly high, but they decrease as the number of criteria increases.

#### 6.4.2 Bias

Assuming that all subjects used equal interval scales (cf. Section 5.3), I calculated the average bias over all squadrons and subjects. The results are presented in Table 6.10.

Model	Goal No.					
	1	2	3	4	5	6
Consultants	0.035	0.057	0.039	0.052	0.056	-0.061
Keeney	0.061	0.063	0.114	0.174	0.177	0.405
-----						
Miller	0.013	0.037	0.015	0.074	0.034	0.076
Tell	0.029	0.036	0.012	0.047	0.040	0.105

Table 6.10     Average bias of the models for different goals.  
Fighter squadron.

The average bias is positive, except for one element, indicating too optimistic an evaluation by the models. We find that the simple additive models (Miller and Tell) both do better than the consultants' and Keeney's flexible models. Keeney's model shows much greater bias for the more complex goals (3 to 6), where almost all models contained interaction terms, than for the simple goals 1 and 2, where a majority of the models were purely additive.

It is also important to note that the bias is small. As the values given at the intuitive evaluation were in multiples of 0.05, we will have to consider an absolute bias less than 0.025 as no bias. From Table 6.10 we can see that in 3 of 24 cases we had a bias as small as this. In 9 of 24 cases the bias was absolutely smaller than 0.05 and in only 6 cases was it greater than 0.1. The bias seems to be independent of the complexity of the evaluated system. At goal 6 the bias is noticeably higher than at the other goals, but this may be due to the very small number of observations.

The average standard deviations indicate the inter-subject differences in bias are small. The differences between the models are also small, except for goal 6 (Table 6.11). As the standard deviation of Keeney's model is no greater than that of the other models, Keeney's model must have a bias constantly higher than the others.

Model	Goal No.					
	1	2	3	4	5	6
Consultants	0.043	0.033	0.040	0.028	0.031	0.081
Keeney	0.047	0.059	0.029	0.033	0.037	0.542
-----						
Miller	0.054	0.027	0.053	0.042	0.037	0.061
Tell	0.042	0.033	0.046	0.037	0.027	0.062

Table 6.11 The average standard deviation of the bias of the models for different goals. Fighter squadron.

The conclusion is that the average bias is negligible. Keeney's model, however, showed a bias much higher than that of the other models.

#### 6.4.3 Variance

To find a measure of the dispersion, I calculated the variance of the four models for every subject and goal (cf. Figure 6.1). To find out whether there were any differences in the variances generated by the models, I used the Wilcoxon signed rank test (Conover, 1971). I then had to assume that the subjects used identical interval scales.<sup>1)</sup>

I tested the hypothesis for all pairs of models and for all goals. The results are presented in Table 6.12, where I have used one matrix for every goal. The result of the tests, i.e. the level of significance or the character N for a non-significant difference, is noted

---

1) Identical conclusions can in fact be drawn by using the sign test. Then we only require that the subjects use interval scales but not that they have to be identical (cf. Section 3.5).

in the row with the largest signed sum of the ranks. Reading the rows of the matrices, we can easily see which models are good (empty rows) or poor (a figure, or the character N), and by reading the columns we learn how good the models are (from the levels of significance). So we see, for example, that at goal 3 the consultants' model was better (had a lower sum of the ranks which implies lower variances) than Miller's model although the difference was not significant. We also see that the consultants' model was significantly outperformed by Keeney's and Tell's models.

	C	K	M	T
Consultants			N	
Keeney	N		N	N
Miller				
Tell	N		.10	

Goal No. 1

	C	K	M	T
Consultants		N		
Keeney				
Miller	.05	.25		
Tell	.25	.20	N	

Goal No. 2

	C	K	M	T
Consultants		.20		.20
Keeney				
Miller	N	.20		N
Tell		N		

Goal No. 3

	C	K	M	T
Consultants		N		N
Keeney				
Miller	N	.25		
Tell		N	N	

Goal No. 4

	C	K	M	T
Consultants		.25		
Keeney			N	N
Miller	.10			
Tell	.15		N	

Goal No. 5

	C	K	M	T
Consultants		N		
Keeney	N		N	N
Miller	.25			N
Tell	N	N		

Goal No. 6

Table 6.12 Pairwise tests of differences in variance between the models. Fighter squadron.

The flexible models above the dotted line show more empty rows, and thus smaller variances, than the additive models below this line. Sixteen comparisons, of which nine were significant, indicated a lower variance for the flexible models. There seems to be no difference between the "practically" and the "theoretically" developed models.

Apart from this, the material seems too weak to permit any further conclusions. There are no other evident tendencies, nor any highly significant differences.

#### 6.4.4 Mean square error

It is intuitively clear that a good model has a small variance. It is perhaps not as evident that a large bias is also a problem. But as the size and the sign of the bias are unknown and will vary from problem to problem, we will not be able to correct for the bias by adding a constant. Thus we prefer a model with a small bias and a small variance.

In statistics it is common to use the mean square error (MSE) as the criterion of a good estimator. The MSE is the mean of the sum of the squared deviations from the intuitive values and it implies that the decision-maker has a quadratic utility function around this value. The MSE can also be written as a sum of two of our criteria, i.e. as

$$\text{MSE} = \text{Variance} + \text{Bias}^2$$

A measure similar to the MSE has been used by Sayeki and Vesper (1973) in a multiple-criteria study.

The MSE was computed for every model, subject, and goal (cf. Figure 6.1). Our hypothesis that there were no differences in MSE between the models was subjected to a pairwise test, using the Wilcoxon signed rank test (Conover, 1971). The assumptions etc. were the same as those described in Section 6.4.3. The results can be seen in Table 6.13.

	C	K	M	T		C	K	M	T		C	K	M	T
Consultants								N	N				.15	
Keeney	.25		N	.20		N		N	N		.20		.05	.15
Miller	N								N					N
Tell	.05		N								N			
	Goal No. 1					Goal No. 2					Goal No. 3			

	C	K	M	T		C	K	M	T		C	K	M	T
Consultants				N				.20	N					N
Keeney	.01		.01	.05		.005		.005	.005		N		N	N
Miller	N			.20					N		N			.10
Tell														
	Goal No. 4					Goal No. 5					Goal No. 6			

Table 6.13 Pairwise tests of differences in MSE between the models. Fighter squadron.

The number of significant differences has now increased. Keeney's model did worse than all the other models; eleven out of eighteen differences were significant. These poor results can be attributed mainly to the large bias in Keeney's model. Tell's model did well. It was better than Miller's model in five out of six cases, better than the consultants' model in four out of six cases, and better than Keeney's model. In seven of these comparisons the difference was significant. The consultants' and Miller's models appeared to be equally good.

The superiority of the flexible models has now vanished. Instead it seems that the additive models are better than the flexible. Nineteen

out of twenty-four comparisons of these models favored the additive model, and in nine of these nineteen comparisons a significant difference was found. But it also seems that the "practical" models have a smaller MSE than the "theoretical" models. Twenty out of twenty-four comparisons indicated this, and of these as many as eleven were found to be significant.

#### 6.4.5 The precision of the models and the number of criteria

Einhorn (1971) put forward the hypothesis that the precision of models would decrease with an increase in the number of criteria. As Goldberg (1971) pointed out, it is not possible to test this hypothesis without examining how consistently subjects can evaluate multiple-criteria objects when the number of criteria grows.

To test Einhorn's hypothesis, I divided the correlation coefficients expressing the precision of the models (Table 6.7) by the correlation coefficients (Pearson's) showing the subject's ability to evaluate the squadrons consistently (Table 5.3). The results can be found in Table 6.14. We can see that the normalized correlation coefficients are around the level 0.75 for all the goals 1 to 4. These multiple-criteria problems were described by two to four criteria. For the more complex goals, 5 (with ten criteria) and 6 (with sixteen criteria), the normalized correlation coefficients tend to fall. Thus Einhorn's hypothesis seems to hold when the number of criteria exceeds approximately eight.

Model	Goal No.					
	1	2	3	4	5	6
Consultants	.71	.85	.70	.82	.65	.34
Keeney	.77	.74	.62	.81	.52	-
-----						
Miller	.76	.82	.78	.70	.55	.53
Tell	.76	.82	.75	.88	.53	.31

Table 6.14 Average product-moment correlation coefficients normalized with respect to correlation coefficients from the test-retest. Fighter squadron.



## 6.5 CONCLUSIONS

We have analyzed the correlations between the utility indices produced by the subjects during the first experiment and those generated by the models. We have also checked the bias, the variance, and the mean square error of the models.

We have found the correlation coefficients to be rather high, but no model showed any higher correlations than the others. The size of the correlation coefficients diminished as the number of criteria increased.

The bias was negligible for all models except Keeney's. The variance of the flexible models seemed slightly lower than that of the additive models. The mean square error indicated, however, that the additive models were better than the flexible. Keeney's model in particular showed poor results. Tell's model seemed to be the best, and there was no difference between the consultants' and Miller's models.

The low precision of Keeney's model may perhaps depend on weaknesses in the estimating method. The analysis revealed that this model showed higher variances in the estimates of the parameters than both Tell's and Miller's methods. Also, the estimating procedure used by Keeney's method to find the unidimensional utility functions showed many inconsistencies.

The experiment also verified a hypothesis put forward by Einhorn (1971) that the predictive power of models would decrease as the number of criteria increased.



## CHAPTER 7

THE PRECISION OF THE MODELS - INDIRECT ESTIMATION

In this chapter we will analyze the precision of the six models presented in Section 3.2 when they are estimated with the help of an indirect technique. We will also compare these results with those obtained in the previous chapter, in order to learn something about the two estimation techniques. For this reason the present analysis will follow the pattern laid down in Chapter 6, to facilitate our comparisons.

7.1 PRESENTATION OF THE EXPERIMENT

The experiments generating the data analyzed in this chapter were presented in Sections 4.5 to 4.7, 5.1, and 6.1.

Each subject's data from the first experiment was divided into two subsamples of equal size and the standard double cross-validation procedure was employed. Thus the first subsample was used to estimate the models by the regression technique, while the second was used to validate the results. All the results - correlation coefficients, variances, mean square errors, etc. - are therefore computed on the second subsample only. Thus, we expect these measures of the indirect models to be more uncertain than those of the direct models, as the results of the latter are calculated on a sample that is twice as large.

I decided not to estimate any models for goals 5 and 6 of the fighter squadron problem (cf. Figure 4.1), because so many observations would have been needed to achieve any degree of accuracy, that the remaining observations would have been inadequate for the validation. Comparisons with the direct models would also have been impossible, as different hierarchical structures would have been used when estimating these models.

The linear model with interaction terms includes several more terms than the others. According to the same arguments that led to the exclusion of goals 5 and 6 in all the indirect models, we can exclude goals 3 and 4 here. Another reason for not estimating the linear model with interaction terms is that the requirements put forward in Section 3.2 make it almost impossible, in anything but an ad hoc way, to make the regression coefficients of the non-interaction terms assume values between zero and one.

In Section 3.2 we specified some model requirements that had to be obeyed. If these are fulfilled, then at least the variance and the mean square error will be greater and the correlation coefficient lower than would otherwise have been the case. This should be remembered in any analysis of the outcomes.

It may be interesting to see how many regression coefficients had to be set at zero according to our theory, and how the number varied between the models. From Table 7.1 we can see that the disjunctive model caused the most trouble. The linear models show surprisingly high figures, while the conjunctive, logarithmic, and exponential models show the lowest.

Model	Goal No.			
	1	2	3	4
Linear	1	2	5	18
Conjunctive	1	1	3	11
Disjunctive	8	8	8	23
Logarithmic	1	0	4	9
Exponential	1	0	5	11
Linear with interaction terms	0	3 <sup>†</sup>	-	-

† To this we will have to add two regression coefficients that had to be set at one.

Table 7.1    Number of regression coefficients set at zero for the six models. Fighter squadron.

Finally it should be mentioned that no time had been fixed for how long the participants would spend on estimating the models. Using the double cross-validation technique, we can set the time required by the indirect method at 30 minutes. This length of time cannot be compared with that of the direct methods. In Section 8.2.2 we will analyze the time requirements of the direct methods.

## 7.2 ANALYZING THE LINEAR MODELS

On an average the two linear models - one with and one without interaction terms - performed equally well. Other studies (see Section 2.1.2) have shown that the interaction terms account for very little of the variance and that very few of these terms are statistically significant.

An analysis of the regression coefficients reveals that only twelve models out of thirty-five (or about 30 percent) contained one significant<sup>1)</sup> interaction term. Eight subjects out of eighteen had at least one model with one significant interaction term, and only four subjects obtained a significant interaction term for both their models. Of the twelve models with a significant interaction term, only seven showed a correlation coefficient that was higher than that of the corresponding pure linear model.

In examining the precision of the models, I included all those with interaction terms irrespective of the level of significance of the terms.

## 7.3 CORRELATION

The average product-moment correlation coefficients for our six models are presented in Table 7.2. The correlations are calculated for every subject on the indices given by him in the first experiment and on the indices produced by each of the models. The coefficients are then averaged over all subjects. The correlations are calculated exclusively on the part of the data that was not used for estimating the models.

---

1) Significant at the 10 percent level.

Model	Goal No.			
	1	2	3	4
Linear	.76	.75	.56	.57
Conjunctive	.70	.68	.56	.57
Disjunctive	.55	.67	.42	.59
Logarithmic	.71	.78	.59	.53
Exponential	.72	.78	.54	.60
Linear with interaction terms	.75	.74	-	-

Table 7.2 Average product-moment correlation coefficients for the six models. Fighter squadron.

We can see that the average correlation coefficients are above 0.70 for most models in the case of the first two goals, and somewhere between 0.50 and 0.60 for goals 3 and 4. Thus we conclude, as we did in the case of the direct models, that the correlation coefficients become smaller as the number of criteria increases. It is interesting to note that the difference between the models tends to vanish as the number of criteria increases.

If we look at the various models, it is immediately obvious that the disjunctive model scores worse than all the others. The two linear models and the exponential model seem to have the highest correlations, followed by the logarithmic and conjunctive.

Comparing the correlation coefficients of these models with those of the direct models (Table 6.7), we find that all the direct models - except Keeney's - show higher coefficients on an average than the best indirect model. For the first goal the direct models perform less well than the best indirect model, but for the more complex goals the difference is striking. Keeney's model produces correlation coefficients well above those of the disjunctive model (the worst of the indirect models), and in parity with those of the conjunctive and logarithmic models.

The number of times that the different models achieve the highest product-moment correlation coefficient can be seen in Table 7.3. We read that, according to this criterion, the linear and exponential models would be preferred. The conjunctive and disjunctive models do not perform well. The linear model with interaction terms surprisingly rarely shows the highest correlation coefficient.

Model	Goal No.				Total
	1	2	3	4	
Linear	5	3	5	5	18
Conjunctive	1	1	1	2	5
Disjunctive	2	2	2	4	10
Logarithmic	5	2	6	2	15
Exponential	3	8	4	3	18
Linear with interaction terms	1	2	-	-	-

Table 7.3 Number of times a model has had the highest product-moment correlation coefficient. Fighter squadron.

Since the correlation coefficients of the models - except in the case of the disjunctive model - lie very close together, the sum of the ranks seems to provide a better criterion than the number of times a model has the highest correlation coefficients. These sums are shown in Table 7.4. We must remember that a good model is characterized here by low scores.

The linear model now receives the lowest score, followed by the exponential and logarithmic models. The conjunctive, and even more the disjunctive, models show very high rank sums. The linear model with interaction terms appears to be as good as the linear models for goals 1 and 2, but they are both surpassed by the exponential model. These results agree with the conclusions we came to on a basis of Table 7.2.

Model	Goal No.			
	1	2	3	4
Linear	46	60	45	39
Conjunctive	72	65.5	51	45
Disjunctive	83	90	73	49
Logarithmic	50	66	45	53
Exponential	49	44.5	56	54
Linear with interaction terms	57	52	-	-

Table 7.4 Sums of the ranks of the product-moment correlation coefficients of the six models. Fighter squadrons.

#### 7.4 BIAS

The average bias calculated over all squadrons and subjects can be found in Table 7.5. Here we must assume that all subjects use at least equal interval scales (cf. Section 5.3).

Model	Goal No.			
	1	2	3	4
Linear	0.034	0.023	0.030	0.009
Conjunctive	-0.009	0.013	0.014	0.009
Disjunctive	-0.032	0.017	0.012	-0.017
Logarithmic	0.021	0.002	0.017	0.021
Exponential	-0.044	-0.022	-0.022	-0.038
Linear with interaction terms	0.000	0.010	-	-

Table 7.5 Average bias of the models for different goals. Fighter squadron.

The bias in the linear and the logarithmic models is positive, while in the exponential model it is negative. The conjunctive and disjunctive models show a relatively small bias, sometimes



positive and sometimes negative. The linear model with interaction terms seems to have the least bias of all the models. The bias here appears to be less than in the models estimated by the direct techniques (cf. Table 6.10).

There are no tendencies towards an increase or decrease in bias as the complexity of the problem grows.

Model	Goal No.			
	1	2	3	4
Linear	0.083	0.037	0.057	0.051
Conjunctive	0.047	0.078	0.039	0.053
Disjunctive	0.061	0.063	0.045	0.067
Logarithmic	0.080	0.071	0.042	0.037
Exponential	0.086	0.043	0.071	0.051
Linear with interaction terms	0.038	0.065	-	-

Table 7.6 The average standard deviation of the bias of the models for different goals. Fighter squadron.

The average standard deviation of the bias, however, seems to be higher for the models estimated by the indirect method than for those estimated by the direct techniques. This may be because of the smaller number of observations used to calculate the standard deviation for the indirect models (cf. Section 7.1), but the deviations must still be considered low.

## 7.5 VARIANCE

For every subject and every goal we calculated the variance of the six regression models (cf. Figure 6.1). To find out whether there were any differences in variance between the models, I used a non-parametric test - the sign test (Conover, 1971). This is a less powerful test than the Wilcoxon signed rank test used in Chapter 6,

but it means that we do not have to assume that the subjects use identical interval scales.

We tested the hypothesis that there was no difference in the variance between the models, for all pairs of models and all goals. The results can be seen in Table 7.7, where there is one matrix for every goal. The result of the tests, i.e. the level of significance or the character N for a non-significant difference, is noted in the row with the largest number of variances exceeding those of the paired model. Thus, the row of a good model will be empty, while the row of a poor model will contain a figure or the letter N. If the letter N is found in both the row and the column of two models, then these models have an equal number of greatest variances - an outcome that is more likely to happen with the sign test than with the Wilcoxon signed rank test. From a study of the columns we can learn from the levels of significance how good the models are.

As we want not only to analyze the six regression models but also to compare these with the four models estimated by the various direct techniques, we will find the matrices in Table 7.7 rather big. However, we can partition every matrix into four submatrices as shown by the heavy lines. The upper left matrix gives us information about the direct models. It can be compared with the matrix in Table 6.12, where a different test has been used.

The lower right matrix tells us about the six regression models. The other two submatrices contain information about the comparisons of the direct and indirect models. As we know that the row of a good model will be empty, it is easy to see from these two submatrices which of the direct and indirect models are the best. Thus if the lower left submatrix is almost empty, then the indirect models have a lower variance; if the upper right submatrix is almost empty, then the direct models have the lowest variance.

## Goal No. 1

	C	K	M	T	Lin	Con	Dis	Log	Exp	Lint
C			N	N					N	
K	N		N	N	.20				N	N
M					N				N	
T	N		N						N	N
Lin	N			N						
Con	.20	.20	.01	.20	.025				.20	.05
Dis	0	.02	.01	.01	.025	.20			.01	.01
Log	.01	.02	.01	.01	.001	.025			.01	.01
Exp	N	N	N	N	.25					
Lint	N		N		N				.20	

C = consultants

K = Keeney

M = Miller

T = Tell

Lin = linear

Con = conjunctive

Dis = disjunctive

Log = logarithmic

Exp = exponential

Lint = linear with  
interaction  
terms

## Goal No. 2

	C	K	M	T	Lin	Con	Dis	Log	Exp	Lint
C		N			.20	N		N	.05	.05
K					N	N		N	N	N
M	N	.20		N	.20	N	N	.20	.02	.05
T	N	.20			.20	N	N	N	.20	N
Lin										N
Con		N			N				N	.20
Dis	.20	N	N	N	.025	.01		N	.025	.001
Log		N			.01	.05			N	.01
Exp		N			.20					.20
Lint										

Table 7.7

Pairwise sign tests of difference in  
variance between the models. Fighter  
squadron.

Goal No. 3

	C	K	M	T	Lin	Con	Dis	Log	Exp	Lint
C		.20	N	.05	N					
K				N						
M		.20		N						
T			N			N				
Lin		N	N	N						
Con	N	N	N		.20					
Dis	N	.05	N	N	.025	.20		.20		
Log	N	N	N	N	.05	.025				
Exp	.10	.01	.20	.20	0	.05	N	.05		
Lint										

C = consultants  
 K = Keeney  
 M = Miller  
 T = Tell

Lin = linear  
 Con = conjunctive  
 Dis = disjunctive  
 Log = logarithmic  
 Exp = exponential  
 Lint = linear with  
 interaction  
 terms

Goal No. 4

	C	K	M	T	Lin	Con	Dis	Log	Exp	Lint
C				N	N					
K	N			N	N					
M	N	N			N	.20	N		N	
T	N	N	N			N				
Lin	N	N		N						
Con	N	N			.025				N	
Dis	N	.20	N	N	.10	N			N	
Log	N	N	N	.20	.10	0	N		.10	
Exp	N	.20		N	N					
Lint										

Table 7.7 continued

The upper left matrices show more N:s than in Table 6.12, indicating that the variances of the compared models surpassed each other on equal number of times. There are still no clear tendencies in any particular direction.

The lower right matrices show that the two linear models have a significantly lower variance than the other models. The exponential model seems to come closest to the linear models, followed by the conjunctive model. The disjunctive and the logarithmic models were significantly outclassed by most of the others.

A comparison of the direct and indirect models shows the direct models to have a smaller variance than the indirect models. The result is seldom significant, but the large number of notations in the lower left submatrices must be considered convincing. Only in the case of the second goal do the two types of model seem to have variances of roughly equal size.

#### 7.6 MEAN SQUARE ERROR

The same analysis as in Section 6.4.4 has been carried out here, except that we have now used the sign test (Conover, 1971). We have tested indirect as well as direct models.

The hypothesis about equal MSE's in all models was tested pairwise, using the sign test for all models and all goals. The results are presented in Table 7.8. We discussed in Section 7.5 how these matrices should be interpreted.

An analysis of the upper left submatrices, containing the results of the tests of the directly estimated models, permits the same conclusions as were drawn in Section 6.4.4.

Turning to the regression models, we find that the two linear models still outclass all the other models, generally significantly. Their superiority appears to increase with the increasing complexity of the problems. This result agrees with the findings of earlier research.

Goal No. 1

	C	K	M	T	Lin	Con	Dis	Log	Exp	Lint
C										N
K	.20		N	.15	.10	.20	N		N	.05
M	N	N			N	N				.10
T	.10		N		N	N				.10
Lin	N		N	N						
Con	N				N					.005
Dis	.001	N	N	.10	.15	.10			.10	.005
Log	.05	N	.10	.10	.10	.001	N		.025	.001
Exp	.025	N	N	.05	.10	.025				.01
Lint					N					

C = consultants  
 K = Keeney  
 M = Miller  
 T = Tell  
  
 Lin = linear  
 Con = conjunctive  
 Dis = disjunctive  
 Log = logarithmic  
 Exp = exponential  
 Lint = linear with  
 interaction  
 terms

Goal No. 2

	C	K	M	T	Lin	Con	Dis	Log	Exp	Lint
C		N	N		.001	.01	.05	.01	.05	.001
K	N		N	N	.10	.10	.20	.20	.20	.20
M		N		N	.01	.05	N	.05	.05	.01
T	N				N	.10	N	.05	.05	.05
Lin										N
Con					N					N
Dis			N		.001	.10		.10	.10	.10
Log					N	.10			N	.025
Exp					.15	N		N		N
Lint					N					

Table 7.8 Pairwise sign tests of difference in MSE between the models. Fighter squadron.

Goal No. 3

	C	K	M	T	Lin	Con	Dis	Log	Exp	Lint
C			.15	N	.10	.10	.15	.15	N	
K	.20		.01	.20	.001	.025	.01	.025	.15	
M				N	.15	N	N	N	N	
T	N				N	N	N	N		
Lin				N						
Con					.05					
Dis				N	.05	.15		.25		
Log				N	.05	.005				
Exp				N	.005	.025	.05	.01		
Lint										

C = consultants

K = Keeney

M = Miller

T = Tell

Lin = linear

Con = conjunctive

Dis = disjunctive

Log = logarithmic

Exp = exponential

Lint = linear with  
interaction  
terms

Goal No. 4

	C	K	M	T	Lin	Con	Dis	Log	Exp	Lint
C			N	N	N	N	N	N	N	
K	.025		.20	.05	.01	.01	.01	.05	.20	
M				N	.025	.10	.10	N	N	
T	N		N		.10	.05		.05	N	
Lin										
Con					.10					
Dis				N	.01	N			.20	
Log					.01	.005	N		N	
Exp					.25	N				
Lint										

Table 7.8 continued

As in the case of variance, we find here too that the disjunctive and logarithmic models are the worst. The conjunctive and the exponential models turned out the same when we studied their variances, but the conjunctive model proves better according to the MSE criterion: the conjunctive model showed the smallest bias of all the models, while the exponential model had the highest.

Finally, comparing the two types of models, we find that the indirect models are significantly better than the direct. The reason for this is the greater bias of the direct models. The indirect models have their greatest bias in the case of goal 1, and there the two types of models seem fairly equal. For the other goals the direct models have a greater bias than the indirect, and so the direct models show the greatest MSE.

Looking at the results in greater detail, we find that Keeney's model is outclassed by the indirect models in almost all comparisons and that the difference is generally highly significant. Of the other three direct-estimated models, Tell's seems to be slightly better than the others, as we would have expected in view of our findings in Chapter 6.

To these findings, we should now add two points that were discussed at the beginning of this chapter. First, the direct models were validated on a sample that was twice as large as the sample used to validate the indirect models. Second, the predictive power of the indirect models would have been even higher if we had not imposed so many restrictions on the regression coefficients.

## 7.7 CONCLUSIONS

The linear model was found to be the best and the disjunctive the worst of the indirect models, irrespective of the criterion used (except the bias). It is more difficult to draw any conclusions about the precision of the conjunctive, exponential, and logarithmic models.



Taking bias, variance, and mean square error as our criteria, we found that they ranked in the order: conjunctive, exponential, logarithmic. All the correlation criteria gave us the ranking: exponential, logarithmic, conjunctive. We can sum this up by saying that the linear model is the best and the disjunctive the worst, while the conjunctive, exponential, and logarithmic models lie between the other two and are indistinguishable. These results agree with those of other studies, e.g. Goldberg (1971). We found no difference between the precision of the pure linear model and the linear model with interaction terms.

In this chapter we also compared direct and indirect models. The indirect models, as expected, proved better than the direct models when the mean square error was taken as the criterion. But in terms of variance and correlation coefficients, the direct models appeared to be better than the indirect.



## CHAPTER 8

OTHER ASPECTS OF THE MODELS

The second and third experiments gave us information about some other aspects of the methods and models. These experiments will be presented briefly in Section 8.1. In the two following sections we will examine the outcomes of the criteria used to evaluate the methods and the models (8.2 and 8.3). In each section we will differentiate between quantitative and qualitative aspects.

8.1 PRESENTATION OF THE EXPERIMENTS

The second experiment was presented in Section 6.1. There we mentioned that the subjects received a booklet consisting of four parts, one for each method, in random order. Immediately following each part was a short questionnaire. Three questions - concerning the believed precision of the model the subject has just constructed, the ease with which it could be understood, and the ease with which the questions posed by the methods could be dealt with - were answered with the help of a nine-grade scale of the Lickert type. The subjects measured the time required by each method.

Another questionnaire was presented to the subject after he had used all four methods. It contained similar questions, but the subject now had to rank the models. The fact that the questions were asked twice made it possible to check the consistency of the subjects' answers, and thus to increase our confidence in the results.

The purpose of the third experiment, carried out fourteen days after the second, was to investigate the subjects' opinions of the models. Three different types of models were tested. The participants received a booklet consisting of three parts (models) in random order. At the end of the booklet was a questionnaire.

To reduce any errors of leniency (Guilford, 1954) we labeled the models A, I, and M.<sup>1)</sup> The A-model uses an additive compound utility function derived from Miller's or Tell's methods, or possibly from Keeney's method if the parameters add up to one.<sup>2)</sup> If the sum of the parameters is not one, Keeney's method will give us a multiplicative model represented by the model M. The I-model has an additive compound utility function with interaction terms, but here all parameters are estimated separately by the consultants' method.

The parameters of the models were fictitious, although reasonable, and identical for all subjects. The subjects received the models in random order and were asked to use them to evaluate three different squadrons. Questions similar to the ones used in the second experiment were posed in a questionnaire at the end of the booklet, and these were to be answered by ranking the models.

## 8.2 CRITERIA REGARDING THE CONSTRUCTION OF MULTIPLE-CRITERIA MODELS

### 8.2.1 Qualitative aspects

When all the methods had been used, the subject was asked some questions whereby he was to rank the methods. These questions and the participants' answers are presented in Table 8.1.

First the subject was asked to make an overall evaluation of the methods. He was free to choose and emphasize attributes as he saw fit. This general question was asked first to prevent the subject from being influenced by the more specific questions that followed.

Our null hypothesis that the ranking of the methods was random was tested by Friedman's two-way analysis of variance (Conover, 1971).

---

1) In this presentation we use the mnemonic names A (for additive), I (for additive with interaction terms), and M (for multiplicative). During the experiment we labeled the models K, L, and M respectively.

2) The chance of the parameters from the consultants' method giving us an additive compound utility function is very small, and we can therefore ignore the possibility.

Attribute (question)	Level of significance			Ranking of the methods		
	Fighter	Reconn.	Total	Fighter	Reconn.	Total
Considering all attributes, i.e. which method the subject wants the Air Force to use	1.0	5.0	0.1	T M C K	M T C K	(TM) C K
Believed precision of the model	5.0	25.0	1.0	T M C K	(TM) C K	T M C K
Ease of understanding the questions posed by the method	0.5	25.0	0.1	T M C K	T M C K	T M C K
Ease of finding the information required by the method	1.0	not	0.1	T M C K	T M C K	T M C K
Number of observations (the lower number only on the believed precision)	14 or 11	6 or 5	20 or 16			

Table 8.1 Level of significance and ranking of the methods based on the ranking questions. Methods/models in parentheses obtained the same sum of the ranks.

The left-hand columns in Table 8.1 give us the levels of significance. In the right-hand columns the methods are ranked according to the sums of their ranks. The "best" method is located to the left and the "worst" to the right.

The low levels of significance for reconnaissance squadrons are probably due to the small size of the sample. The question of the believed confidence in the precision of the models also produced low levels of significance. This can probably be explained by the subjects' inability to see at this stage what the models would look like. As yet, they had only answered the questions posed by the methods, and they probably found it hard to see how their answers could form a model. This explanation seems to be confirmed by the fact that fewer subjects answered this question than answered the others.

If we look at the sum of the ranks of the methods, we will find that Tell's method is considered the best in eleven elements and Keeney's the worst in all twelve elements of the table. Miller's method seems to be preferred to the consultants' in all elements. To find the difference between the methods that caused the significant results, we must turn to a technique for making multiple comparisons at a given level of significance.

The last three of the questions analyzed above were also asked in direct connection with each method. Here a nine-grade scale was used, and the subject was asked to indicate his attitude by a mark on a nine-grade scale. The scale ranged from very easy (to understand etc.) to very difficult. The hypothesis that the methods are equal (have the same mean) was tested with the help of a multivariate technique, namely the analysis of repeated measurement (Morrison, 1967). The results are presented in Table 8.2.

Attribute (question)	Level of significance			Ranking according to averages		
	Fighter	Reconn.	Total	Fighter	Reconn.	Total
Believed precision of the model	5.0	not	0.5	T M C K	M T (CK)	T M C K
Ease of understanding the questions posed by the method	0.5	1.0	0.5	T C M K	C T M K	T C M K
Ease of finding the information required by the method	2.5	not	0.5	T C M K	C T M K	T C M K
Number of observations	16 or 15	6 or 5	22, 21 or 20			




 significant difference at the 5 % level  
 significant difference at the 10 % level  
 significant difference at the 20 % level

Table 8.2 Significances and ranking of the methods based on the scale questions. Methods/models in parentheses obtained the same sum of the ranks.

In seven of the nine elements of the table, significant differences were found. The resemblance with Table 8.1 is great. Both show lower levels of significance for reconnaissance squadrons, and for the question about the believed precision of the models.

If we look at the rankings of the methods according to their averages, we find that in most of the elements Tell's method is considered the best and Keeney's the worst. Of the other two methods, the consultants' seems the easier to use but more confidence is felt in Miller's. Thus, the results are very similar to those derived from the ranking questions.

Our purpose in this analysis is to study which method or methods might have caused the significant differences. We analyzed all six pure contrasts of the type  $\mu_i - \mu_j = 0$  for all questions and all types of squadrons. The significant differences are marked by arcs in the right-hand columns in Table 8.2.

This analysis shows that Tell's and Keeney's methods generally differ significantly. From the totals we can also see that the consultants' and Miller's methods are clustered together half way between Tell's and Keeney's, and that the order between them seems random, i.e. they are very similar.

### 8.2.2 Quantitative aspects

The hypothesis that the methods require an equal amount of time was tested with the help of an analysis of repeated measurement (Morrison, 1967).

The time requirements for fighter and reconnaissance squadrons are very different, as the fighter-evaluation problem is more complex. There is thus no point in combining the data into a total.

Tell's method requires less time than the others and Keeney's requires the most - 16 and 32 minutes respectively for the fighter-squadron problem. The differences between these two methods and

between Tell's and Miller's methods are the only significant pure contrasts.

It is interesting to compare these time requirements with those of the indirect methods. To estimate all six methods by the indirect method we need at least 16 evaluated squadrons. As it took 5.5 minutes on an average to evaluate one squadron, the indirect method will require about 90 minutes. This should be compared with the 32 minutes required by the most time-consuming direct method. Thus we conclude that the time requirement of the direct methods is, as expected, much lower than that of the indirect methods.

Attribute	Level of significance		Ranking according to averages	
	Fighter	Reconn.	Fighter	Reconn.
Time requirement	0.5	1.0	T C M K	T C M K
Number of observations	13	6		

— significant difference at the 5 % level

- - - significant difference at the 10 % level

Table 8.3 Significance levels and ranking of time requirements.<sup>1)</sup>

### 8.2.3 Some comments made by the participants

The subjects were asked to comment on the methods while they were using them. Their response will give us more insight than the answers to our questions alone. The more remarkable tendencies are the following.

- 1) As mentioned above the procedure for finding the utility functions by the consultants' method was never used in the preliminary test. The time for this method is thus underestimated. A more correct ranking is T M C K for fighter squadrons, as M and C were rather close but far away from K, and this is the ranking that will be used in the summary.



The consultants' and Keeney's methods were criticized for posing unrealistic or impossible questions by nine and ten subjects respectively. It was said, for example, that the situation will never occur in which all crews can take off but none can land. The reason may be technical or educational (you always learn to land before you learn to take off). More subtle examples were pointed out by several subjects.

Several subjects asked us to explain Keeney's method during the test (this accounted for five of seven questions asked in the course of the experiment) and two participants noted on their forms that they did not understand the technique.<sup>1)</sup>

Three subjects said it was easy to answer the questions in the consultants' method. It would be interesting to know whether this was because this method evaluated extreme combinations (all or nothing), or because the subjects found this type of evaluation easy to make.

One subject said he appreciated the coefficients of Miller's method by which it was possible to adjust for the interpretative quality of the performance measures. Some participants said they preferred ranking and pairwise comparisons (Miller's method), while others found it easier to use probabilities (Tell's method).

One subject indicated that Tell's method had a great advantage in its probabilities. These should be presented to the squadron leaders, he said, so they could be used to indicate priorities between various elements in the crew-training.

#### 8.2.4 Conclusions

The study shows very clear results. Tell's method received the most positive ratings and Keeney's method the most negative on all

---

1) Colleagues at the Stockholm School of Economics had praised the description of Keeney's method.

the attributes that were used to analyze the subjects' attitudes to the methods. Miller's and the consultants' methods were ranked in between these two extremes, and were generally indistinguishable. Time requirements produced the same result as the other criteria.

From the participants' spontaneous comments we were able to confirm our belief that the consultants' and Keeney's methods involved questions that were unrealistic and therefore very difficult or impossible to answer. The experiment also confirmed our hypothesis that Keeney's lottery technique is hard to explain.

### 8.3 CRITERIA REGARDING THE USE OF MULTIPLE-CRITERIA MODELS

#### 8.3.1 Qualitative aspects

The hypothesis that the models were equal, i.e. that the subjects' rankings were random, was tested by the Friedman two-way analysis of variance (Conover, 1971). The results are presented in Table 8.4.

The low levels of significance seem to be due to few observations.

We note that the subjects, as in the second experiment, find it hard to express any opinion on the believed precision. There were relatively few answers to this question.

The ranking shows a high level of agreement and the overall judgment rated the additive model (A) very favorably.

The right-hand columns show us the rankings that may have produced the models. We can see that these rankings are very similar to all those presented earlier. Keeney's model, however, seems more appreciated than his method.

#### 8.3.2 Quantitative aspects

The time required by these models can be mathematically expressed and was not measured. It takes more time to perform a division than a multiplication, and both require more time than an addition or a subtraction. This holds for calculations by man (with or without the help of a machine) as well as by computer.

Attribute (question)	Level of significance		Ranking of models by sums of ranks		Ranking of methods that may have generated the tested models. Ranking made by sums of ranks.	
	Fighter	Total	Fighter	Total	Fighter	Total
Considering all attributes, i.e. which model the subject wants the Air Force to use	0.5	0.5	A M I	A M I	(M T) K C	(M T) K C
Believed precision of the model	not	not	A I M	A I M	(M T) C K	(M T) C K
Ease of understanding the instructions of the model	2.5	2.5	A M I	A M I	(M T) K C	(M T) K C
Ease of performing the instructions given by the model	0.5	2.5	A M I	A M I	(M T) K C	(M T) K C
Number of observations	11 or 7	14, 12 or 10				

Table 8.4 Levels of significance and ranking by sums of the ranks for the evaluation models.<sup>1)</sup> Models in parentheses obtained the same sum of the ranks.

The time required for evaluating a goal by the three models can be expressed as

$$T_A = n \cdot u + n \cdot m + (n-1)a \quad (8-1)$$

$$\begin{aligned} T_M &= n \cdot u + n \cdot m + n \cdot s + (n-1)m + s + d = \\ &= n \cdot u + (n+1)s + (2n-1)m + d \end{aligned} \quad (8-2)$$

and assuming that there is at least one positive and negative coefficient,

$$T_I = n \cdot u + \sum_{k=0}^n \frac{n}{k} \cdot k \cdot m + (n-2)a + s \quad (8-3)$$

1) The reconnaissance squadron model is not reported separately, because of the very small number of observations (3 or 1).

$u$  = time required for finding the utility index of one criterion  
 $m$  = time required for multiplying two numbers  
 $d$  = time required for dividing one number by another  
 $a$  = time required for adding one number to another<sup>1)</sup>  
 $s$  = time required for subtracting one number from another  
 $T_i$  = total time required for evaluating one goal using model  $i$ ,  
 $i = A, I, M$   
 $n$  = number of criteria

For the I-model the time needed for calculations can be substantially reduced if some coefficients are zero, but there will still be a lot of multiplications to do and the number will grow exponentially as the number of criteria increases.

It is easy to see that the A-model requires less time than the M-model and that the M-model takes less time than the I-model. The difference between the models will increase as more criteria are used.

To illustrate these conclusions, I have expressed the calculation time as a function of the number of criteria (see Figure 8.1). However, I have excluded the term  $n \cdot u$  from all formulas, i.e. the time it takes to find the utility indices of all  $n$  criteria, as the size of this term is difficult to estimate. The relative importance of this term will depend on whether man or machine makes the calculations, but in both cases it would be small. The relationship between the time required for an addition, a subtraction, etc. is the same as the specified instruction times for the IBM system 370 Model 158.

---

1) The formulas are correct for computers or adding machines. For conventional addition on paper, the marginal time for adding one number is smaller than the average time  $a$ .

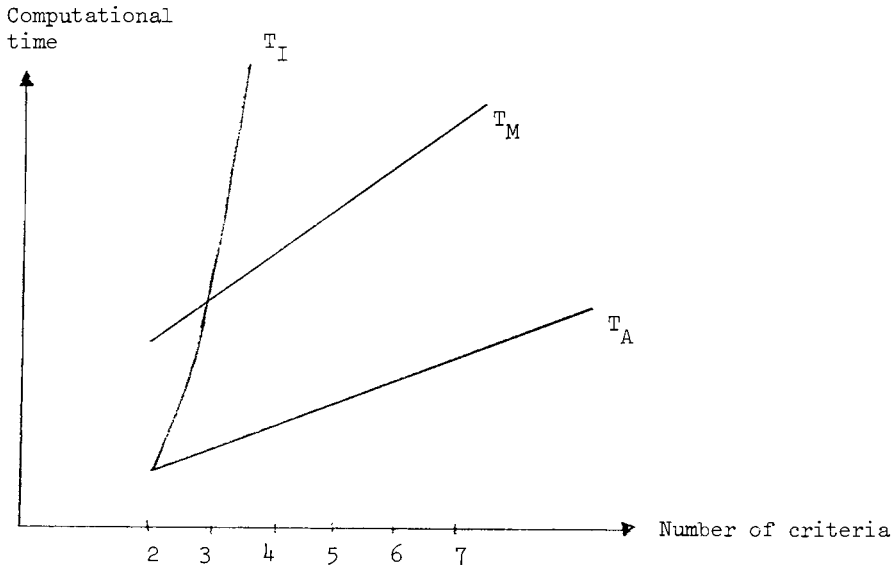


Figure 8.1 Computational time as a function of the number of criteria.

### 8.3.3 Some comments made by the participants

The participants were asked to say whether they had used the evaluation model developed by the Air Staff before. They were also asked to compare this model with the three models tested in the third experiment.

Several subjects did not answer this question, probably because they had not used the Air Force model. (They had nothing with which to compare the tested models.) Of the subjects who answered the question, the great majority had no experience of the Air Force model, and those who had (three subjects) found the additive (A) model (which is in fact the Air Force model) better than "the Air Force model"! This finding suggests that we can ignore any risk of the results having been influenced by prior knowledge of any method or model.

#### 8.3.4 Conclusions

The additive model, represented by Miller's and Tell's models, received the best ranking on all attributes, followed by Keeney's and the consultants' models. The analysis of the time requirements led to the same ranking. These results agree with the hypotheses put forward in Section 3.7.

## CHAPTER 9

THE EFFECT OF UNCERTAINTY ON THE SELECTION OF A UTILITY MODEL  
 USING DIRECT ESTIMATION

=====

In this chapter we will analyze the effect of uncertainty in the estimated parameters and utility functions. It is implicitly assumed that, using a direct estimation method, the decision-maker can accurately state his preferences. In Section 9.1 we will examine the assumption of regression analysis (an indirect technique) and in Section 9.2 we will look at some psychological studies of men's ability to make judgments. In Section 9.3 I will present some studies of other problems, in which the effect of random errors in estimates has been analyzed. In Section 9.4 I will study what effect uncertainty in the estimates may have on the selection of a utility model using direct methods.

#### 9.1 THE INDIRECT METHOD OF ESTIMATING UTILITY MODELS

In Section 2.2.1 we presented the regression technique for use in estimating utility models. According to this technique we assume one model to be correct. The parameters of this model are estimated so as to minimize the sum of the square of the differences between the utility index given by the decision-maker and the index generated by the model. This deviation is also called the random error of the model, and it is generally assumed that it is attributable to the dependent variable only, i.e. to the utility index, as it is usually assumed that there are no errors in the criteria.

The random error may have two sources: measurement errors, and stochastic errors due to incomplete reproduction of the correct model (Press, 1972; Wonnacott and Wonnacott, 1972). In our experiment the subjects were given information on a specified number of criteria for each squadron and were asked to evaluate the squadron on a basis of this information only. Thus errors must be the result of the

measurements (in a wide sense), because the model consists of as many criteria as the decision-maker uses.

Thus we observe that the regression approach allows for measurement errors in the dependent variable, i.e. it recognizes the decision-maker's inability to evaluate the multiple-criteria object quite correctly, something the direct approaches do not.

At this point we should look at a number of psychological studies that bear witness to man's inability to handle complex sets of data adequately, and which therefore show how important it is to investigate what influence random errors in the values given by a decision-maker can have on the selection of a model.

## 9.2 THE ESTIMATION PROCESS

It is a well-known psychological fact that people do not always make the same choice, or assign the same rating to an object, if they are placed in the same choice situation at different times. This seems to hold even if the circumstances of the choice appear to be absolutely identical in all relevant respects, and if the time elapsing between the two choices is very small (Davidson and Marschak, 1959). It is this phenomenon that makes the use of test-retest procedures necessary in any study of human behavior.

In Section 1.3 we suggested that one reason for constructing evaluation models was the limited ability of human beings to handle complex sets of stimuli. For example, Whitmore and Cavadias (1974, p. 616) have explicitly considered this when they evaluated their multiple-criteria model as they assumed the evaluator's estimate to be "subject to inherent statistical error". But it is not only man's inability to make these evaluations that gives rise to random errors. The preferences of the subjects for certain numbers (Edwards, 1954), or the introduction of rounding off figures (Becker, DeGroot, and Marschak, 1964) all make errors in the decision-maker's answers more likely.



Few writers on multiple-criteria methods discuss this uncertainty problem, but Churchman and Ackoff (1953) are aware of it, and suggest that estimation methods should provide an estimate of the accuracy of the judgments concerned.

These and other psychological studies show that all answers given by man are subject to random errors. And the same naturally applies to man's ability to give reliable weights and utility functions as required by the direct approach we are discussing here. There are, however, some other theories that ought to be mentioned as they give some interesting information regarding our estimation problem.

The theory of fuzzy sets (Bellman and Zadeh, 1970) has influenced some theoretical work in which the decision-maker's inability to correctly state his preferences is explicitly considered. For example, Zeleny (1973) and Steuer (forthcoming) suggest programming methods where the decision-maker is asked to specify interval criterion weights only, since these writers feel it is too much to expect the decision-maker to give point estimates.

Random utility models represent another theory that is applicable to our problem. This theory argues that the alternatives or the objects to be evaluated are so complex that one person can only consider at any time a subset of the criteria characterizing the alternatives. Thus the weights expressed will have no fixed value but will fluctuate according to the decision-maker's mood etc, regardless of any change in his preferences (Block and Marschak, 1960; Luce and Suppes, 1965; and Becker and McClintock, 1967). Shepard (1964) provides psychological support for this model.

So far our discussion has concerned the (measurement) errors that anyone may make. Although in this study we have restricted ourselves to individual decision-making, it seems appropriate to say something here about group estimates, as this is probably the type of estimate most commonly used in real-life problems.

When a group of people has to estimate a utility function or the relative weights of different criteria for an organization, the opinions they express will generally diverge somewhat. This may depend on the sort of factors we have discussed above, but it may also stem from the fact that different people represent different functions in the organization and are therefore more concerned with some criteria than with others (Souder, 1975). Variations of this kind can often be reduced by the use, for example, of the Delphi technique (Quade, 1968). Or the decision-makers can be asked individually to estimate the weights etc., after which the average of their judgements can be taken as the estimate of the group as a whole (Churchman and Ackoff, 1953; Miller, 1970). In both these cases the estimate of the group will be a random variable.

All this suggests that it is realistic to consider all estimates given by the decision-maker as random variables. Thus we write the expressed utility index associated with criterion  $i$  when it takes on the value  $x_i$  as

$$u_i(x_i) = v_i(x_i) + \epsilon_1 \quad (9-1)$$

and the expressed weight of criterion  $i$  as

$$a_i = \alpha_i + \epsilon_2 \quad (9-2)$$

where  $v_i(x_i)$  and  $\alpha_i$  are the decision-maker's true values and  $\epsilon_1$  and  $\epsilon_2$  are two error terms.

We will now look at some few studies in which estimates have been regarded as random variables.

### 9.3 STUDIES OF THE EFFECT OF ESTIMATION ERRORS ON MODEL PERFORMANCE

There are thus many grounds for believing that people make errors when they are evaluating objects or stating preferences. We can

now look at what are, so far as I know, the only studies in which the effects of estimation errors have been analyzed. These studies are concerned with the effects of errors in the model, and should be distinguished from studies of the effects of errors in the data that describes the objects to be evaluated.

Dyer (1974) has made a mathematical analysis of the effect of estimation errors (approximation errors and stochastic errors) when estimating the gradient used in some interactive programming methods. The motivation of the study was the results of some experiments that showed that the subjects had some difficulty in performing their task, "so that errors in these estimates should be expected" (Dyer, 1974, p. 160).

In an attempt to provide curriculum-planning information for elementary school principals, Dyer et al. (1973) derived utility estimates from which unidimensional utility functions could be derived. As they received many inconsistent answers, the authors found they had either to consider these as errors or to define preference and indifference in terms of probabilities of choice. The second of these alternatives was simulated by introducing a stochastic error term associated with the underlying utility function, i.e. as in formula (9-1).

Six differently shaped utility functions were simulated, assuming the error term ( $\epsilon_1$ ) to be normally distributed with a mean of zero. The authors also examined changes in the shape of these functions by letting the standard deviation assume three different values, 0.01, 0.05, and 0.1.

Some studies of the effect of estimation errors have been made in the field of portfolio theory. Most papers in which mean-variance portfolio models are discussed, disregard the problem of estimating the parameters needed for the model and simply take the parameters as known. Kalyon (1971), however, has studied the effect of un-

certainty regarding the means, but he disregards the uncertainty about the variances and covariances used in the model. Frankfurter, Phillips, and Seagle (1971) have studied the simultaneous effect of errors in the estimates of means, variances, and covariances of security returns.

Frankfurter et al. assume that the returns ( $R$ ) on  $s$  securities have a multivariate normal distribution

$$f(R) = f(R_1, R_2, \dots, R_s) \quad (9-3)$$

with means

$$\mu = (\mu_1, \mu_2, \dots, \mu_s) \quad (9-4)$$

and a variance-covariance matrix

$$\Sigma = \left\| \sigma_{ij} \right\| = \left\| \rho_{ij} \sigma_i \sigma_j \right\|. \quad (9-5)$$

The vector

$$\lambda = (\lambda_1, \lambda_2, \dots, \lambda_s) \quad (9-6)$$

indicates the portions of the portfolio that are invested in the  $s$  different securities.

The mean ( $E$ ) and the variance ( $V$ ) of the portfolio are found by

$$E = \mu \lambda' \quad (9-7)$$

and

$$V = \lambda \Sigma \lambda' \quad (9-8)$$

where  $\lambda'$  is the transpose of the row vector  $\lambda$ .

The information available about security returns, whether it be observed or subjective time series data, is used to estimate the means ( $\mu$ ) and the variance-covariance ( $\Sigma$ ). But this information,

according to Frankfurter et al., constitutes only a sample of a multidimensional process. Thus they regard the means and the variance-covariance of the security returns as random variables,  $\tilde{\mu}$  and  $\tilde{\Sigma}$  respectively.<sup>1)</sup> Accordingly the mean and the variance of the portfolios will be random variables,

$$\tilde{E} = \tilde{\mu}\lambda' \quad (9-9)$$

and

$$\tilde{V} = \lambda\tilde{\Sigma}\lambda' \quad (9-10)$$

Assuming the return of the  $i$ -th security to be

$$R_i = \mu_i + \epsilon_i \quad (9-11)$$

where the error  $\epsilon_i$  is normally distributed

$$\epsilon_i \in N(0, \sigma_i^2) \quad (9-12)$$

the authors made a simulation of the mean-variance portfolio model. They found that estimation errors in  $\mu$  and  $\Sigma$  can cause the analyst to select substantially inferior portfolios in a great majority of cases, and they conclude that "the impact of estimation error is so strong that the usefulness of present mean-variance approaches to portfolio selection is brought into question" (Frankfurter et al., 1971, p. 1251).

Finally we can mention a study by Lord (1962). Although this study concerned measurement errors, it was conducted in a manner very similar to ours and therefore deserves to be included in this review. Lord showed that when the errors in the criteria increased, the appropriate conjunctive step function could very well be approximated by a linear function.

#### 9.4 THE EFFECT OF ESTIMATION ERRORS ON THE SELECTION OF A UTILITY MODEL USING DIRECT ESTIMATION

We have seen in the previous sections that the estimates made by decision-makers should be regarded as random variables. This assumption

---

1) The tilde is used to indicate a random variable.

was made in the regression technique. Direct techniques do not allow for errors in the estimates since they follow an axiomatic approach. But even axiomatic models involve estimation and they will therefore be affected by estimation errors. The question will then arise: will our preference for simple additive models be greater than our preference for more complex models, such as an additive model with interaction terms, if all estimates are considered as random variables? Will not the interaction terms add variance, thus reducing the precision of the interaction models? This is the problem we will be discussing in the present chapter.

#### 9.4.1 The problem

We can take the case of a two criteria problem, but this could easily be extended to include more criteria.

Let the correct model be<sup>1)</sup> 2)

$$T(x) = \alpha_1 u_1(x_1) + \alpha_2 u_2(x_2) + (1-\alpha_1-\alpha_2)u_1(x_1)u_2(x_2) \quad (9-13)$$

with

$$0 \leq u_1(x_1), u_2(x_2), \alpha_1, \alpha_2 \leq 1$$

- 
- 1) We will use Greek letters for all population data, while Roman letters will be reserved for sample data or estimates.
  - 2) Had we assumed the correct model to be additive, we would have had no problem, since Keeney's method and the methods estimating additive models would produce identical models.

Estimating<sup>1)</sup> this model by Keeney's method gives us an additive model with interaction terms ( $U_I(x)$ ), called an interaction model,

$$\begin{aligned} U_I(x) &= a_1 u_1(x_1) + a_2 u_2(x_2) + (1-a_1-a_2)u_1(x_1)u_2(x_2) = \\ &= a_1 u_1(x_1) + a_2 u_2(x_2) + u_1(x_1)u_2(x_2) - a_1 u_1(x_1)u_2(x_2) - \\ &\quad - a_2 u_1(x_1)u_2(x_2) \end{aligned} \quad (9-14)$$

Assuming that the decision-maker intuitively normalizes the parameters when using a method that estimates additive models only, or that he explicitly normalizes the parameters generated by Keeney's method, we will get an additive model ( $U_A(x)$ ) of the form

$$U_A(x) = \frac{a_1}{a_1+a_2} u_1(x_1) + \frac{a_2}{a_1+a_2} u_2(x_2) \quad (9-15)$$

In both our models we will require

$$0 \leq u_1(x_1), u_2(x_2), a_1, a_2 \leq 1$$

If the weights are regarded as random variables and if we assume random errors in the unidimensional utility functions, then the compound utility will also be a random variable. Choosing the best model is now a more complicated problem and we must discuss the decision-rule to be used in comparing the two models.<sup>2)</sup>

#### 9.4.2 Decision-rule

When we compare the two models (9-14) and (9-15) it is evident that the additive model is biased while the interaction model is not. I believe, however, that the variance of the additive model is generally smaller than that of the interaction model because of the smaller number of terms in the additive model. In the two-criteria problem the interaction model has three stochastic terms more than the additive

---

1) In this chapter we pretend that the estimation methods are equally effective. Thus all differences between the models are attributable to the models only.

2) An interesting extension of this study would be to examine the effect on the two types of models if the correct model had assumed some other form.

model, and this number increases exponentially as the number of criteria increases. These extra terms will contribute to the total variance, as variances are additive.

Thus, choosing between the additive model (9-15) and the interaction model (9-14) is not a simple matter, as we have to compare bias with variance. The problem is similar to the general scientific problem of choosing between validity and reliability. We can also find such problems in portfolio theory, for example, where we have to choose between the expected return and the risk (variance) of investments (Robichek and Myers, 1965); and in statistics, where we must choose between biased linear estimators and least square estimators (no bias but a fairly high variance) when the data is not orthogonal (Marquardt, 1970). This problem, which is in fact a multi-criterion problem, cannot be solved without introducing the user's utility function for bias and variance in the output data of the evaluation models.

Not having specified any user, I decide here to apply a quadratic utility function around the true value, which is the value given by our correct model. This implies the use of the mean square error (MSE) as our criterion. It is quite common to assume the utility function to be quadratic<sup>1)</sup> (Chernoff, 1960; Hoerl and Kennard, 1970; Andersson, 1972). Another reason for using the MSE is its nice mathematical properties. Our decision-rule will now be to choose the model with the smallest MSE.

The MSE can be decomposed into the sum of the variance and the square of the bias, and for the additive model we will get

$$\text{MSE}(U_A) = V(U_A) + E(U_A)^2$$

and for the interaction model

$$\text{MSE}(U_I) = V(U_I) + E(U_I)^2 = V(U_I)$$

---

1) To indicate that other utility functions may be relevant, I refer again to the Air Force evaluation problem. There an unsymmetric utility function may be better, as it may be considered a more serious error to over-estimate than to under-estimate the capability of the combat units.



as the interaction model is unbiased. This implies for our problem that we will have to compare the MSE of the additive model with the variance of the interaction model. We hypothesized above that the variance of the additive model was the smallest. Following our criterion, we can then say that the additive model will be preferred if the square of its bias is smaller than the difference between the variances.

#### 9.4.3 An analytical approach

To solve the stated problem analytically, we will have to calculate the expected value and the variance of the additive (9-15) and the interaction models (9-14). The calculations for the interaction model are fairly simple, although the expression of the variance consists of more than 40 terms of the fourth order. To calculate the variance of the additive model is a more cumbersome business, as we have to deal with ratios of random variables. Another problem is the large number of terms, and particularly the large number of terms of high orders, which renders it very difficult to make any inferences. Because of these obstacles, I will turn to simulations.

To make it easier to understand some of the results of the simulation, however, we can expand the variance of one of the models somewhat. The variance of the interaction model will be

$$\begin{aligned}
 V(U_I(x)) &= V[a_1 u_1(x_1) + a_2 u_2(x_2) + u_1(x_1)u_2(x_2) - \\
 &\quad - a_1 u_1(x_1)u_2(x_2) - a_2 u_1(x_1)u_2(x_2)] = \\
 &= V(a_1 u_1(x_1)) + \left\{ V[a_2 u_2(x_2) + u_1(x_1)u_2(x_2) - \right. \\
 &\quad \left. - a_1 u_1(x_1)u_2(x_2) - a_2 u_1(x_1)u_2(x_2)] + \text{Cov} \right\} = \\
 &= E(a_1)^2 \cdot V(u_1(x_1)) + E(u_1(x_1))^2 \cdot V(a_1) + V(u_1(x_1))V(a_1) + E
 \end{aligned} \tag{9-16}$$

where Cov stands for all the covariances and R stands for the expression in the big brackets. The expansion of the variance of a product of two random variables is based on Goodman (1960).

We can see that the variance of the interaction model (and the additive model, had we looked at it) depends on the expected value of the utility indices.<sup>1)</sup> This will give us eight variables to include in our study  $a_1, a_2, u_1(x_1), u_2(x_2), V(a_1), V(a_2), V(u_1(x_1)),$  and  $V(u_2(x_2))$ .

#### 9.4.4 A simulation approach

Having found it difficult to solve the problem analytically, we turn to simulation. We regard the decision-maker's revealed utility of each of the unidimensional criteria as well as his revealed parameters as samples from normally distributed and independent variables. We write them

$$a_i = \alpha_i + \varepsilon_{ai} \quad \text{for } i = 1, 2 \quad (9-17)$$

and

$$u_i(x_i) = u_i(x_i) + \varepsilon_{ui} \quad \text{for all } x_i \text{ and } i = 1, 2 \quad (9-18)$$

with

$$\varepsilon_{ai} \in N(0, \sigma_a^2) \text{ and } \varepsilon_{ui} \in N(0, \sigma_u^2)$$

Thus we assume the variances to be equal for all the parameters ( $\sigma_a^2 = V(a) = V(a_1) = V(a_2)$ ) and utility functions ( $\sigma_u^2 = V(u(x)) = V(u_1(x_1)) = V(u_2(x_2))$ ) respectively. The variances of the parameters, on the other hand, may assume values different from those of the functions.

To determine a suitable value of the standard deviations we turned to our experimental study. It indicated that, in estimating the parameters according to Miller's or Tell's methods, the dispersion among the subjects resulted in a standard deviation of about 0.10.

---

1) Note that we do not treat the criterion  $x_i$  as a random variable. Thus our conclusions will hold only for fixed values  $x_i$ . In Section 9.4.7 we will regard  $x_i$  as a random variable.

The standard deviation was greater, 0.15 - 0.20, when the subjects used Keeney's method (cf. Section 6.2.2). I have not from the experimental data calculated any measure of the subjects' dispersion when estimating the unidimensional utility functions. Dyer et al. (1973) used the standard deviations 0.01, 0.05, and 0.10 in their simulations.

As a compromise I selected the standard deviation 0.10 as the normal value for the parameters as well as for the utility functions. To check the sensitivity in the chosen value of the standard deviation, I also made some simulations with standard deviations at 0.05 and 0.15. The lower of these is closer to the values used by Dyer et al. (1973), and the higher appears to be a more accurate reflection of the values obtained in our experiment.

I restricted the estimates to the range zero to one,  $0 \leq a_i, u_i(x_i) \leq 1$ , for  $i = 1, 2$ .

If any  $u_i(x_i)$  fell outside this interval, I generated a complete new set of  $u_i(x_i)$ 's. This means that the normal distribution of  $u_i(x_i)$  was truncated at zero and one. These truncated distributions were of course only found when  $v_i(x_i)$  assumed values close to zero or one. The same approach has been used by Dyer et al. (1973).

If any  $a_i$  fell outside the permissible region, I decided to change it to the end-value of the interval. This means that the normal distribution was truncated, and that the entire truncated probability was assigned to the end-value of the interval. This approach seemed more appropriate for the parameters. These truncated functions, however, seemed only to affect the results when  $\alpha_1$  and  $\alpha_2$  were both around 0.20, and then only slightly. Such low values of  $\alpha_1$  and  $\alpha_2$  were used only in one simulation.

The assumptions and the approach applied in this simulation can perhaps be seen more clearly in Figure 9.1. In the first quadrant the objects to be evaluated are represented by points. It is in

this quadrant that we will draw the indifference curves (compound utility function) when using for example regression analysis. The second and fourth quadrants will contain the unidimensional utility functions of the criteria  $x_2$  and  $x_1$  respectively. Finally, in the third quadrant, we illustrate the compound utility function derived by the direct methods by some indifference curves.

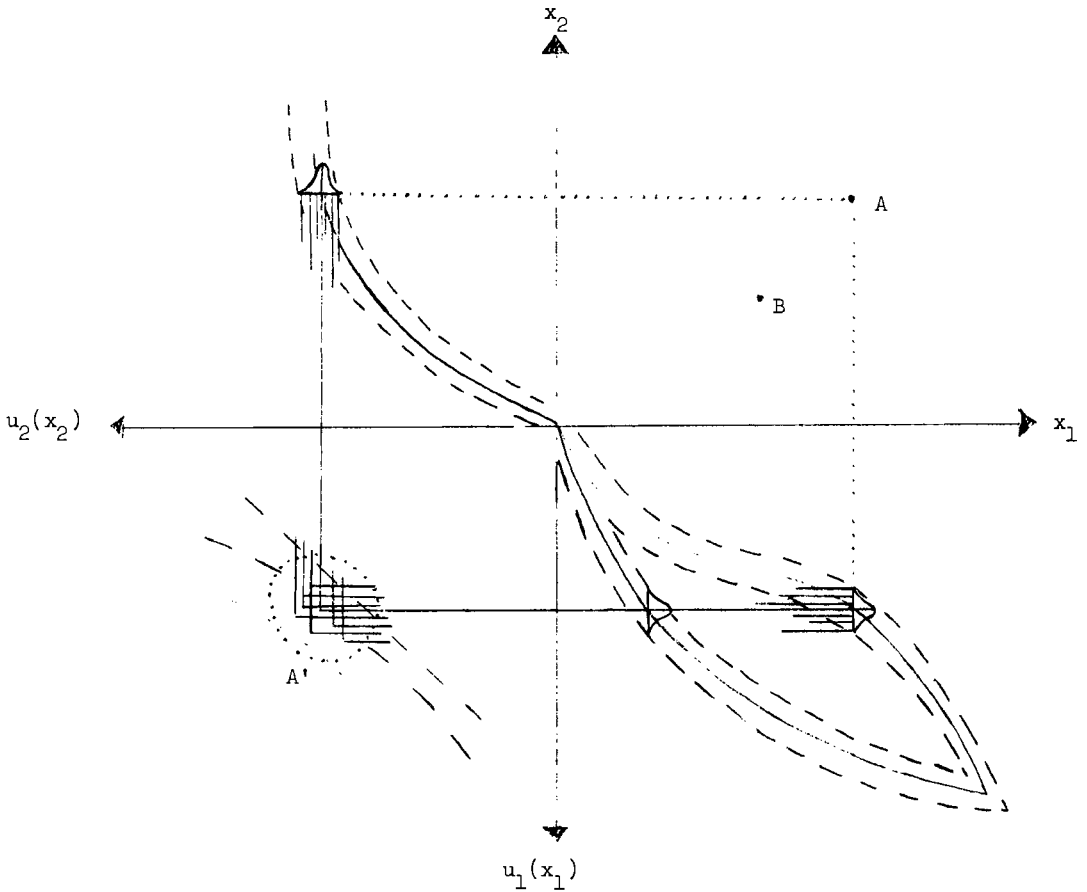


Figure 9.1 Illustration of the problem with uncertainty in the parameters and the unidimensional utility functions.

In the second quadrant we have illustrated one unidimensional utility function and indicated the dispersion around this line. We have used the same assumption as in regression analysis, i.e. that the error is normally distributed around the true line and that the variance is constant and independent of the utility index.

Finally we should note that there is uncertainty not only in the second and fourth quadrants but in the third quadrant as well. There, the weights of the compound utility function are random variables.

In Figure 9.1 we have illustrated an object A, described by the two criteria  $x_1$  and  $x_2$ . The criterion  $x_2$  is transformed by the utility function in the second quadrant to a unidimensional utility index  $u_2(x_2)$ , the value of which varies because of the uncertainty in the utility function. This is illustrated by the many vertical lines drawn from the utility function. The same holds for the criterion  $x_1$ . The compound utility index will vary not only because the two unidimensional utility indices may assume different values, but also because the indices will be weighted to a sum by varying weights or models. This is illustrated by some indifference curves in the third quadrant. All this uncertainty in the estimates implies that point A may assume many utility indices. This is illustrated by the circle marked A' in the third quadrant, containing all the utility indices that may be assigned to the alternative A.

We do not want to make any more assumptions than necessary, and so nothing will be assumed in this simulation about the form of the unidimensional utility function. All we need are the assumptions of the dispersion presented above. As is illustrated in the fourth quadrant, the necessary distribution of the utility index may stem from very different utility functions. This implies that at this

moment the values of the criteria are uninteresting to our present purpose, and will henceforth be omitted.

#### 9.4.5 Method

The principle has been to keep  $u_1$ ,  $u_2$ ,  $V(a)$ , and  $V(u)$  fixed for every simulation and to have the program modify the  $\alpha_1$  and  $\alpha_2$  from an initial set of values in order to trace the lines where the MSE of the two models are equal. Thus every simulation generates one area where the additive model was preferred according to our criterion, and another area where the interaction model was found superior.

We can now examine our simulation in greater detail. In Figure 9.2 the simulation program has been reproduced as a block diagram.

I began every simulation by initiating the values of  $u_1$ ,  $u_2$ ,  $V(a)$ , and  $V(u)$ . The initial values of  $\alpha_1$  and  $\alpha_2$  were generally set at 0.2 and 0.8 respectively. For several reasons there was generally little point in investigating the more extreme combinations of  $\alpha_1$  and  $\alpha_2$ . First of all the model would tend to become lexicographic if the parameters were too extreme. We would also encounter problems with our truncated distributions, as it would be difficult to see which distribution we had actually used. Also, the cost of the simulations made it necessary to restrict the study to the most interesting combinations only. Finally, in the experiment I never in fact encountered any combinations of weights more extreme than 0.2 and 0.8.

Next we generated the estimates, that is  $a_1$ ,  $a_2$ ,  $u_1$ , and  $u_2$ . These were checked against the requirements discussed above; if they did not fulfil the requirements, new values were generated. The compound utility index was then calculated with the help of the additive model,  $U_A$ , and the interaction model,  $U_I$ .

New estimates were generated and utility indices calculated until

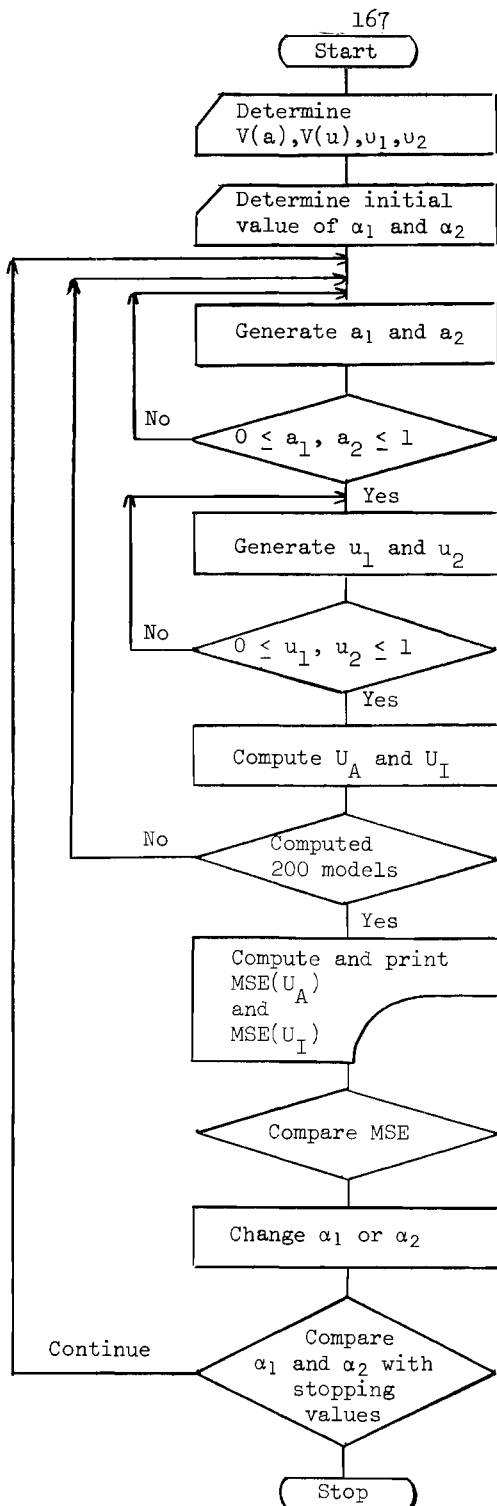


Figure 9.2 Block diagram illustrating the simulation.

200 indices of each model had been obtained. I then calculated the bias, the variance, and the MSE of each of the two models. I compared the two MSE and, depending on the outcome of the comparisons, changed the value of either  $\alpha_1$  or  $\alpha_2$ . To one of these, 0.02 was either added or subtracted, so that the line  $\text{MSE}(U_A) = \text{MSE}(U_I)$  was being followed as closely as possible. New estimates were generated and utility indices calculated, and so on. The simulation continued as long as  $0.2 \leq \alpha_1, \alpha_2 \leq 0.8$ . In some cases other stopping rules were used.

#### 9.4.6 Results

We will first look at the deterministic solution. Figure 9.3 illustrates this case. If there is no uncertainty in the estimation procedure, then the additive model will be preferred only when the true model is additive, that is when  $\alpha_1$  and  $\alpha_2$  add up to one. This line in Figure 9.3 represents all such combinations. This deterministic line will be drawn in all figures as a line of reference.

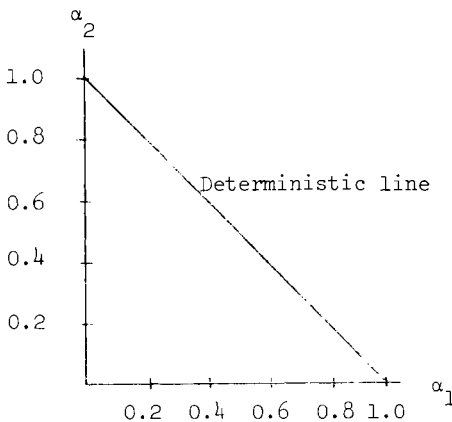


Figure 9.3 The deterministic solution



Let us now turn to the results of the simulations. To begin with, we will keep the standard deviations at their normal level, which was 0.10.

Figure 9.4 illustrates the case when  $v_1 = v_2 = 0.90$ . The MSE of the additive model is smaller than the variance (or MSE) of the interaction model in the area between the two lines. This area will be called the additive area. Thus, for combinations of  $\alpha_1$  and  $\alpha_2$  in this area, the additive model produces a smaller MSE than the interaction model and is preferred according to our criterion.

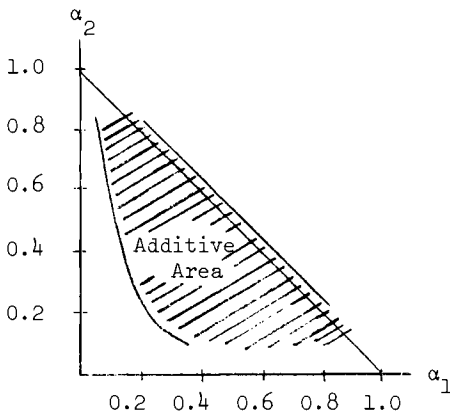


Figure 9.4 The additive area for  $v_1 = v_2 = 0.90$  and  $\sigma_a = \sigma_u = 0.10$ .

We can see from Figure 9.4 that when the parameters add up to more than one, we will use the interaction model only when the sum is greater than 1.05. If the sum is smaller than one, we will also have to consider the relationship between the parameters. If they are equally important, we will prefer the additive model when the sum of the parameters is as little as 0.38. When the parameters differ in importance their sum - which is the criterion of when we

should start using the additive model - has to increase as one of the parameters becomes more important.

Let us now study how the additive area is affected by changes in  $v_1$  and  $v_2$ . We will keep the standard deviations at the normal level and we will keep  $v_1 = v_2$ . To complete Figure 9.5, the deterministic line from Figure 9.3 and the additive area of  $v_1 = v_2 = 0.90$  from Figure 9.4 have been inserted.

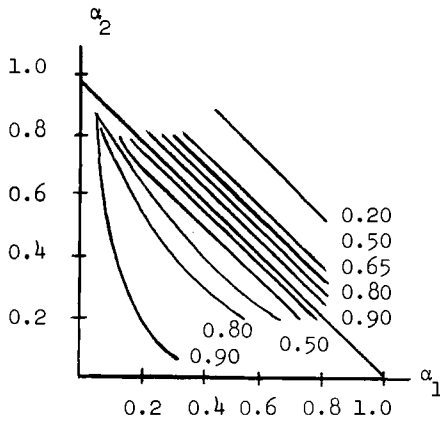


Figure 9.5 The additive area for  $(v_1, v_2)$ ,  $(0.9, 0.9)$ ,  $(0.8, 0.8)$ ,  $(0.65, 0.65)$ ,  $(0.5, 0.5)$ , and  $(0.2, 0.2)$  with  $\sigma_a = \sigma_u = 0.10$ .

We note that the area where the additive representation is superior becomes more symmetrical with the deterministic line as the means of the utility indices approach 0.50. The additive area will be  $0.85 \leq \alpha_1 + \alpha_2 \leq 1.14$  when  $v_1 = v_2 = 0.50$ .

When the means of the utility indices decrease from 0.50, the asymmetry increases again. But the area of the additive model will now reflect the areas for utility indices greater than 0.50. We also note that the additive area seems to be symmetrical with the  $45^\circ$ -ray from the point of origin for any value of the utility indices.

Next we investigated the effect on the additive area of variations in the means of the utility indices. This is illustrated in Figure 9.6. We fixed  $u_1$  at 0.90 and varied  $u_2$ . The standard deviations were still kept at their normal level.

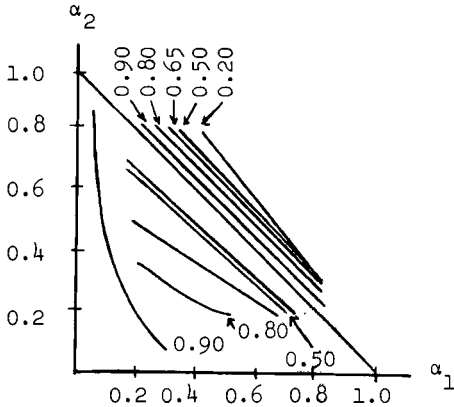


Figure 9.6 The additive area for  $u_1 = 0.90$  and  $u_2 = 0.90, 0.80, 0.65, 0.50,$  and  $0.20$  with  $\sigma_a = \sigma_u = 0.10$ .

The most noticeable change is that the additive area is no longer symmetrical with the  $45^\circ$ -ray from the point of origin. Instead the lines separating the additive and the interaction areas seem to be rays from the point  $\alpha_1 = 1.0$  and  $\alpha_2 = 0.0$ . These lines now seem to turn upwards towards the deterministic line very much later.

Comparing the additive areas of Figures 9.6 and 9.5, we note that the additive areas above the deterministic line have become smaller, while those below the line have become larger. Thus the additive area seems to have moved towards the point of origin. The shift seems to increase with the difference between the means of the utility indices.

We examined the effect of various precision in the estimates by varying the standard deviations. For the parameters three different

levels of the standard deviation have been used, 0.05, 0.10, and 0.15, i.e. one level above and one below our normal level. We have simulated for three different combinations of the means of the utility indices (Figures 9.7, 9.8, and 9.9).

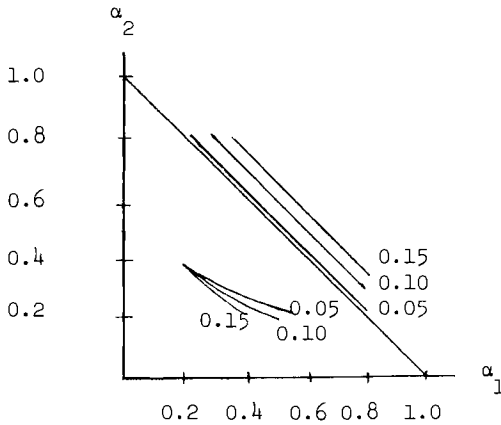


Figure 9.7 The additive area for  $v_1 = 0.90$  and  $v_2 = 0.80$ , with  $\sigma_a = 0.05, 0.10$ , and  $0.15$  and  $\sigma_u = 0.10$ .

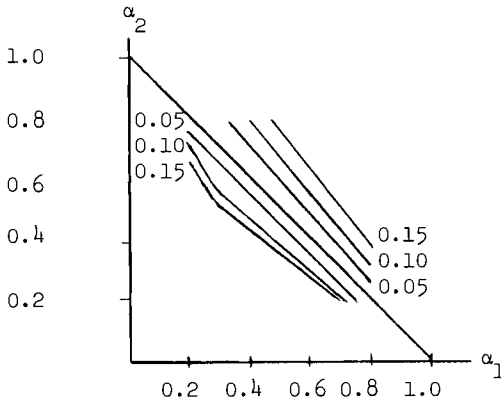


Figure 9.8 The additive area for  $v_1 = 0.65$  and  $v_2 = 0.20$  with  $\sigma_a = 0.05, 0.10$ , and  $0.15$  and  $\sigma_u = 0.10$ .

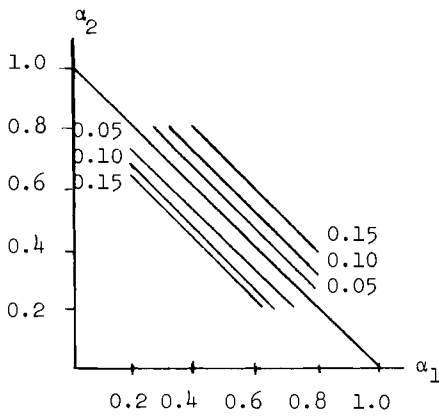


Figure 9.9 The additive area for  $u_1 = u_2 = 0.50$  with  $\sigma_a = 0.05, 0.10,$  and  $0.15$  and  $\sigma_u = 0.10$ .

First we note that the additive area behave in the same way that we have found in the previous simulations. We see that the additive area is symmetric when the means of the utility indices are equal (Figure 9.9) and becomes more assymetric as the means of the utility indices differ from each other. The additive area also moves towards the point of origin as the means of the utility indices depart. All this confirms our findings from Figures 9.3 to 9.6, but we also note that all these conclusions hold when we change the precision of the estimates of the parameters. We also observe that the additive area increases as the standard deviation gets greater - a tendency that was expected.

Turning finally to changes in the standard deviation of the utility indices, we find the results of these simulation in Figure 9.10. The standard deviations of the parameters have been kept at the normal level and the means of the utility indices were both retained at 0.50, so that the results can be compared with those shown in Figure 9.9.

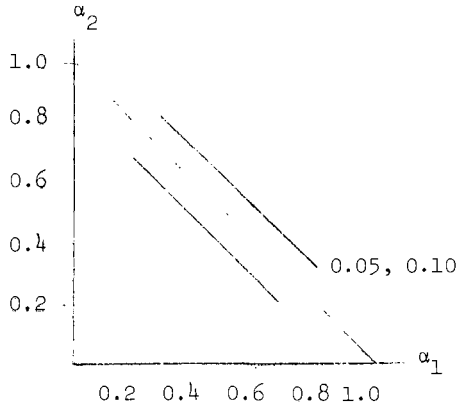


Figure 9.10 The additive area for  $v_1 = v_2 = 0.50$  with  $\sigma_a = 0.10$  and  $\sigma_u = 0.05$  and  $0.10$ .

The simulations with the standard deviations at 0.05 and 0.10 produced no change at all in the additive area.

The simulations indicate that the area, i.e. the combinations of the parameters  $\alpha_1$  and  $\alpha_2$  where the mean square error of the additive model is smaller than that of the interaction model, is fairly large. The size and location of this area is affected by the size of the means of the utility indices and the difference between them, as well as by the size of the standard deviation of the parameters. However, the area seems to be invariant to changes in the standard deviation of the utility indices.

Thus the simulations suggest that when there are errors in the estimates it is often better to choose a simple model (the additive) than to try to estimate the correct model (the additive with interaction terms). These results agree with the findings of Yntema and Torgerson (1961), who found that the average correlation between the true scores and the values generated by a model decreased when the interaction effects were added to the main effects, despite the fact that the subjects in their experiment had in some way taken the interactions into account when making their evaluations.

Figures 9.4 to 9.10 illustrate the results of our simulations. We kept the values of the utility indices fixed, and varied the values of the parameters. By choosing a value for each of the parameters, we can use our results to give us Figure 9.11. This figure has been "turned upside down", as it represents the third quadrant in Figure 9.1.

In Figure 9.11 the utility indices will be our variables. The lines in the figure separate the area - to the left and below the lines - where the additive model is preferred to the interaction model. The lines are approximate, as they are drawn solely from the results of simulations represented by the dots. We can see that for the parameters  $\alpha_1 = 0.30$  and  $\alpha_2 = 0.55$ , the additive area is smaller than the interaction area, while the opposite is true for the parameters  $\alpha_1 = 0.30$  and  $\alpha_2 = 0.65$ .

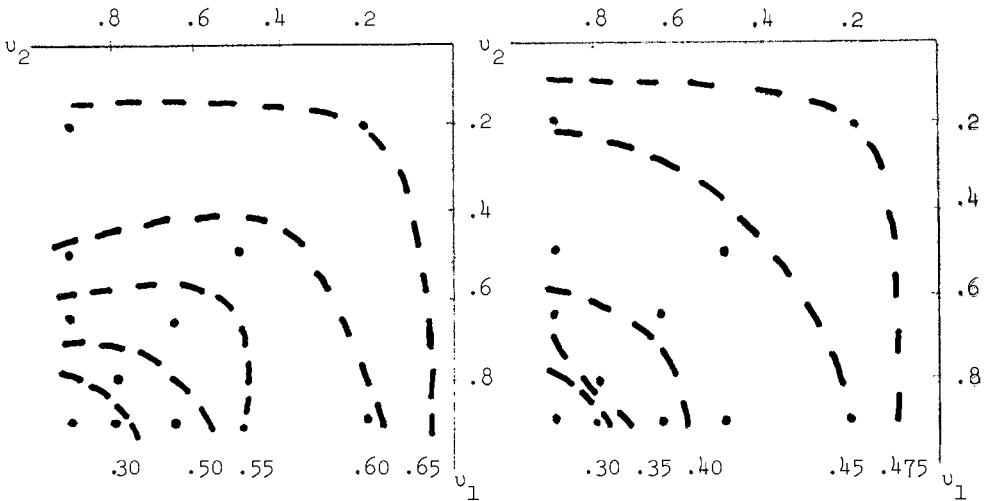


Figure 9.11 The additive area (below and to the left of the lines) for different values of the parameters.

In (a) we have  $\alpha_1 = 0.30$  and  $\alpha_2$  is varied

In (b) we have  $(\alpha_1, \alpha_2)$  (0.30, 0.30), (0.35, 0.35), (0.40, 0.40), (0.45, 0.45), and (0.475, 0.475).

For the parameter values  $\alpha_1 = 0.30$  and  $\alpha_2 = 0.70$  in Figure 9.11a and for  $\alpha_1 = \alpha_2 = 0.5$  in Figure 9.11b the additive model will be preferred for all values of  $u_1$  and  $u_2$ , because the line separating the additive area from the interaction area will coincide with the axes.

#### 9.4.7 Which model will be chosen?

Figures 9.4 to 9.11 only showed the areas in which one model was preferred to another. But to be able to say which model is the best, we have to know the distribution of the utility indices and this distribution will vary from problem to problem. Knowing the distribution of the utility indices, we can calculate the probability of each model - the additive or the interaction - having the greatest precision. This information can be used as the basis on which to choose a model.

It is generally easier to estimate the distribution of the criteria values and to transform it into a distribution of the utility indices, than to estimate the distribution of the utility indices direct. This means that we will now have to incorporate the values of the criteria,  $x_1$  and  $x_2$ , into the analysis.

As the final evaluation of the models is a practical problem, we cannot make any general statements about the precision of the two models. To show how this kind of analysis can be made, and to throw a little light on the results of our simulations, we can carry out two simple numerical examples.

##### 9.4.7.1 Example 1

Let us assume that the two criteria are statistically independent, and that the (marginal) probability density function of the first criterion is

$$f_1(x_1) = 2x_1, \quad 0 \leq x_1 \leq 1 \quad (9-19)$$



and that of the second criterion is

$$f_2(x_2) = 2x_2, \quad 0 \leq x_2 \leq 1. \quad (9-20)$$

This implies that the mean of the simultaneous function

$$f(x_1, x_2) = 4x_1x_2 \quad (9-21)$$

is  $\bar{x}_1 = 2/3$  and  $\bar{x}_2 = 2/3$ . These values are not unrealistic for the Air Force problem examined in this study.

Let us further assume the two unidimensional utility functions to be

$$v_1(x_1) = x_1^{2/3}, \quad 0 \leq v_1(x_1) \leq 1 \quad (9-22)$$

and

$$v_2(x_2) = x_2^{1/2}, \quad 0 \leq v_2(x_2) \leq 1. \quad (9-23)$$

It is now possible to determine the marginal density function of the utility indices. These are found by inserting the formulae (9-22) and (9-23) into (9-19) and (9-20) respectively, and transforming these functions so that they fulfil the requirements of a density function. These functions will become

$$g_1(v_1(x_1)) = 3v_1(x_1)^2 \quad (9-24)$$

and

$$g_2(v_2(x_2)) = 4v_2(x_2)^3 \quad (9-25)$$

The means of these density functions will be

$$\bar{v}_1(x_1) = 3/4 \quad \text{and} \quad \bar{v}_2(x_2) = 4/5$$

As the functions are independent, we can write the simultaneous density function as

$$g(v_1(x_1), v_2(x_2)) = 12v_1(x_1)^2 v_2(x_2)^3 \quad (9-26)$$

Let us now examine for some values of the parameters  $\alpha_1$  and  $\alpha_2$  which model is to be preferred, i.e. which one is most likely to produce the lowest MSE. To illustrate the procedure let us take the parameter combinations shown in Figure 9.11a. To simplify the calculations I have approximated the regions where the additive model is preferred, as in Table 9.1.

The probability of the additive model having the lowest mean square error, which can be seen in the right hand column in Table 9.1 for some values of  $\alpha_1$  and  $\alpha_2$ , is found by

$$P(\text{MSE}_A < \text{MSE}_I) = \int_{v_1^*(x_1)}^1 \int_{v_2^*(x_2)}^1 12v_1(x_1)^2 v_2(x_2)^3 dv_1 dv_2 \quad (9-27)$$

where the values of  $v_1^*(x_1)$  and  $v_2^*(x_2)$  can be found in the column in the middle of Table 9.1.

Parameter values		Additive model preferred when $v_1(x_1)$ and $v_2(x_2)$ are greater than		Probability of the additive model having the lowest MSE
$\alpha_1$	$\alpha_2$	$v_1^*(x_1)$	and $v_2^*(x_2)$	
0.30	0.30	0.80	0.80	0.79
0.30	0.50	0.65	0.65	0.95
0.30	0.55	0.50	0.50	0.99
0.30	0.60	0.35	0.35	1.00
0.30	0.65	0.10	0.10	1.00
0.30	0.70	0.00	0.00	1.00

Table 9.1 The probability of the additive model having the lowest MSE for some different parameter values.

With the assumed density function, this gives probability values that are very high and very insensitive to variations in at least one of the parameter values.

#### 9.4.7.2 Example 2

Let us now choose another density function to illustrate how a model can be selected in a practical situation. The function used in Example 1 was not unrealistic for the Air Force evaluation problem, but let us now take a less extreme function, e.g. the normal distribution.

Let us assume immediately that we have found the marginal density functions of the utility indices to be normally distributed with the means  $\bar{v}_1(x_1) = 0.5$  and  $\bar{v}_2(x_2) = 0.5$  and with standard deviations of 0.1, i.e.

$$v_1(x_1) \in N(0.5, 0.01) \quad (9-28)$$

and

$$v_2(x_2) \in N(0.5, 0.01) \quad (9-29)$$

These functions correspond to the density functions (9-24) and (9-25) in our first example. The probability of the additive model having the lowest mean square error is given in Table 9.2 for the same parameter values as in the previous example.

Parameter values		Additive model preferred when $v_1(x_1)$ and $v_2(x_2)$ are greater than		Probability of the additive model having the lowest MSE
$\alpha_1$	$\alpha_2$	$v_1^*(x_1)$	and $v_2^*(x_2)$	
0.30	0.30	0.80	0.80	0
0.30	0.50	0.65	0.65	0.02
0.30	0.55	0.50	0.50	0.42
0.30	0.60	0.35	0.35	0.93
0.30	0.65	0.10	0.10	1.00
0.30	0.70	0.00	0.00	1.00

Table 9.2 The probability of the additive model having the lowest MSE for some different parameter values.

We find that the probabilities are much lower in this example. This is due partly to the lower means of the marginal density functions of the utility indices, and partly to the difference in shape between the two simultaneous density functions.

#### 9.4.7.3 Conclusions

We cannot definitely state which model is the best without specifying, among other things, the probability of various combinations of the utility indices. These probabilities will of course always be specific to the particular problem concerned, but for any problem the procedure outlined above can be used to evaluate the appropriateness of the models available. This procedure was illustrated by two simple numerical examples.

In the first example, we would have preferred the additive model for parameter values as small as  $\alpha_1 = 0.30$  (fixed) and  $\alpha_2 = 0.10$  (varied), if we had decided to select the model most likely to have the lowest MSE. In the second example, where we used a less extreme density function, we would still prefer the additive model for many parameter values (for values around  $\alpha_1 = 0.30$  and  $\alpha_2 = 0.57$ ). Had there been no uncertainty, we would have chosen the additive model only when the parameters added up to one, i.e. when  $\alpha_1 = 0.30$  and  $\alpha_2 = 0.70$  in our two examples.

Our general conclusion is that the additive model is preferred in many more situations than had been expected, when there is uncertainty in the estimates.

So far we have only looked at the case of a two-criteria problem, but I expect the additive model to outclass the interaction model even more clearly as the number of criteria increases. I base this claim on the fact that the number of interaction terms contributing to the variance of the interaction model will grow exponentially as the number of criteria increases. It would be interesting to investigate this hypothesis, while also trying to see how our belief in the superiority of the additive model is affected by other correct models, apart from the one used in the present study. However, this could be a rewarding line for future inquiry.

## 9.5 SUMMARY AND CONCLUSION

The traditional way of discovering a decision-maker's preferences for different criteria has been to use some indirect method, such as regression analysis. During recent years, however, certain direct methods have been receiving considerable attention. These two approaches differ, among other things, in their treatment of uncertainty in the decision-maker's estimates. Psychological studies have shown that people's estimates of various phenomena are affected by (random) errors, but the indirect methods are the only ones that explicitly take this into consideration. In this chapter we have studied the effect that these errors would have on our selection of a utility model when a direct estimation method is being used.

We assumed an additive model with interaction terms to be correct. In the case of a two-criteria problem we then investigated which of two models - one with and one without interaction terms - had the highest precision, given the presence of random errors in the decision-maker's estimates. We took the mean square error as our criterion of precision.

We found the analytical approach to be fruitless, and so we turned to simulations. For given values of the criteria we found that in many cases the simple additive model was superior to the additive model with interaction terms. The superiority of the simple additive model increased as the mean of either of the utility indices began to deviate from 0.50, and decreased as the sum of the two weights began to deviate from one. Naturally the superiority of this model also increased if the estimation error in the weights increased.

It is only possible to consider the values of the criteria in specific problems where the distribution of these criteria values can be estimated. Two examples with different distributions were analyzed and they indicated that the simple additive model would be preferred in many more situations than had been expected, when there was uncertainty in the estimates of the unidimensional utility functions and the parameters.



## CHAPTER 10

SUMMARY AND FUTURE RESEARCH10.1 SUMMARY

The aim of this study has been to analyze some suggested approaches to the solution of multiple-criteria problems. The approaches that have interested us here, involve the determination of a model of the decision-maker's preferences and the use of this model to evaluate different multiple-criteria objects or alternatives. To estimate the components of the model we have to choose a suitable method.

I decided to concentrate on direct methods, i.e. methods that derive the decision-maker's preferences from direct interrogation, since these methods have been the object of considerable attention in management literature during the last few years. Several reports have appeared on the implementation of various direct methods, but there have been no comparative studies. I therefore decided to carry out an experiment to discover some of the characteristics of different direct methods and to find out something about the models they produce.

The dominating criterion for a "good" model or method is its precision, which is often defined as the degree of conformity with the subject's intuitive evaluation. I did not want to restrict myself to this criterion alone, since several other attributes, such as the decision-maker's confidence in the results or the ease with which the models and methods can be used, are very important when applying multiple-criteria methods. I therefore also investigated aspects of this kind.

As one of the traditional approaches to the determination of a decision-maker's preferences has been regression analysis, I decided to include this in my study. Regression analysis is an indirect

method, because it deduces preferences from the behavior of the decision-makers. I used this method to estimate six commonly used models, and then compared the precision of these models with that of the tested direct models.

I decided to test the methods and their models on a real-life economic problem. The introduction of the new Planning, Programming, and Budgeting System in the Swedish Air Force called for annual evaluations of the Air Force squadrons, and I chose this as my problem in view of several desirable properties it possessed: for instance the problem had already been structured by the Air Staff in a way that gave me an opportunity to test the methods and models on several different subproblems of varying complexity. The Air Force problem also provided me with many subjects who could take part in the experiment, thereby increasing my confidence in the results. I do not consider this problem to be unique to the defense sector. It could just as well have concerned the evaluation of a computer department, for example.

I selected four direct methods as the maximum number of methods I could test. Two of these were chosen by the Air Staff. These had been developed within the defense organization (one by a group of consultants and one by Tell) but were unknown to the participants in the experiment. The remaining methods (one by Keeney and one by Miller) were chosen from the literature of the subject.

After analyzing the methods and their models, I deduced some hypotheses about the expected results of the experiment (Section 3.7). These are shown on the left in Table 10.1, while the actual findings of the study are noted on the right.

The first attribute to evaluate is the precision of the models. Precision is defined as the degree of concordance between the utility indices produced by the model and those given by the decision-maker. We found that the indirect models in the experiment had a



Attribute	Ranking of methods									
	Expected					Results				
Precision	R	C	K	M	T	R	T	(MC)	K	
Believed precision		T	M	C	K					
- during construction						T	M	C	K	
- in use						(TM)	C	K		
<u>During construction</u>										
Ease of understanding the questions		(CT)		M	K	T	(MC)		K	
Ease of answering the questions		(MT)		(CK)		T	(MC)		K	
Time requirement		(MT)		K	C	T	M	C	K	
<u>In use</u>										
Ease of understanding the instructions		(MT)		K	C	(MT)		K	C	
Ease of performing the instructions		(MT)		K	C	(MT)		K	C	
Time requirement		(MT)		K	C	(MT)		K	C	

(C = consultants', K = Keeney's, M = Miller's, R = regression, and T = Tell's methods respectively.)

Table 10.1 A summary of expected and real outcomes on the attributes used to evaluate the studied multiple-criteria models and methods. The best outcomes to the left.

precision matching that of the direct models when we took the correlation coefficients or the variance as our criterion. Because the direct models exhibited greater bias, the indirect models revealed greater precision than the direct models when we took the mean square error as our criterion.

Of the six indirect models, the linear models proved to have the greatest precision and the disjunctive model the least, irrespective of which criterion we used except in the case of the bias.

The two linear models proved equally good. The conjunctive, exponential, and logarithmic models lay in between the others and could not be separated from one another. This is a result in accordance with that of other studies.

Of the direct models, I had expected the two additive models (Miller's and Tell's) to have a great appeal, although their precision and the decision-makers' belief in their precision could be expected to be lower than that of the two additive models with interaction terms. However, the experiment revealed a unanimous preference for pure additive models. Tell's method, which had been developed in the organization but was unknown to the subjects participating in the experiment, came out best or tied with the other additive model (Miller's) on all criteria. Keeney's method was the least preferred method, although his model was found easier to use than that of the consultants.

If we take all the attributes into consideration, it seems clear that pure additive models are preferred to more complex models when they are derived by direct methods. It also seems evident that the "practical" method is preferred to the "theoretical" method in the class of pure additive models, and in the class of additive models with interaction terms.

This experiment showed clearly that simple models - such as the additive - had a higher precision than more complex models and that these models and their estimation methods received high rankings on all other attributes. However, we must bear in mind that this was a static experiment where it was not possible to study, for example, any effects of learning. It is not impossible that a greater element of teaching or more experience of using the methods or models over a longer period, may change the rankings on some criteria.

The experiment also confirmed a hypothesis that has been discussed in other studies, namely that an increase in the number of criteria lowers the precision of the model.

In Section 3.3.3 we put forward the hypothesis that the von Neumann-Morgenstern lottery-technique is hard to understand and to use. Our experiment seems to have generated a good deal of material to support this hypothesis. We found (a) that the method required substantially more time than the other methods. This is also expressed by the subjects' attitudes as measured by (b), the ease of understanding the questions posed by the method, and by (c), the ease of answering these questions. The impression that the method is difficult is reinforced by (d), the demand for an explanation of this method during the experiment, (e), the many negative comments given by the participants and (f), the great variance in the outcomes of the lottery-technique.

The unexpectedly good results of the pure additive models requires further investigation. One cause could be the existence of uncertainty in the preferences expressed by the decision-makers. Several psychological studies have shown that people's estimates of various phenomena are affected by errors. The indirect methods take explicit account of the uncertainty in the data provided by the decision-maker and try to minimize its influence on the model, while the direct methods do not take uncertainty in the estimates into consideration.

I decided to investigate the effect that estimation errors would have on the selection of a multiple-criteria model. I restricted my study to a two-criteria problem, in which I assumed an additive model with interaction terms to be the correct model. In an estimation of this model, there would be errors in the unidimensional utility functions and in the weights. Using a simulation approach, we were able to study a pure additive model and an additive model with interaction terms, to see which of them would achieve the greatest precision. Precision was measured by the mean square error.

For given values of the criteria we found the Purely additive model to be superior in many cases to the additive model with interaction

terms. This superiority increased when the mean of the utility indices began to deviate from 0.50, but it decreased when the sum of the two criteria began to deviate from one, and when the estimation error in the weights fell. When we incorporated the values of the criteria into the analysis, we still found the simple additive model to be preferred in more situations than had been expected.

The results are only tentative, as we have analyzed only one problem, i.e. one particular correct compound utility function of only two criteria. The results of these simulations indicate that in a great many situations where there are errors in the estimates and where a more complex model than the pure additive model would normally be chosen, a model with greater precision would result if we chose the purely additive model instead.

To sum up the experiment and the simulation study: both indicate that simple models, such as the linear and the additive models, seem to attract users and to be as accurate as the more complex models.

## 10.2 FUTURE RESEARCH

In the area of multi-criteria evaluations much remains to be done. During the last few years many new approaches and methods have been presented to help decision-makers to solve multiple-criteria problems. At present we know relatively little about how these methods work, as the number of reported applications, experiments, etc. is fairly low. In my opinion, this is the type of study that is needed now, to increase our knowledge of existing methods and to provide a basis for the development of new methods.

In the present experiment the comparison of direct and indirect methods was not of primary importance. The indirect models were derived from available data, and it was thus not possible to obtain information

about the decision-makers' attitudes to the two techniques. It would be very interesting, however, to compare these two approaches in more detail and particularly to focus on the difference between them. Such a study could be designed so as to tell us something about the trade-offs between, for example, precision (one of the advantages of the indirect methods) and time-requirements (an advantage that is claimed for the direct methods).

The present study also leaves some minor but very important issues for future research. Thus all the comments made in connection with the presentation of the four direct methods contain several testable hypotheses. For example, there was the question raised in connection with Miller's method, as to whether paired comparisons will produce weights different from those produced by methods where the decision-maker maintains his overall view while assigning his weights; and the question that appeared in the review of Keeney's method, as to whether the use of non-extreme criteria values would produce better estimates of unidimensional utility functions and weights than Keeney's use of extreme values; and the question about the significance of the two alternative simplifications introduced in Tell's method.

It would be extremely valuable to broaden the perspective from multiple-criteria control to include other decision-making situations as well, and to test the direct and indirect approaches with interactive techniques, for example. Such a test would automatically emphasize other criteria, besides the accuracy of the methods, since they attack and solve the problems in such different ways. This would provide a natural follow-up to the present study of direct and indirect utility models and to the comparative studies of interactive methods of Dyer (1973) and Wallenius (1975).

Many comparative studies of multiple-criteria methods have been made as laboratory experiments, but in the present study I have based my experiment on a real-life problem. The problem was taken from the Swedish defense organization, and its connection with economic decisions was that the utility indices generated were going to be used together with

cost estimates and other information in the budgeting process. A natural continuation would be to study methods in real-life multiple-criteria problems in business firms, e.g. in sales-allocation problems, investment decisions, and production planning, since such studies would be directed more towards the implementation of multiple-criteria decision-making methods.

As regards the simulations, these appear to contain the seeds of many big research projects. If human error in making estimates is to be explicitly allowed for, then we must first of all find out more about the kind of errors people tend to make, and how these errors will affect the various types of model.

I personally intend to continue investigating what effect uncertainty in the estimates has on the selection of a multiple-criteria model, and among the aspects I hope to study further are:

- (a) what effect would various distributions of the criterion values have on the conclusions?
- (b) what effect would other estimation methods have on the conclusions?
- (c) what effect would other correct models have on the conclusions?
- (d) what effect would an increase in the number of criteria have on the conclusions?

REFERENCES

- ABELSON, R.P., and TUKEY, J.W., 1959, Efficient Conversion of Non-metric Information into Metric Information. Proceedings of the Social Statistics Section, American Statistical Association, 226-30
- ABELSON, R.P., and TUKEY, J.W., 1963, Efficient Utilization of Non-numerical Information in Quantitative Analysis: General Theory and the Case of Simple Order. The Annals of Mathematical Statistics, 34, 1347-69
- ACKOFF, R.L., 1962, Scientific Method. Wiley, New York
- ACKOFF, R.L., 1967, Management Misinformation Systems. Management Science, 14, No. 4, B147-56
- ACKOFF, R.L., 1970, A Concept of Corporate Planning. Wiley, New York
- AHLBERG, G., 1972, Beslutsfattande - teori och praktik. EFI, Stockholm School of Economics, Stockholm
- ARCHER, E.J., BOURNE, L.E., and BROWN, F.G., 1955, Concept Identification as a Function of Irrelevant Information and Instructions. Journal of Experimental Psychology, 49, 153-64
- ASHTON, R.H., 1974, The Utilization and Expert Judgements: A Comparison of Independent Auditors with Other Judges. Journal of Applied Psychology, 59, No. 4, 437-44
- ANDERSSON, Å., 1972, En ny metod för kommunal planering. Nordisk Operationsanalytisk Konferens, Göteborg
- AUBIN, J-P., and NÄSLUND, B., 1972, An Exterior Branching Algorithm. European Institute for Advanced Studies in Management, Working Paper 72-42, Brussels
- BAKER, B.O., HARDYCK, C.F., and PETRINOVICH, L.F., 1966, Weak Measurements vs. Strong Statistics; An Empirical Critique of S.S. Stevens' Proscriptions on Statistics. Educational and Psychological Measurement, 26, No. 2, 291-309
- BAUMOL, W.J., 1959, Business Behavior, Value and Growth. MacMillan, New York
- BECKER, G.M., DeGROOT, G.M., and MARSCHAK, J., 1964, Measuring Utility by a Single-Response Sequential Method. Behavioral Science, 9, 226-32
- BECKER, G.M., and McCLINTOCK, C.G., 1967, Value: Behavioral Decision Theory. Annual Review of Psychology, 18, 239-86

- BELLMAN, R.E., and ZADEH, L.A., 1970, Decision-Making in a Fuzzy Environment, Management Science, 17, No. 4, B141-64
- BIRNBAUM, M.H., 1973, The Devil Rides Again: Correlation as an Index of Fit. Psychological Bulletin, 79, No. 4, 239-42
- BIRNBAUM, M.H., 1974, Reply to the Devil's Advocates: Don't Confound Model Testing and Measurement. Psychological Bulletin, 81, No. 11, 854-59
- BLAU, P.M., 1956, Bureaucracy in Modern Society. Random House, New York
- BLOCK, H.D., and MARSCHAK, J., 1960, Random Orderings and Stochastic Theories of Responses. In Olkin, I. and others (eds.), Contributions to Probability and Statistics. Stanford University Press, Stanford, Calif.
- BOGGESS, W.P., 1967, Screen-Test Your Credit Risk. Harvard Business Review, Nov-Dec, 113-22
- BONEAU, C.A., 1960, The Effect of Violations of Assumptions Underlying the t Test. Psychological Bulletin, 57, No. 1, 49-64
- CARO, F.G., 1971, Evaluation Research: An Overview. In Caro, F.G. (ed.), Readings in Evaluation Research. Russell Sage Foundation, New York
- CHERNOFF, H., 1960, A Compromise Between Bias and Variance in the Use of Nonrepresentative Samples. In Olkin, I. and others (eds.), Contributions to Probability and Statistics. Stanford University Press, Stanford, Calif.
- CHURCHMAN, C.W., and ACKOFF, R.L., 1954, An Approximate Measure of Value. Operations Research, 2, 172-80
- CONOVER, W.J., 1971, Practical Nonparametric Statistics. Wiley, New York
- COOLEY, W.W., and LOHNES, P.R., 1971, Multivariate Data Analysis. Wiley, New York
- COOMBS, C.H., 1964, A Theory of Data. Wiley, New York
- COOMBS, C.H., and RAO, R.C., 1955, Nonmetric Factor Analysis. Engineering Research Bulletin No. 38, Engineering Research Institute, University of Michigan, Ann Arbor, Mich.
- CYERT, R.M., and MARCH, J.G., 1963, A Behavioral Theory of the Firm. Prentice Hall, Englewood Cliffs, N.J.
- DAVIDSON, D., and MARSCHAK, J., 1959, Experimental Tests of a Stochastic Decision Theory. In Churchman, C.W., and Ratoosh (eds.), Measurement. Definitions and Theories. Wiley, New York



- DAWES, R.M., 1971, A Case Study of Graduate Admissions: Application of Three Principles of Human Decision Making. American Psychologist, 26, 180-88
- DAWES, R.M., 1973, Objective Optimization under Multiple Subjective Functions. In Cochrane, J.L., and Zeleny, M. (eds.), Multiple Criteria Decision Making. University of South Carolina Press, Columbia, S.C.
- DAWES, R.M., and CORRIGAN, B., 1974, Linear Models in Decision Making. Psychological Bulletin, 81, No. 2, 95-106
- DEAN, J., 1956, Managerial Economics. Prentice-Hall, Englewood Cliffs, N.J.
- DeNEUFVILLE, R., and KEENEY, R.L., 1972, Use of Decision Analysis in Airport Development for Mexico City. In Drake, A.W., Keeney, R.L., and Morse, P.M. (eds.), Analysis of Public Systems. MIT Press, Cambridge, Mass.
- DeSOTO, C.B., 1961, The Predilection for Single Orderings. Journal of Abnormal Social Psychology, 62, 16-23
- DIGGORY, T.J., 1963, The Installation of Budgetary Controls in Small Companies. Cost and Management, 37, No. 9. 393-401
- DRAPER, N.R., and SMITH, H., 1966, Applied Regression Analysis. Wiley, New York
- DRUCKER, P.F., 1954, The Practice of Management. Harper and Brothers, New York
- DYER, J.S., 1973, An Empirical Investigation of a Man-Machine Interactive Approach to the Solution of the Multiple Criteria Problem. In Cochrane, J.L., and Zeleny, M. (eds.), Multiple Criteria Decision Making. University of South Carolina Press, Columbia, S.C.
- DYER, J.S., 1974, The Effects of Errors in the Estimation of the Gradient on the Frank-Wolfe Algorithm, with Implications for Interactive Programming. Operations Research, 22, No. 1, 160-74
- DYER, J.S., FARRELL, W., and BRADLEY, P., 1973, Utility Functions for Test Performance. Management Science, 20, No. 4, Part I, 507-19
- EASTON, A., 1965, Advertising Decision Involving Multiple Criteria. Proceedings from the 11th Annual Conference, October 5, 1965, Advertising Research Foundation, New York
- EASTON, A., 1966, A Forward Step in Performance Evaluation, Journal of Marketing, 30, 26-32

- EASTON, A., 1973, Complex Managerial Decisions Involving Multiple Objectives. Wiley, New York
- ECKENRODE, R.T., 1965, Weighting Multiple Criteria. Management Science, 12, No. 3, 180-92
- EDWARDS, W., 1954, The Theory of Decision Making. Psychological Bulletin, 51, No. 4, 380-417
- EILON, S., 1972, Goals and Constraints in Decision Making. Operational Research Quarterly, 23, 8-15
- EINHORN, H.J., 1970, The Use of Nonlinear, Noncompensatory Models in Decision Making. Psychological Bulletin, 73, No. 3, 221-30
- EINHORN, H.J., 1971, Use of Nonlinear, Noncompensatory Models as a Function of Task and Amount of Information. Organizational Behavior and Human Performance, 6, 1-27
- EINHORN, H.J., 1972, Expert Measurement and Mechanical Combination. Organizational Behavior and Human Performance, 7, 86-106
- FABRYCKY, W.J., and THUESEN, G.J., 1974, Economic Decision Analysis. Prentice-Hall, Englewood Cliffs, N.J.
- FEINBERG, A., 1972, An Experimental Investigation of an Interactive Approach for Multi-Criterion Optimization with an Application to Academic Resource Allocation. Ph.D. dissertation, University of California, Los Angeles, Calif.
- FIELD, J.E., 1969, Toward a Multi-level, Multi-goal Information System. Accounting Review, 44, No. 3, 593-99
- FISHBURN, P.C., 1967, Methods of Estimating Additive Utilities. Management Science, 13, No. 7, 435-53
- Försvarets Materielverk, 1974 a och b, Särskild inspektion av underhållsfunktionen 1972-1974, Bilaga B och C. FMV-A:P H A78:2/74 och A79:2/74
- FRANKFURTER, G.M., PHILLIPS, H.E., and SEAGLE, J.P., 1971, Portfolio Selection: The Effects of Uncertain Means, Variances, and Covariances. Journal of Financial and Quantitative Analysis, 4, Dec., 1251-62
- FRENCKNER, T.P., 1953, Syfta företagen mot högsta möjliga vinst. FFI at the Stockholm School of Economics, Stockholm
- GLENNING, C., 1975, Sluta säger Riksrevisionsverket. Vi Bilägare, nr 16, 10-11

- GOLDBERG, L.R., 1968, Simple Models or Simple Processes? Some Research on Clinical Judgements. American Psychologist, 23, 483-96 .
- GOLDBERG, L.R., 1971, Five Models of Clinical Judgements: An Empirical Comparison between Linear and Nonlinear Representations of the Human Inference Process. Organizational Behavior and Human Performance, 6, 458-79
- GOODMAN, L.A., 1960, The Exact Variance of Products. American Statistical Association Journal, December, 708-13
- GRANICK, D., 1954, Management of the Industrial Firm in the USSR. Columbia University Press, New York
- GRAYSON, C.J. Jr., 1960, Decisions under Uncertainty. Drilling Decisions by Oil and Gas Operators. Harvard Business School, Boston, Mass.
- GREEN, B.F. Jr., 1968, Descriptions and Explanations: A Comment on Papers by Hoffman and Edwards. In Kleinmuntz, B. (ed.), Formal Representation of Human Judgement. Wiley, New York
- GREEN, P.E., and CARMONE, F.J., 1970, Multidimensional Scaling and Related Techniques in Marketing Analysis. Allyn and Bacon, Boston, Mass.
- GREEN, P.E., and CARMONE, F.J., 1974, Evaluation of Multiple-attribute Alternatives: Additive versus Configural Utility Measurement. Decision Sciences, 5, No. 2, 164-81
- GREEN, P.E., and WIND, Y., 1973, Multiattribute Decisions in Marketing: A Measurement Approach. Dryden Press, Hinsdale, Ill.
- GUILFORD, J.P., 1954, Psychometric Methods. McGraw-Hill, New York
- HAMMOND III, J.S., 1967, Better Decisions with Preference Theory. Harvard Business Review, Nov-Dec, 123-41
- HAYES, J.R., 1962, Human Data Processing Limits in Decision Making. Electronic Systems Division Report ESD-TDR-62-48, July, cited in Shepard (1964)
- HEISER, H.C., 1959, Budgeting. Principles and Practice. Ronald Press, New York
- HERTZ, D.B., 1964, Risk Analysis in Capital Investment. Harvard Business Review, Jan-Feb, 95-106
- HILL, R.E., 1974, An Empirical Comparison of Two Models for Predicting Preferences for Standard Employment Offers. Decision Sciences, 5, No. 2, 243-54

- HILLIER, F.S., 1963, The Derivation of Probabilistic Information for the Evaluation of Risky Investments. Management Science, 9, 443-57
- HITCH, C.J., and McKEAN, R.N., 1960, The Economics of Defense in the Nuclear Age. Harvard University Press, Cambridge, Mass.
- HOEPFL, R.T., and HUBER, G.P., 1970, A Study of Self-Explicated Utility Models. Behavioral Science, 15, 408-14
- HOERL, A.E., and KENNARD, R.W., 1970, Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12, No. 1, 55-67
- HOFFMAN, P.J., 1960, The Paramorphic Representation of Clinical Judgement. Psychological Bulletin, 57, No. 2, 116-31
- HOFFMAN, P.J., 1968, Cue-Consistency and Configurality in Human Judgement. In Kleinmuntz, B. (ed.), Formal Representation of Human Judgement. Wiley, New York
- HOFFMAN, P.J., SLOVIC, P., and RORER, L.G., 1968, An Analysis-of-Variance Model for the Assessment of Configurality Cue Utilization in Clinical Judgement. Psychological Bulletin, 69, 338-49
- HUBER, G.P., 1968, Multiplicative Utility Models in Cost Effectiveness Analysis. Journal of Industrial Engineering, 19, No. 3, XVII-XIX
- HUBER, G.P., 1974, Multi-attribute Utility Models: A Review of Field and Field-Like Studies. Management Science, 20, No. 10, 1393-402
- HUBER, G.P., DANESHGAR, R., and FORD, D.L., 1971, An Empirical Comparison of Five Utility Models for Predicting Job Preferences. Organizational Behavior and Human Performance, 6, 267-82
- HUBER, G.P., SAHNEY, V.K., and FORD, D.L., 1969, A Study of Subjective Evaluation Models. Behavioral Science, 14, 483-9
- HUMPHREYS, P., 1975, Applications of Multiattribute Utility Theory. Paper presented to 5th Research Conference on Subjective Probability, Utility, and Decision Making. Darmstadt
- JOHNSEN, E., 1968, Studies in Multiobjective Decision Models. Studentlitteratur, Lund
- KALYMON, B.A., 1971, Estimation Risk in the Portfolio Selection Model. Journal of Financial and Quantitative Analysis, 4, Jan, 559-82

- KEENEY, R.L., 1969, Multidimensional Utility Functions: Theory, Assessment, and Application. Technical Report 43, Operations Research Center, MIT, Cambridge, Mass.
- KEENEY, R.L., 1972, An Illustrated Procedure for Assessing Multi-attributed Utility Functions. Sloan Management Review, Fall, 37-50
- KEENEY, R.L., 1973, A Decision Analysis with Multiple Objectives: The Mexico City Airport. The Bell Journal of Economics and Management Science, 4, No. 1, 101-17
- KEENEY, R.L., 1975, Examining Corporate Policy Using Multiattribute Utility Analysis. International Institute for Applied Systems Analysis, RM-75-36, Schloss Laxenberg
- KEENEY, R.L., and RAIFFA, H., 1972, A Critique of Formal Analysis in Public Decision Making. In Drake, A.W., Keeney, R.L., and Morse, P.M. (eds.), Analysis of Public Systems. MIT Press, Cambridge, Mass.
- KORT, F., 1968, A Nonlinear Model for the Analysis of Judicial Decisions. American Political Science Review, 62, 546-55
- LORD, F.M., 1962, Cutting Scores and Errors of Measurement. Psychometrica, 27, No. 1, 19-30
- LUCE, R. D., and SUPPES, P., 1965, Preference, Utility, and Subjective Probability. In Luce, R.D., Bush, R.R., and Galanter, E. (eds.), Handbook of Mathematical Psychology, Vol. III, Chapter 19, 249-410. Wiley, New York
- MacCRIMMON, K.R., 1968, Decisionmaking among Multiple-Attribute Alternatives: A Survey and Consolidated Approach. RAND, RM-4823-ARPA, Santa Monica, Calif.
- MacCRIMMON, K.R., 1973, An Overview of Multiple Objective Decision Making. In Cochrane, J.L., and Zeleny, M. (eds.), Multiple Criteria Decision Making. University of South Carolina Press, Columbia, S.C.
- MacCRIMMON, K.R., 1974, Managerial Decision-Making. In McGuire, J.W. (ed.), Contemporary Management. Issues and Viewpoints. Prentice-Hall, Englewood Cliffs, N.J.
- MacCRIMMON, K.R., and SIU, J.K., 1974, Making Trade-Offs. Decision Sciences, 5, No. 4, 680-704
- MacCRIMMON, K.R., and TODA, M., 1969, The Experimental Determination of Indifference Curves. Review of Economic Studies, 36, No. 4, 433-50

- McGUIRE, J.W., 1964, Theories of Business Behavior. Prentice-Hall, Englewood Cliffs, N.J.
- McKEAN, R.N., 1958, Efficiency in Government through Systems Analysis. Wiley, New York
- McKENNEL, A., 1970, Attitude Measurement: Use of Coefficient Alpha with Cluster of Factor Analysis. Sociology, 4, 227-45
- MARCH, J.G., and SIMON, H.A., 1958, Organizations. Wiley, New York
- MARKOWITZ, H., 1952, Portfolio Selection. The Journal of Finance, 7, No. 1, 77-91
- MARQUARDT, D.W., 1970, Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. Technometrics, 12, No. 3, 591-612
- MARQUIS, D.G., 1968, Individual and Group Decisions Involving Risk, Industrial Management Review, 9, No. 3, 69-75
- MARSCHAK, J., 1955, Elements for a Theory of Teams. Management Science, 1, No. 2, 127-37
- MARSCHAK, J., and RADNER, R., 1972, Economic Theory of Teams. Yale University Press, New Haven, Conn.
- MEEHL, P.E., 1954, Clinical versus Statistical Prediction. University of Minnesota Press, Minneapolis, Minn.
- MERTZ, W.H., and DOHERTY, M.E., 1974, The Influence of Task Characteristics on Strategies of Cue Combination. Organizational Behavior and Human Performance, 12, 196-216
- MILLER, D.W., and STARR, M.K., 1960, Executive Decisions and Operations Research. Prentice-Hall, Englewood Cliffs, N.J.
- MILLER, D.W., and STARR, M.K., 1967, The Structure of Human Decisions. Prentice-Hall, Englewood Cliffs, N.J.
- MILLER, G.A., 1956, The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. Psychological Review, 63, No. 2, 81-97
- MILLER, J.R., III, 1966, The Assessment of Worth: A Systematic Procedure and its Experimental Validation. Ph.D. dissertation, MIT, Cambridge, Mass.
- MILLER, J.R., III, 1970, Professional Decision-Making; A Procedure for Evaluating Complex Alternatives. Praeger Publishers, New York

- MOORE, J.R., Jr., and BAKER, N.R., 1969, Computational Analysis of Scoring Models for R and D Project Selection. Management Science, 16, No. 4, B212-32
- MORRISON, D.F., 1967, Multivariate Statistical Methods. McGraw-Hill, New York
- NADLER, G., HUBER, G.P., and SAHNEY, V.K., 1967, A Study of Measuring the Quality of Patient Care. In Innovation and Enterprise, the 18th Annual Institute Conference and Convention of the AIIE, New York
- NÄSLUND, B., 1974, Interactive Methods in Multiple Criteria Optimization. EFI Working Paper No. 6004, Stockholm School of Economics, Stockholm
- NÄSLUND, B., 1975, The Importance of Goal Analysis in Strategic Planning. EFI Working Paper No. 6025, Stockholm School of Economics, Stockholm
- NUNNALLY, J.C., 1967, Psychometric Theory. McGraw-Hill, New York
- NYSTEDT, L., 1975, Personbedömning och klinisk bedömning. In G. Zetterblom (ed.), Psykologin i samhället. Aktuell svensk forskning. AWE/Gebers, Stockholm
- NYSTEDT, L., and MAGNUSSON, D., 1975, Integration of Information in a Clinical Judgment Task, An Empirical Comparison of Six Models. Perceptual and Motor Skills, 40, 343-56
- OSGOOD, C.S., SUCI, G.J., and TANNENBAUM, P.H., 1957, The Measurement of Meaning. University of Illinois Press, Urbana, Ill.
- PFLOMM, N.E., 1963, Managing Capital Expenditure. Studies in Business Policy no. 107, National Industrial Conference Board, Inc., New York
- PPBG, 1974, Skrivelse från Ledningsgruppen för fortsatt utveckling av försvarets planerings- och programbudgetsystem. 1974-10-01, dnr 93, Bihang B
- PRESS, S.J., 1972, Applied Multivariate Analysis. Holt, Rinehart & Winston, New York
- QUADE, E.S., 1968, When Quantitative Models are Inadequate. In Quade, E.S., and Boucher, W.I. (eds.), Systems Analysis and Policy Planning. American Elsevier, New York
- RAIFFA, H., 1968, Decision Analysis; Introductory Lectures on Choices and Uncertainty. Addison-Wesley, Reading, Mass.

- RAIFFA, H., 1969, Preferences for Multi-Attributed Alternatives. RAND RM 5868-DOT/RC, Santa Monica, Calif.
- RAPOPORT, A., 1960, Fights, Games, and Debates. The University of Michigan Press, Ann Arbor, Mich.
- RIDGEWAY, V.F., 1956, Dysfunctional Consequences of Performance Measurements. Administrative Science Quarterly, 1, 240-7
- ROBICHECK, A.A., and MYERS, S.C., 1965, Optimal Financing Decisions. Prentice-Hall, Englewood Cliffs, N.J.
- ROSE, G.L., 1974, Assessing the State of Decision Making. In McGuire, J.W. (ed.), Contemporary Management. Issues and Viewpoints. Prentice-Hall, Englewood Cliffs, N.J.
- ROUSSEAS, S.W., and HART, A.G., 1951, Experimental Verification of a Composite Indifference Map. Journal of Political Economy, 59, 288-318
- ROY, B., 1971, Problems and Methods with Multiple Objective Functions. Mathematical Programming, 1, No. 2, 237-66
- SAMUELSON, P.A., 1947, Foundations of Economic Analysis. Harvard University Press, Cambridge, Mass.
- SAWYER, J., 1966, Measurement and Prediction, Clinical and Statistical. Psychological Bulletin, 66, No. 3, 178-200
- SAYEKI, Y., and VESPER, K.H., 1973, Allocations of Importance in a Hierarchical Goal Structure. Management Science, 19, No. 6, 667-75
- SCHEFFE, H., 1959, The Analysis of Variance. Wiley, New York
- SCHLAIFER, R., 1969, Analysis of Decisions under Uncertainty. McGraw-Hill, New York
- SCHWARTZ, S.L., VERTINSKY, I., ZIEMBA, W.T., and BERNSTEIN, M., 1975, Some Behavioural Aspects of Information Use in Decision Making: A Study of Clinical Judgements. Paper presented at the Multiple Criteria Decision-Making Conference at Jouy-en-Josas, sponsored by the European Institute for Advanced Studies in Management in Brussels
- SEVON, G., 1974, Värderingar och subjektiva sannolikheter. Meddelanden från forskningsinstitutet vid Svenska Handelshögskolan, no. 1, Helsinki
- SHARPE, W.F., 1964, Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. The Journal of Finance, 19, No. 3, 425-42



- SHEPARD, R.N., 1964, On Subjectively Optimum Selections Among Multi-attribute Alternatives. In Shelly, M.W., and Bryan, G.L. (eds.), Human Judgements and Optimality. Wiley, New York
- SIEGEL, S., 1956, Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill, New York
- SIMON, H.A., 1957, Models of Man. Wiley, New York
- SLOVIC, P., 1969, Analyzing the Expert Judge: A Descriptive Study of a Stockbroker's Decision Processes. Journal of Applied Psychology, 53, 255-63
- SLOVIC, P., FLEISSNER, D., and BAUMAN, W.S., 1972, Analyzing the Use of Information in Investment Decision Making: A Methodological Proposal. Journal of Business, 45, 283-301
- SLOVIC, P., and LICHTENSTEIN, S., 1971, Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment. Organizational Behavior and Human Performance, 6, 649-744
- SOELBERG, P.O., 1967, Unprogrammed Decision Making. Industrial Management Review, 8, No. 2, 19-29
- SOU 1968:1, Ekonomisystem för försvaret. Allmänna Förlaget, Stockholm
- SOU 1969:24, Ekonomisystem för försvaret 2. Allmänna Förlaget, Stockholm
- SOU 1969:25, Planering och budgetering inom försvaret. Allmänna Förlaget, Stockholm
- SOUDER, W.E., 1972, A Scoring Methodology for Assessing the Suitability of Management Science Models. Management Science, 18, No. 10, B526-43
- SOUDER, W.E., 1975, Achieving Organizational Consensus with Respect to R&D Project Selection Criteria. Management Science, 21, No. 6, 669-81
- STANLEY, J.K., 1974, A Cardinal Utility Approach for Project Evaluation. Socio-Economic Planning Sciences, 8, 329-38
- STARR, M.K., 1974, Management Science and Management. In McGuire, J.W., (ed.), Contemporary Management. Issues and Viewpoints. Prentice-Hall, Englewood Cliffs, N.J.
- STEUER, R.E., forthcoming, Linear Multiple Objective Programming with Interval Criterion Weights. Management Science.

- STEVENS, S.S., 1968, Measurement, Data Theory, and the Schemapiric View. Science, 161, No. 3894, 849-56
- SUMMERS, D.A., TALIAFERRO, J.D., and RICHGERS, D.J., 1970, Subjective vs Objective Description of Judgment Policy. Psychonomic Science, 18, 249-50
- SWALM, R.O., 1960, Utility Theory - Insights into Risk Taking. Harvard Business Review, Nov-Dec, 123-36
- TELL, B., 1974, Formulär använda vid undersökning av fyra multimålmeter. EFI Working Paper, Stockholm School of Economics, Stockholm
- TELL, B., 1975, An Experimental Study of Four Multiple-Criteria Evaluation Methods. EFI Working Paper No. 6028, Stockholm School of Economics, Stockholm
- TERRY, H., 1963, Comparative Evaluation of Performance Using Multiple Criteria. Management Science, 9, No. 3, 431-41
- THIRIEZ, H., and HOURI, D., 1975, Multi-Person, Multi-Criteria Decision-Making: A Sample Approach. Paper presented at the Multiple Criteria Decision-Making Conference at Jouy-en-Josas, sponsored by the European Institute for Advanced Studies in Management in Brussels
- THURSTONE, L.L., 1931, The Indifference Function. Journal of Social Psychology, 2, 139-67
- Tre Konsulter AB, 1972a, Metoder och modeller för värdering av system - särskilt materialorienterade system inom arméförband. Rapport 2303, 1972-03-13
- Tre Konsulter AB, 1972b, Metodik för utvärdering av kvalitativ tillgänglighet för elektronisk materiel. Rapport 2103, 1972-10-09
- TURBAN, E., and METERSKY, M.L., 1971, Utility Theory Applied to Multivariable System Effectiveness Evaluation. Management Science, 17, No. 12, B817-28
- VONNEUMANN, J., and MORGENTERN, O., 1947, Theory of Games and Economic Behavior, 2nd ed.. Princeton University Press, Princeton, N.J.
- VERTINSKY, I., and WONG, E., 1975, Eliciting Preferences and the Construction of Indifference Maps: A Comparative Empirical Evaluation of Two Measurement Methodologies. Socio-Economic Planning Sciences, 9, 15-24
- WALLENIUS, J., 1975, Interactive Multiple Criteria Decision Making: An Investigation and an Approach. The Helsinki School of Economics, Helsinki

- WALLIS, W.A., and FRIEDMAN, M., 1942, The Empirical Derivation of Indifference Functions. In Lange, O., McIntyre, F., and Yntema, F. (eds.), Studies in Mathematical Economics and Econometrics, In Memory of Henry Schultz. Chicago Illustrated Press, Chicago, Ill.
- WHITE, C.M., 1960, Multiple Goals in the Theory of the Firm. In , Boulding, K.E., and Spivey, W.A., Linear Programming and the Theory of the Firm. MacMillan, New York
- WHITMORE, G.A., and CAVADIAS, G.S., 1974, Experimental Determination of Community Preferences for Water Quality - Cost Alternatives. Decision Sciences, 5, No. 4, 614-31
- WIGGINS, N., 1973, Individual Differences in Human Judgments: A Multi-Variate Approach. In Rapoport, L., and Summers, D.A.(eds.), Human Judgment and Social Interaction. Holt, Rinehart, and Winston, New York
- WILHELM, J., 1975, Objectives and Multi-Objective Decision Making under Uncertainty. Lecture Notes in Economics and Mathematical Systems No. 112. Springer-Verlag, Berlin
- WILLIAMSON, O.E., 1964, The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm. Prentice-Hall, Englewood Cliffs, N.J.
- WONNACOTT, T.H., and WONNACOTT, R.J., 1972, Introductory Statistics, 2nd edition. Wiley, New York
- YNTEMA, D.B., and TORGERSON, W.S., 1961, Man-Computer Cooperation in Decisions Requiring Common Sense. IRE Transactions on Human Factors in Electronics, HFE-2, 20-6
- ZELENY, M., 1973, Compromise Programming. In Cochrane, J.L., and Zeleny, M. (eds.), Multiple Criteria Decision Making. University of South Carolina Press, Columbia, S.C.



LIST OF REPORTS PUBLISHED SINCE 1970 BY THE ECONOMIC RESEARCH INSTITUTE  
AT THE STOCKHOLM SCHOOL OF ECONOMICS

Unless otherwise indicated, these reports are published in Swedish.

	Sw.Crs. (circa)
<u>1976</u>	
BRODIN, B., Product Development Processes - A study of product development in Swedish companies, a marketing approach. Stockholm 1976.	85:- (part I part II)
Part I - Analysis	51:-
Part II - Case Descriptions	51:-
HEDEBRO, G. & NOWAK, K., Developing Countries, Study Groups in Sweden - Three surveys in Gothenburg 1975. Stockholm 1976. <sup>2)</sup>	
Report No. 1: Description of the Participants	25:50
Report No. 2: The Participants and the Work of the Study Group	17:85
Report No. 3: Six Months After the Study Group	15:30
Report No. 4: Analysis and Discussion	15:30
LINDH, L. G., Trade Deals in the Packets Goods Industry. Stockholm 1976.	42:50
RYDEN, I., Transportation Cost and Regional Development - A theoretical analysis of freight subsidies as a means of regional policy. Stockholm 1976.	39:95
STOLT, B., Information Exchange Between Citizens and Public Authorities - A utopia? Stockholm 1976. <sup>2)</sup>	40:80
STOLT, B., Information Exchange Between Citizens and Public Authorities: The descriptive, stipulative and normative framework for information exchange between citizens and public authorities and a classification system for the political resources of the individual Stockholm 1976. <sup>2)</sup>	25:50
TELL, B., A Comparative Study of Some Multiple-Criteria Methods. Stockholm 1976. <sup>1)</sup>	39:95
<u>1975</u>	
ASPLUND, G., Strategy Formulation - An intervention study of a complex group decision process. Stockholm 1975. <sup>1)</sup>	50:-
BIRGEGÅRD, L-E., The Project Selection Process in Developing Countries - A study of the public investment project selection process in Kenya, Zambia and Tanzania. Stockholm 1975. <sup>1)</sup>	45:-

1) Published in English.

2) Mimeographed report.

LIST OF REPORTS PUBLISHED SINCE 1970 (cont.)

	Sw.Crs. (circa)
<u>1975</u>	
DOCHERTY, P. & STJERNBERG, T., Decision Making in Contingencies - Perception and action. Stockholm 1975. <sup>1)</sup>	51:-
FJAESEAD, B. & HOLMLÖV, P. G., Swedish Newsmen's Views on the Role of the Press - Paper presented at the 30th annual conference of the American association for public opinion research, June 1975. Stockholm 1975. <sup>1)2)</sup>	17:85
FORSGÅRDH, L-E. & HERTZEN, K., Information, Expectations and Stock Prices - A study of the Swedish stock market. Stockholm 1975.	56:10
GAVATIN, P., Budget Simulation - Meaning, efficiency and implementation. Stockholm 1975. (In collaboration with Sveriges Mekanförbund.)	51:-
JOHANSSON, S-E., Rate of Return, Growth and Capital Structure of the Firm. Stockholm 1975. (Published by IFL.) <sup>2)</sup>	12:75
JONSSON, E., The Social Cost of Traffic Accidents, Industrial Injuries and Diseases Caused by Tobacco Smoking - A study of medical costs, production losses and other welfare costs. Stockholm 1975.	19:55
JONSSON, E., Health Economy - some articles on current issues. Stockholm 1975. (Off-print from Landstingets Tidskrift, 1968:2, 1968:4, 1974:4, 1974:6 and 1975:2.) <sup>2)</sup>	10:-
JONSSON, E., Put a Price on Working Conditions - Comparison between fees and costs of industrial injuries. Stockholm 1975.	11:50
JULANDER, C-R., Consumers' Saving and Efficiency of Income Use - A behavioural science approach to the study of consumers income use. Stockholm 1975.	46:75
LYBECK, J. A., A Disequilibrium Model of the Swedish Financial Sector. Stockholm 1975. <sup>1)</sup>	61:20
ÅHREN, P., Economic Evaluation Methods in Community Planning. Stockholm 1975.	34:-
<u>1974</u>	
AHLMARK, D., Product-Investment-Financing - A contribution towards a theory of the firm as a product-centered financial system. Tables incl. Stockholm 1974. <sup>2)</sup>	60:-
ASPLUND, G., Uncertainty Factors in the Company and its Environment - A taxonomic study. Stockholm 1974.	45:-
BACK, R., DALBORG, H. & OTTERBECK, L., Industrial Location Patterns - A multidimensional analysis of relationships between firms and regions. Stockholm 1974. <sup>1) 2)</sup>	40:-
BERG, C., The Participation System. A clinical study of a growing company. Stockholm 1974.	40:-
BERGMAN, L. & BERGSTROM, C., Energy Policy and Energy Use. Stockholm 1974. <sup>2)</sup>	17:85
BERGSTRAND, J., Budgetary planning - A summary of two previous books. Stockholm 1974. (In collaboration with Sveriges Mekanförbund) <sup>1)</sup>	10:-
BERGSTRAND, J., Budgetary Planning II - The formulation of budgetary relationships. Stockholm 1974. (In collaboration with Sveriges Mekanförbund.)	30:-

1) Published in English.

2) Mimeographed report.

LIST OF REPORTS PUBLISHED SINCE 1970 (cont.)

	Sw.Crs. (circa)
<u>1974</u>	
The Motor Insurance Department - A report from an experiment with autonomous working groups in a Swedish Insurance Company. 2nd ed. Stockholm 1974. <sup>2)</sup>	17:85
BLOMQUIST, S., Taxation of Different Forms of Household Savings. Stockholm 1974. <sup>2)</sup>	30:-
DALBORG, H., Research and Development - Organization and location. Stockholm 1974. <sup>1)</sup>	30:60
HEDLUND, G. & OTTERBECK, L., The Multinational Corporation, the Nation State and the Trade Unions. Stockholm 1974. (Published by P A Norstedt & Söner.)	44:20
HÄGG, C., Periodic Payment Variations - A study in time series analyses of the firm. Stockholm 1974.	34:-
LIND, R., The Effect on Tax Revenues of Changes in Collective Traffic Fares - A quantification of the effects on the tax revenues of the state, the county councils and the primary municipalities of changes in the fares for local traffic in greater Stockholm. Stockholm 1974. <sup>2)</sup>	25:50
MAGNUSSON, Å., Budgetary Control - Analysis of budget deviations. Stockholm 1974. (In collaboration with Sveriges Mekanförbund.)	45:-
PAULANDER, H., Distributed Lags. Stockholm 1974. <sup>1)</sup> 2)	17:85
SJÖSTRAND, S-E., Company Organisation in the Swedish Building Industry. Stockholm 1974.	15:30
SWAHN, H., The Railway in Frykdalen. Stockholm 1974. <sup>2)</sup>	34:-
THORSLUND, S., Experiences in Information on Traffic via Mass Media - Problems for road users' and pedestrians before, during and after the change-over to right-hand driving as an object for information activities. Stockholm 1974. <sup>2)</sup>	17:-
VALDELIN, J., Product Development and Marketing - An investigation of product development processes in Swedish companies. Stockholm 1974.	35:70
de VYLDER, S., Chile 1970-73 - The political economy of the rise and fall of the unidad popular. Stockholm 1974. (Published by Unga Filosofers Förlag) <sup>1)</sup>	17:-
<u>1973</u>	
AHNSTRÖM, L., Governing and Administrative Activities in Western Europe - A study of their economic-geography. Stockholm 1973. (In collaboration with Almqvist & Wiksell)	51:-
ARVIDSSON, G., Internal Transfer Negotiations - Eight experiments. Stockholm 1973. <sup>1)</sup>	34:-
BACK, R., Coordination of Decisions. Stockholm 1973.	25:50
BACK, R., DALBORG, H. & OTTERBECK, L., A Summary of the Research Program: Location and Development of Economic Structures. Stockholm 1973.	5:10
BERGSTRAND, J., Budgetary Planning - Methods, cases and descriptive models. Stockholm 1973. (In collaboration with Sveriges Mekanförbund.)	35:-

1) Published in English.

2) Mimeographed report.

LIST OF REPORTS PUBLISHED SINCE 1970 (cont.)

	Sw.Crs. (circa)
<u>1973</u>	
FJAEASTAD, B., A Supplement to Mass Media and Business. Stockholm 1973.	8:50
GAVATIN, P., MAGNUSSON, Å. & SAMUELSON, L., Budgeting and Planning in the USA - A travel report. Stockholm 1973. (In collaboration with Sveriges Mekanförbund.)	40:-
HAMMARKVIST, K-O., Adoption of New Products on the Building Market. Stockholm 1973. (In collaboration with Industrins Byggmaterialgrupp.) <sup>2)</sup>	25:50
JONSSON, E., The Costs and Revenues in Local Government 1973 - 1977. Stockholm 1973. (Off-print from SOU 1973:21) <sup>2)</sup>	10:-
JONSSON, E., Local Government Economy - some articles on urgent questions. Stockholm 1973. (Off-print from Landstingets Tidskrift, 1972:9, 1972:13/14, 1973:1 and Kommunal Tidskrift, 1972/20.) <sup>2)</sup>	10:-
JONSSON, E., Local Government Finances - five articles with some results and conclusions. Stockholm 1973. (Off-print from Kommunal Tidskrift, 1972, No. 15, 16, 17, 18 and 20.) <sup>2)</sup>	10:-
LUNDBERG, D., Mass Media Credibility and the Use of Information - An experiment concerning the importance of the function of information for credibility effects. Stockholm 1973. <sup>2)</sup>	20:-
NORSTRÖM, G., A Study of the Transport Geography of Sweden's Foreign Trade. Stockholm 1973.	55:25
OTTERBECK, L., Location and Strategic Planning - Towards a contingency theory of industrial location. Stockholm 1973. <sup>1)</sup>	22:10
SAMUELSON, L., Efficient Budgeting - An analysis of decisions of budgeting design. Stockholm 1973. (In collaboration with Sveriges Mekanförbund.)	45:-
SJÖSTRAND, S-E., Company Organization - A taxonomic approach. Stockholm 1973.	51:-
WIBBLE, A., Selective Economic Policy. Stockholm 1973. <sup>2)</sup>	34:-
ÖSTMAN, L., The Development of Accounting Report - An imperical study with special emphasis and the views of recipients, and on the planning and follow-up of report systems in companies with computer-based accounting. Stockholm 1973.	38:25
<u>1972</u>	
AHLBERG, G., Decision-Making - Theory and practice. Stockholm 1972. <sup>2)</sup>	21:25
BERTMAR, L., Effects of the Investment Tax of 1970 - An empirical study. Stockholm 1972. <sup>2)</sup>	30:-
Budgeting Practice. Stockholm 1972. (In collaboration with Sveriges Mekanförbund.)	25:-
DOCHERTY, P., The Management of Contingencies. Stockholm 1972. <sup>1)</sup>	46:75
ELMGREN-WARBERG, J., Short-term Liquidity Planning - Development of descriptive models. Stockholm 1972. <sup>2)</sup>	34:-
FJAEASTAD, B. & NOWAK, K., Mass Media and Business. Stockholm 1972. (In collaboration with Studieförbundet Näringsliv och Samhälle.)	29:35
HEDBERG, S., Courses of Action when Appropriations are Exceeded in Public Administration - Descriptive studies of rules and practices. Stockholm 1972. <sup>2)</sup>	38:25

1) Published in English.

2) Mimeographed report.



LIST OF REPORTS PUBLISHED SINCE 1970 (cont.)

	Sw.Crs. (circa)
<u>1972</u>	
JENNERGREN, L. P., Studies in Mathematical Theory of Decentralized Resource-Allocation. Stockholm 1972. <sup>1)2)</sup>	25:50
JONSSON, E., Automatic Changes of Budgets - A study of the automatic change of local government finances during the period 1953-65. Stockholm 1972. <sup>2)</sup>	20:-
JONSSON, E., Local Government Finances, Part I, II and Tables. Stockholm 1972.	95:-
KRISTENSSON, F., People, Firms and Regions - A structural economic analysis. 2nd ed. Stockholm 1972. (In collaboration with Almqvist & Wiksell.)	23:80
LINDSTRÖM, B., Foreign Transactions and Business Cycles - Swedish experience during the Bretton-Woods period. Stockholm 1972. <sup>2)</sup>	25:-
MAGNUSSON, Å. & SAMUELSON, L., Budgeting - Studies in some companies. Stockholm 1972. (In collaboration with Sveriges Mekanförbund.)	15:-
NOWAK, K., The Psychological Study of Mass Communication Effects - On the validity of laboratory experiments and an attempt to improve ecological validity. Stockholm 1972. <sup>1)2)</sup>	46:75
RYDEN, I., An Extension of the Icebreaker Fleet in the Gulf of Bothnia - its effects in terms of reduced cost to shippers and its relevance to regional policy. Stockholm 1972. <sup>2)</sup> (Published by Norrlandsfonden)	-
STÅHL, I., Bargaining Theory. Stockholm 1972. <sup>1)</sup>	50:-
THORNGREN, B., Studies in Location - Analysis of regional structures. Stockholm 1972.	33:-
TJERNSTRÖM, S., Studies of Local Government Costs - Estimates of long-range cost functions for old aged homes and elementary education. Stockholm 1972. <sup>2)</sup>	30:-
<u>1971</u>	
ABRAHAMSON, A., GORPE, P. & NYGREN, B., Japan - Economy and Politics. Stockholm 1971.	18:70
ARVIDSSON, G., Pricing of Internal Transfers - Theory and practice. Stockholm 1971. (In collaboration with Sveriges Mekanförbund.)	65:-
ARVIDSSON, G., LARSSON, L. & SUNDQUIST, L., Transfer Pricing - An empirical investigation. Stockholm 1971. <sup>2)</sup>	17:-
BJÖRKMAN, J., Short-term Effects of Traffic Information. Stockholm 1971.	40:-
von ESSEN, G., Investments in Forestry - A profits analysis. Stockholm 1971.	30:-
GUSTAFSSON, L., Negotiations. Stockholm 1971.	17:50
HULTEN, O. & LUNDBERG, D., Using Gratification Studies in Mass Media Research. Stockholm 1971. <sup>2)</sup>	9:-
JANSSON, J. O., Pricing of Street-Space. Stockholm 1971.	30:-
JONSSON, E., Two Models of Budget Process in the Local Government. Stockholm 1971.	30:-
LUNDBERG, E., et.al., Swedish Fiscal Policy in Theory and Practice. Stockholm 1971.	50:-

1) Published in English.

2) Mimeographed report.

# LIST OF REPORTS PUBLISHED SINCE 1970 (cont.)

	Sw.Crs. (circa)
<u>1971</u>	
MATTSSON, L-G., People and Business in a Society of Communications - Contributions to the debate on consumer influence in consumer issues and the policy of firms. Stockholm 1971.	16:50
PERSSON, S., Decision Tables - 1. Description of the relation between decision rules. Lund 1971.	45:75
STOLT, B., The Congruity Principle and the Effect of Relation Strength. Stockholm 1971. <sup>1)2)</sup>	15:-
UGGLA, C., Problems Associated with Corporate Foreign Exchange Positions - An attempt in microeconomic conceptualization. Stockholm 1971.	40:-
<u>1970</u>	
AHLBERG, G. & SUNDQUIST, S-I., Traditional Costing Methods and Linear Programming. Stockholm 1970.	35:-
BACK, R., DALBORG, H. & OTTERBECK, L., Location and Development of Economic Structures. Stockholm 1970.	40:-
BACKELIN, T. & LUNDBERG, E., Economic Policy in Process of Change. Stockholm 1970.	26:-
BERGSTRAND, J., GAVATIN, P., MAGNUSSON, Å. & SAMUELSON, L., Budgeting - A survey of existing literature. Stockholm 1970. (In collaboration with Sveriges Mekanförbund.)	60:-
GUSTAFSSON, L., Decision Making through Negotiations. Stockholm 1970. <sup>2)</sup>	20:-
LAINE, V., Linear Programming in Management - Model construction and possibilities for application. Stockholm 1970.	30:-
MELESKO, S., The Formulation of Goals and Goal Problems in Organization. Stockholm 1970. <sup>2)</sup>	25:-
NOWAK, K., Inducing Resistance to Persuasion - Generality and specificity in the effects of defense-by-refutation. Stockholm 1970. <sup>1)2)</sup>	5:-
NYSTRÖM, H., Retail Pricing. Stockholm 1970. <sup>1)</sup>	45:-
ODHNOFF, J. & OLOFSSON, C-G., Complex Planning Processes - An analysis in connection with some case studies. Stockholm 1970. <sup>2)</sup>	15:-
SANDELL, R. G., Situational Factors in Choice Behavior - Four research papers. Stockholm 1970. <sup>1)2)</sup>	15:-
SEIPEL, C-M., A Study of Consumer Reactions to Promotional Activities in the Form of Gifts and Premiums. Stockholm 1970. <sup>2)</sup>	20:-
SEIPEL, C-M., Premiums - Forgotten by theory. Stockholm 1970. <sup>1)2)</sup>	9:-
STAEL VON HOLSTEIN, C-A. S., Assessment and Evaluation of Subjective Probability Distributions. Stockholm 1970. <sup>1)</sup>	31:50
SUNDQUIST, S-I. & ÖSTERLUND, J-E., Decision-Making and EDP Simulation. Stockholm 1970. <sup>2)</sup>	25:-
WÄRNERYD, K-E., CARLMAN, B., CARLZON, J., CASSEL, U., NOWAK, K. & THORSLUND, S., Advertising and Attention - Some articles. Stockholm 1970. <sup>2)</sup>	16:-
ÖLANDER, F., HJELMSTRÖM, E., LILLIESKÖLD, J. J. & PERSSON, A., A Consumer Taste Test - A comparison between blind-test and stated references. Stockholm 1970. <sup>2)</sup>	9:-
ÖLANDER, F. & SEIPEL, C-M., Psychological Approaches to the Study of Saving. Urbana, Ill. 1970. EFI/University of Illinois, Bureau of Economic and Business Research, Urbana, Ill. <sup>1)</sup>	17:50

1) Published in English.

2) Mimeographed report.



