INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106

EVALUATION OF PROBABILITY FORECASTS IN METEOROLOGY

by

Allan Hunt Murphy

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Atmospheric and Oceanic Science)
in The University of Michigan
1974

Doctoral Committee:

Professor Edward S. Epstein, Co-Chairman
Professor William A. Ericson, Co-Chairman
Professor Aksel C. Wiin-Nielsen
Doctor Robert G. Miller, Life Insurance Market
and Research Association

This thesis is dedicated to my parents -

Ethan Allen Murphy and Mina Dorothea (Hunt) Murphy

ACKNOWLEDGMENTS

The author would like to thank all those individuals who have contributed to this study. In particular, the guidance, assistance, and patience of Professor Edward S. Epstein, Co-Chairman of the Doctoral Committee, is greatly appreciated. The author also wishes to express his appreciation to Professor William A. Ericson and to Professor Aksel C. Wiin-Nielsen and Dr. Robert G. Miller, who served as Co-Chairman and Members, respectively, of the Doctoral Committee, for their interest and encouragement. Special thanks are due Professor E. Wendell Hewson and Professor James G. Wendel, former members of the Doctoral Committee, for their assistance during the early years of the study.

Many of the author's colleagues and friends contributed significantly to this study through substantive discussions and through expressions of interest and encouragement. The author would particularly like to express his appreciation to Dr. Eugene W. Bierly and Dr. Chien-hsiung Yang for their "support" over an extended period of time.

Dr. John W. Firor and Dr. Peter A. Gilman are due special thanks for providing the author with the opportunity to complete this work.

Financial and material support for this study have been provided by a number of institutions and organizations, including the University of Michigan, the National Center for Atmospheric Research, the National Science Foundation, the Travelers Research Center, Inc., and the National Weather Service. The author is extremely grateful

for this support.

Special thanks are due Ms. Gloria S. Parker who typed several drafts of the thesis and the final manuscript.

Finally, and most importantly, the author expresses his gratitude to his wife, Shelly, and to his children, Kenneth, Christopher, Peter, and Andrea, whose patience, understanding, and encouragement were essential to the completion of this work.

TABLE OF CONTENTS

	Page
DEDICATION	ii
ACKNOWLEDGMENTS	iii
LIST OF TABLES	x
LIST OF FIGURES	xiii
LIST OF APPENDICES	xxi
CHAPTER ONE. INTRODUCTION	1
1.1 Forecast Evaluation in Meteorology	2
1.2 Probability Forecasting in Meteorology	7
1.3 Related Developments in Other Fields	14
1.4 The Study	19
1.4.1 The Problem	19
1.4.2 Scope of Thesis	20
1.4.3 Content of Thesis	23
CHAPTER TWO. FORECAST EVALUATION: ITS NATURE AND PURPOSES	26
2.1 Nature of Forecast Evaluation	26
2.2 Purposes of Forecast Evaluation	28
2.2.1 Empirical Evaluation	28
2.2.2 Operational Evaluation	30
CHAPTER THREE. PROBABILITY FORECASTS: THEIR NATURE AND ATTRIBUTES	32
3.1 The Nature of Probability Forecasts	33
3.1.1 Probabilistic Considerations	33
3.1.2 Meteorological Considerations	34
3.1.2.1 Meteorologists	34
3.1.2.2 Forecasting Procedures	34
3.1.2.3 Predictands	34
3.1.2.4 Forecasts	36
3.1.2.5 Observations	37

Page

3.2 The Attributes of Probability Forecasts 37

3.2.1 Normative Attributes: Description and Measurement 37

3.2.1.1 Consistency 38
3.2.1.2 Honesty and "Proper" and Improper Scoring Rules 39

3.2.2 Substantive Attributes: Description 41

3.2.2.1 Empirical Evaluation 42
3.2.2.2 Operational Evaluation and Utility 53

3.2.3 The Normative and Substantive Attributes of Primary Concern 54

CHAPTER FOUR. EVALUATION FRAMEWORKS 56

4.1 A Framework for the Representation of Probability Forecasts 56
4.2 A Framework for Empirical Evaluation 60
4.3 A Framework for Operational Evaluation 62

4.3.1 A Model for Decision Making 62

4.3.1.1 The SEU Model: A Description 62
4.3.1.2 The SEU Model: A Brief Critique 65
4.3.1.3 A Modified SEU Model 67

4.3.2 The Framework for Operational Evaluation 69

4.4 A Comprehensive Framework for Evaluation 77

CHAPTER FIVE. THE FRAMEWORKS: SPECIFICATION OF PRIMITIVE ELEMENTS 78

5.1 The Primitives: Classification 78
5.2 The Quantitative Primitives: Description 79

5.2.1 The Forecast (or Probabilities) 79
5.2.2 The Observation 80
5.2.3 The Utilities 80

5.3 The Primitives: Specification 86

5.3.1 Qualitative Primitives 88
5.3.2 Quantitative Primitives 88

5.3.2.1 The Forecast (or Probabilities) 88
5.3.2.2 The Observation 90
5.3.2.3 The Utilities 90

Page

CHAPTER SIX. EMPIRICAL EVALUATION MEASURES 95

6.1 Some Basic Concepts 95

6.1.1 Scales of Goodness 95
6.1.2 "Proper" and Improper Measures 96
6.1.3 Equivalent Measures 97
6.1.4 Absolute and Relative Measures 98

6.2 Distance Measures 100

6.2.1 A Family of Distance Measures 100
6.2.2 The Absolute Distance Measure 101
6.2.3 A Partition of the Distance Measure 103
6.2.4 The Relative Distance Measure 108

6.3 Distance Measures and Other Measures: Some Comparisons 114

6.3.1 Measures of Validity 114

6.3.1.1 Absolute Measures 114
6.3.1.2 Relative Measures 124

6.3.2 Measures of Bias 134
6.3.3 Measures of Resolution 136
6.3.4 Measures of Sharpness 138

6.4 Empirical Evaluation Measures: A Brief Summary 140

CHAPTER SEVEN. OPERATIONAL EVALUATION MEASURES 148

7.1 Knowledge of Utilities Complete 148

7.1.1 The Utility Measure: Definition 148
7.1.2 The Utility Measure: Examples 150

7.2 Knowledge of Utilities Incomplete 153

7.2.1 Utility and Expected-Utility Measures 153
7.2.2 The Expected-Utility Measures: Definition 154
7.2.3 The Utility and Expected-Utility Measures: Examples 155

7.2.3.1 A General Two-Action, Two-State Decision Situation 156
7.2.3.2 A Special Two-Action, Two-State Decision Situation 165
7.2.3.3 A Special Three-Action, Three State Decision Situation 171

Page

7.3 Some Properties of Utility and Expected-Utility Measures 179

7.3.1 Utility Measures 179
7.3.2 Expected-Utility Measures 182
7.3.3 Utility and Expected-Utility Measures: A Summary 185

CHAPTER EIGHT. EMPIRICAL AND OPERATIONAL EVALUATION MEASURES: SOME RELATIONSHIPS 188

8.1 General Decision Situations 189

8.1.1 Individual Forecasts 190
8.1.2 Collections of Forecasts 192
8.1.3 Some Examples 193

8.1.3.1 Two-State (N=2) Situations 194
8.1.3.2 N-State (N>2) Situations 196

8.1.4 Some Speculations 200

8.2 Special Decision Situations 203

8.2.1 A Two-Action, Two-State Situation 204
8.2.2 A M-Action, M-State Situation 206
8.2.3 General Relationships between Validity and Utility Measures: Discussion 209

CHAPTER NINE. CONCLUSION 211

9.1 Summary and Discussion 211
9.2 Some Extensions 218

9.2.1 A Potpourri of Extensions 218
9.2.2 An Alternative Framework 220

9.3 A Postscript 225

APPENDICES 228

BIBLIOGRAPHY 252

LIST OF TABLES

Table Number	Table Legend	Page
6-1.	A sample collection of forecasts and the relevant observations in the three-state situation.	109
6-2.	The partition of the distance measure $\overline{\Delta^2}(\underline{r},\underline{d})$ for the sample collection of forecasts presented in Table 6-1.	110
6-3(a).	The name, form, and properties of the absolute and relative measures of (primary) validity considered in this study (see subsection 6.3.1).	143-144
6-3(b).	The name, form, and properties of the measures of bias, resolution, and sharpness considered in this study (see subsections 6.3.2, 6.3.3, and 6.3.4, respectively).	146-147
7-1.	The bounds on, or values assumed by, the utility measure $U(\underline{r},\underline{d})$ in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2, when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of a bounded interval measure. (a) General intervals. (b) Specific intervals.	158
7-2.	The bounds on, or values assumed by, the utility measure $U(\underline{r},\underline{d})$ in the special two-action, two state decision situation described by the key utility matrix (c) in Figure 5-3, when the evaluator's knowledge of the utility x is	166

Table Number	Table Legend	Page
7-2. (cont.)	expressed in terms of a bounded interval measure. (a) General interval. (b) Specific interval.	
7-3.	The expressions for the expected-utility measure EU($\underline{r}$,$\underline{d}$) in the special three-action, three-state decision situation described in Figure 7-9, when the evaluator's knowledge of the utility x is expressed in terms of a uniform distribution $x \sim \beta(1,1)$.	176
7-4.	The name, form, and properties of the operational evaluation measures considered in this study (see subsections 7.1.1 and 7.2.2).	187
8-1.	A sample collection of three forecasts prepared by forecasters A and B, the relevant observations, and the validities and utilities of the forecasts according to the measures D($\underline{r}$,$\underline{d}$) and U($\underline{r}$,$\underline{d}$), respectively, for two-state situations. The utilities relate to the decision situation presented in Figures 4-6 and 5-1a.	195
8-2.	A sample collection of forecasts prepared by forecasters A and B, the relevant observations, and the validities and utilities of the forecasts according to the measures D($\underline{r}$,$\underline{d}$) and U($\underline{r}$,$\underline{d}$), respectively, for three-state situations. The utilities relate to the decision situation presented in Figures 4-8 and 5-1b.	197

Table Number		Page
F-1.	The expected-utility measure EU($\underline{r}$,$\underline{d}$) in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2 when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of a bounded interval - uniform measure.	244
G-1.	The integral expressions for the expected-utility measure EU($\underline{r}$,$\underline{d}$) in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2 with independent beta measures on the utilities x_1 and x_2.	246
G-2(a).	The specific expression for the expected-utility measure EU($\underline{r}$,$\underline{d}$) in the two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2 with independent beta measures on the utilities x_1 and x_2 assuming positive integer parameter values. Case 1: $0 \leq r_1 \leq 1/2$.	248
G-2(b).	The specific expression for the expected-utility measure EU($\underline{r}$,$\underline{d}$) in the two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2 with independent beta measures on the utilities x_1 and x_2 assuming positive integer parameter values. Case 2: $1/2 \leq r_1 \leq 1$.	249

LIST OF FIGURES

Figure Number	Figure Legend	Page
4-1.	The sets (a) R^2 and (b) R^3 in a rectangular coordinate system.	58
4-2.	The sets (a) R^2 and (b) R^3 in length and area coordinate systems, respectively.	59
4-3.	The framework for empirical evaluation in the (a) two-state and (b) three-state situations.	61
4-4.	The SEU model of a M-action, N'-state decision situation.	66
4-5.	The modified SEU model of a M-action, N-state decision situation.	70
4-6.	A specific three-action, two-state decision situation.	72
4-7.	The framework for operational evaluation for the three-action, two-state decision situation described in Figure 4-6.	73
4-8.	A specific four-action, three-state decision situation.	75
4-9.	The framework for operational evaluation for the four-action, three-state decision situation described in Figure 4-8.	76
5-1.	The standard utility matrices corresponding to the decision situations depicted in (a) Figure 4-6 and (b) Figure 4-8.	82

Figure Number	Figure Legend	Page
5-2.	The six key utility matrices which represent the twenty-four possible utility matrices in the general two-action, two-state decision situation.	84
5-3.	The three key utility matrices which represent the twelve possible utility matrices in the special two-action, two-state decision situation.	85
5-4.	The cost-loss ratio decision situation.	87
6-1.	The absolute distance measure $D(\underline{r},\underline{d})$ in the two-state situation. The solid curve denotes $D_1(\underline{r})$ and the dashed curve denotes $D_2(\underline{r})$.	104
6-2.	The representation of the terms in the partition of the measure $\Delta^2(\underline{r},\underline{d})$ within the simplex R^3 for a subset of the collection of ten forecasts presented in Table 6-1. The five (K=5) forecasts in the four subcollections (T=4) for which t = 4, 5, 6, and 7 are depicted (see Table 6-2). The contributions to the terms $\Delta^2(\underline{r},\underline{d})$, $\Delta^2(\bar{\underline{d}},\underline{d})$, $\Delta^2(\underline{r}^t,\bar{\underline{d}}^t)$, and $\Delta^2(\bar{\underline{d}}^t,\bar{\underline{d}})$ are represented by solid (——), dashed (– – –), dashed-dotted (— · —), and dashed-long-and-short (— – —) lines, respectively. The numbers in parentheses adjacent to certain lines indicate the number of times that this particular distance occurs for this subset of forecasts (the absence of a number indicates that the distance occurs only once).	111

Figure Number	Figure Legend	Page
6-3.	The absolute validity measures $D(\underline{r},\underline{d})$ [$=Q'(\underline{r},\underline{d})=$ $RPS(\underline{r},\underline{d})$], $L'(\underline{r},\underline{d})$, and $S(\underline{r},\underline{d})$ in the two-state situation. The solid curves denote $D_1(\underline{r})$, $L_1'(\underline{r})$, and $S_1(\underline{r})$ and the dashed curves denote $D_2(\underline{r})$, $L_2'(\underline{r})$, and $S_2(\underline{r})$.	123
6-4.	The validity measure $D(\underline{r},\underline{d})$ in the three-state situation when state s_1 obtains (i.e., when $d_1= 1$).	125
6-5(a).	The relative validity measures $D(\underline{\pi},\underline{r},\underline{d})$, $RSS(\underline{\pi},\underline{r},\underline{d})$, $IR(\underline{\pi},\underline{r},\underline{d})$, and $MSS(\underline{\pi},\underline{r},\underline{d})$ in the two-state situation when $\pi_1= 0.2$. The solid curves and lines denote $D_1(\underline{\pi},\underline{r})$, $RSS_1(\underline{\pi},\underline{r})$, $IR_1(\underline{\pi},\underline{r})$, and $MSS(\underline{\pi},\underline{r})$ and the dashed curves and lines denote $D_2(\underline{\pi},\underline{r})$, $RSS_2(\underline{\pi},\underline{r})$, $IR_2(\underline{\pi},\underline{r})$, and $MSS_2(\underline{\pi},\underline{r})$.	132
6-5(b).	The relative validity measures $D(\underline{\pi},\underline{r},\underline{d})$, $RSS(\underline{\pi},\underline{r},\underline{d})$, $IR(\underline{\pi},\underline{r},\underline{d})$, and $MSS(\underline{\pi},\underline{r},\underline{d})$ in the two-state situation when $\pi_1= 0.5$. The solid curves and lines denote $D_1(\underline{\pi},\underline{r})$, $RSS_1(\underline{\pi},\underline{r})$, $IR_1(\underline{\pi},\underline{r})$, and $MSS_1(\underline{\pi},\underline{r})$ and the dashed curves and lines denote $D_2(\underline{\pi},\underline{r})$, $RSS_2(\underline{\pi},\underline{r})$, $IR_2(\underline{\pi},\underline{r})$, and $MSS_2(\underline{\pi},\underline{r})$.	133
6-6.	The sharpness measures $D'(\underline{r},\underline{e})$ and $IM'(\underline{r})$ in the two-state situation. The solid curve denotes $D'(\underline{r},\underline{e})$ and the dashed curve denotes $IM'(\underline{r})$.	141
7-1.	The utility measure $U(\underline{r},\underline{d})$ in the decision situation depicted in Figure 4-6 with standardized utilities (see Figure 5-1a). The solid lines denote $U_1(\underline{r})$, while the dashed lines denote $U_2(\underline{r})$. The large dots denote the values of $U(\underline{r},\underline{d})$ at each "step" (see text).	151

Figure Number	Figure Legend	Page
7-2.	The utility measure $U(\underline{r},\underline{d})$ in the decision situation depicted in Figure 4-8 with standardized utilities (see Figure 5-1b). The figures in square brackets represent the values of $U_1(\underline{r})$, $U_2(\underline{r})$, and $U_3(\underline{r})$ when $\underline{r} \in R_m^3$ (m=1,2,3,4).	152
7-3.	The utility measure $U(\underline{r},\underline{d})$ in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2, when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of the specific bounded interval measure $0.1 \leq x_1 \leq 0.6$ and $0.3 \leq x_2 \leq 0.8$ (see Table 7-1b). (a) $U_1(\underline{r})$. (b) $U_2(\underline{r})$.	159
7-4.	The expected-utility measure $EU(\underline{r},\underline{d})$ in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2, when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of the specific bounded interval - uniform measure $f(x_1,x_2) = 4$ for $0.1 \leq x_1 \leq 0.6$ and $0.3 \leq x_2 \leq 0.8$ and $f(x_1,x_2) = 0$ otherwise. The solid curve denotes $EU_1(\underline{r})$ and the dashed curve denotes $EU_2(\underline{r})$.	161
7-5(a).	The expected-utility measure $EU(\underline{r},\underline{d})$ in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2, when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of the specific	162

Figure Number	Figure Legend	Page
7-5(a). (cont.)	beta measures $x_1 \sim \beta(a_1,b_1)$ and $x_2 \sim \beta(a_2,b_2)$ for which $a_i = b_i$ $(i=1,2)$. (1) $a_i = b_i = 1$. (2) $a_i = b_i = 2$. (3) $a_i = b_i = 5$. The solid curves denote $EU_1(\underline{r})$ and the dashed curves denote $EU_2(\underline{r})$.	
7-5(b).	The expected-utility measure $EU(\underline{r},\underline{d})$ in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2, when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of the specific beta measures $x_1 \sim \beta(a_1,b_1)$ and $x_2 \sim \beta(a_2,b_2)$ for which $2a_i = b_i$ $(i=1,2)$. (1) $2a_i = b_i = 2$. (2) $2a_i = b_i = 4$. (3) $2a_i = b_i = 8$. The solid curves denote $EU_1(\underline{r})$ and the dashed curves denote $EU_2(\underline{r})$.	163
7-5(c).	The expected-utility measure $EU(\underline{r},\underline{d})$ in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2, when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of the specific beta measures $x_1 \sim \beta(a_1,b_1)$ and $x_2 \sim \beta(a_2,b_2)$ for which $2a_1 = b_1$ and $a_2 = 2b_2$. (1) $2a_1 = b_1 = a_2 = 2b_2 = 2$. (2) $2a_1 = b_1 = a_2 = 2b_2 = 4$. The solid curves denote $EU_1(\underline{r})$ and the dashed curves denote $EU_2(\underline{r})$.	164
7-6.	The utility measure $U(\underline{r},\underline{d})$ in the special two-action, two-state decision situation described by the key utility matrix (c) in Figure 5-3, when the evaluator's knowledge of the utility x is expressed in terms of the specific bounded interval measure $0.1 \leq x \leq 0.6$ (see Table 7-2b). (a) $U_1(\underline{r})$. (b) $U_2(\underline{r})$.	168

Figure Number	Figure Legend	Page
7-7.	The expected-utility measure $EU(\underline{r},\underline{d})$ in the special two-action, two-state decision situation described by the key utility matrix (c) in Figure 5-3, when the evaluator's knowledge of the utility x is expressed in terms of the specific bounded interval - uniform measure $f(x) = 2$ for $0.1 \leq x \leq 0.6$ and $f(x) = 0$ otherwise. The solid curve denotes $EU_1(\underline{r})$ and the dashed curve denotes $EU_2(\underline{r})$.	170
7-8(a).	The expected-utility measure $EU(\underline{r},\underline{d})$ in the special two-action, two-state decision situation described by the key utility matrix (c) in Figure 5-3, when the evaluator's knowledge of the utility x is expressed in terms of the specific beta measures $x \sim \beta(a,b)$ for which $a = b$. (1) $a = b = 1$. (2) $a = b = 2$. (3) $a = b = 5$. (4) $a = b = 10$. The solid curves denote $EU_1(\underline{r})$ and the dashed curves denote $EU_2(\underline{r})$.	172
7-8(b).	The expected-utility measure $EU(\underline{r},\underline{d})$ in the special two-action, two-state decision situation described by the key utility matrix (c) in Figure 5-3, when the evaluator's knowledge of the utility x is expressed in terms of the specific beta measures $x \sim \beta(a,b)$ for which $2a = b$. (1) $2a = b = 2$. (2) $2a = b = 4$. (3) $2a = b = 8$. The solid curves denote $EU_1(\underline{r})$ and the dashed curves denote $EU_2(\underline{r})$.	173
7-9.	A special three-action, three-state decision situation.	174

Figure Number	Figure Legend	Page
7-10.	The framework for operational evaluation for the special three-action, three-state decision situation described in Figure 7-9 when (a) $0 \leq x \leq 1/2$ and (b) $1/2 \leq x \leq 1$.	175
7-11(a).	The expected-utility measure $EU_1(\underline{r})$ in the special three-action, three-state decision situation described in Figure 7-9, when the evaluator's knowledge of the utility x is expressed in terms of a uniform probability distribution $x \sim \beta(1,1)$. $EU_1(\underline{r})$ becomes $EU_3(\underline{r})$ when the labels on the vertices (1,0,0) and (0,0,1) are interchanged.	177
7-11(b).	The expected-utility measure $EU_2(\underline{r})$ in the special three-action, three-state decision situation described in Figure 7-9, when the evaluator's knowledge of the utility x is expressed in terms of a uniform probability distribution $x \sim \beta(1,1)$.	178
7-12.	The form of the isograms of expected utility (i.e., the "lines" of equal expected utility) within the framework of the simplex R^3 for the three-action, three-state decision situation described in Figure 7-9. The term "none" means that all of the points in the region are of equal expected utility.	180
8-1.	The representation of the forecasts presented in Table 8-2 within the framework of the simplex R^3 and the decision situation presented in Figures 4-8 and 5-1b.	198

Figure Number	Figure Legend	Page
8-2.	The decision situation presented in Figures 4-8 and 5-1b depicted within the framework of the simplex R^3. Each arc represents a set of forecasts of equal validity according to the distance measure $D_2(\underline{r})$. These arcs were chosen to include the forecasts $\underline{r}_1^A$ and $\underline{r}_1^B$ (see Table 8-2 and Figure 8-1).	201
8-3.	The cost-loss matrix in the M-action, M-state generalization of the cost-loss ratio decision situation when M = 5. The variable x represents the cost-loss ratio (i.e., x = C/L).	208
9-1.	The framework for the representation of forecasts and observations in three-state (N=3) situations, the equilateral triangle R^3 (dashed lines), depicted in a cartesian coordinate system. The corresponding framework for the representation of cumulative forecasts and observations, a right, isosceles triangle Π^3 (heavy solid lines) on the top face of the unit cube (light solid lines), is also depicted in this figure.	222
9-2.	The cumulative forecast $\underline{R}$ = (0.2,0.7,1.0) depicted within the framework for the representation of cumulative forecasts and observations in three-state (N=3) situations, a right, isosceles triangle Π^3 in a two-dimensional cartesian coordinate system. The dashed lines represent the distances between the cumulative forecast $\underline{R}$ and the three cumulative observations $\underline{D}_1$= (1,1,1), $\underline{D}_2$= (0,1,1), and $\underline{D}_3$= (0,0,1). The squares of these distances are $RPS_1(\underline{R})$, $RPS_2(\underline{R})$, and $RPS_3(\underline{R})$, respectively.	224

Figure Number		Page
9-3.	The forms of the matrix $\underline{A}$ in Equation (9-7) when N = 5 for (a) the distance measure $D(\underline{r},\underline{d})$ and (b) the ranked probability score $RPS(\underline{r},\underline{d})$.	226

LIST OF APPENDICES

Appendix Number	Appendix Title	Page
A	The class of distance measures $\Delta^k(\underline{r},\underline{d})$: "proper" and improper scoring rules	228
B	The information ratio $IR(\underline{\pi},\underline{r},\underline{d})$: an improper scoring rule	231
C	The matrix skill score $MSS(\underline{\pi},\underline{r},\underline{d})$: a proper scoring rule	234
D	The utility measure $U(\underline{r},\underline{d})$: a proper scoring rule	236
E	The expected-utility measure $EU(\underline{r},\underline{d})$: a "proper" scoring rule	238
F	The expected-utility measure $EU(\underline{r},\underline{d})$ in a general two-action, two-state decision situation with a bounded interval - uniform measure on the utilities x_1 and x_2	242
G	The expected-utility measure $EU(\underline{r},\underline{d})$ in a general two-action, two-state decision situation with independent beta measures on the utilities x_1 and x_2	245
H	The conditions under which the expected-utility measure $EU(\underline{r},\underline{d})$ is linearly related to the distance measure $\Delta^2(\underline{r},\underline{d})$ in a special two-action, two-state decision situation	250

CHAPTER ONE. INTRODUCTION

This study is concerned with the problem of evaluating, i.e., assessing the "goodness" of, probability forecasts. In a meteorological context, this problem arises at the "intersection" of two areas of research, namely forecast evaluation and probability forecasting. In order to provide an appropriate setting for the material to be found in subsequent chapters of this dissertation, brief surveys of the relevant research in these two areas are presented in Sections 1.1 and 1.2, respectively.

Forecast evaluation and probability forecasting are, of course, topics of concern not only to meteorologists. Probabilists, statisticians, economists, management scientists, operations researchers, psychologists, and others are becoming increasingly interested in problems related to forecast evaluation and/or probability forecasting, and recent developments in these fields have important implications for the problem of concern in this study. The most relevant developments in these areas and their implications for the evaluation of probability forecasts in meteorology are briefly discussed in Section 1.3.

In Section 1.4, we describe the problem of concern in this study in greater detail, outline the organization and contents of the dissertation, and present some preliminary concepts and terminology.

1.1 Forecast Evaluation in Meteorology

Forecast evaluation has had a long and controversial history in meteorology, and the literature in this area is quite extensive. As a result, a comprehensive survey of this literature is beyond the scope of this study. Moreover, since we are concerned with the evaluation of probability forecasts, significant portions of this literature (e.g., the portion concerned with the evaluation of categorical forecasts) are not directly relevant. Thus, we shall restrict this survey to the evaluation of probability forecasts and to the process of evaluation itself. For general surveys and/or reviews of the literature in the area of forecast evaluation, we refer the reader to Brier and Allen (1951, 1958), Dobryshman (1972), Gringorten (1951), and Johnson (1957). For an annotated bibliography on forecast evaluation, refer to Meglis (1960).

The evaluation process has been discussed by Brier and Allen (1951, 1958), Johnson (1957), and Murphy and Epstein (1967a) among others. An important step in this process consists of identifying the purposes of evaluation. In this regard, two general purposes have been identified, and these purposes yield two distinct types of evaluation. In this discussion, we shall refer to these forms of evaluation as scientific evaluation and economic evaluation. The former is concerned with the "accuracy," "reliability," "skill," etc., of forecasts, while the latter is concerned with the "value" of forecasts. Scientific evaluation is often referred to as forecast

verification. The purposes and types of evaluation and the attributes of the forecasts of interest for each type of evaluation will be considered in detail in Chapters 2 and 3, respectively.

In the early probability forecasting experiments (see Section 1.3), forecast evaluation consisted simply of comparing, in a qualitative manner, the forecast probabilities and the observed relative frequencies. Such comparisons were made by Cooke (1906a), Hallenbeck (1920), and Brier (1944). The first quantitative measure specifically designed to evaluate probability forecasts was the Brier, or probability, score (Brier, 1950). This measure is simply the mean square error of the forecast, where the forecast and the observation are discrete probability distributions defined on a set of mutually exclusive and collectively exhaustive categories, or states, of the relevant variable. Subsequently, Sanders (1958, 1963) described a method of partitioning the probability score into two distinct measures for variables with two states. This method of partitioning the probability score has recently been extended to N-state ($N \geq 2$) situations in which the forecasts are considered to be both scalars and vectors (Murphy, 1972a,b). In addition, Murphy (1973b) has formulated a new vector partition of the probability score which yields three distinct measures.

Evaluation measures such as the probability score are "absolute" measures and, as such, are concerned with measuring the attributes of the forecasts (e.g., "accuracy") on an absolute scale.

Meteorologists, on the other hand, are often also concerned with the "accuracy" of their forecasts relative to the "accuracy" of forecasts prepared by a "standard" procedure such as climatology (e.g., in order to compare forecasts prepared under different meteorological conditions). These "relative" measures are generally referred to as "skill scores." A ratio skill score based upon the probability score, together with the probability score itself in both its partitioned and unpartitioned forms, are currently being used by the National Weather Service to evaluate its probability forecasts of precipitation occurrence (e.g., Hughes, 1967, 1968; Derouin and Cobb, 1970). The relative merits of ratio and difference skill scores have recently been considered by Murphy (1973a, 1974).

Other evaluation measures formulated in a meteorological context include the "matrix skill score" (Gringorten, 1951, 1965, 1967), the "information ratio" (Holloway and Woodbury, 1955), and the "validity measure" (Miller, 1962). These measures, which have not been used extensively by meteorologists, are briefly examined and compared by Murphy and Epstein (1967a).

The probability score and the other measures mentioned heretofore are based upon the assumption that the variable of concern is an unordered (i.e., nominal) variable or that, if the variable is ordered (for example, on an ordinal or interval scale), that order is not a relevant consideration in the evaluation process. On the other hand, many meteorologists have felt that forecasts for ordered

variables which are, in some sense, "closer" to the observation should receive a "better" score. The ranked probability score, which was formulated by Epstein (1969) within the context of a decision situation (see below), was the first evaluation measure for probability forecasts which possessed the property of being "sensitive to distance." Subsequently, a family of measures which is sensitive to distance has been described by Stael von Holstein (1970b). Murphy (1970, 1971) has compared the ranked probability score and the probability score and demonstrated that the ranked probability score can be expressed in a particularly simple form.

With the exception of the discussion in the previous paragraph, heretofore we have been concerned exclusively with the scientific aspects of evaluation. Frameworks within which economic evaluation could be undertaken were implicit, if not explicit, in the early work on "meteorological decision making" by Brier (1944), Thompson (1952), and Thompson and Brier (1955) and in the more recent work in this area by Epstein (1962), Glahn (1964), Gringorten (1958, 1959), and Nelson and Winter (1964) among others. In this regard, Gringorten (1958) was perhaps the first meteorologist to explicitly assess the "value" of forecasts within such a framework. Murphy (1966, 1969a) has used such a framework to formulate measures of the "value" of probability forecasts in simple decision situations. This framework provides a means of assessing the "value" of such forecasts in situations in which knowledge of the "utilities" associated with the

consequences (by the decision maker or decision makers) is incomplete. In this regard, Murphy (1966) has shown that the probability score is a measure of "value" when knowledge of the decision maker's "cost-loss ratio" can be expressed in terms of a uniform (probability) distribution. This approach can be extended to more complex decision situations, and the ranked probability score (Epstein, 1969) and the family of sensitive to distance measures described by Stael von Holstein (1970b) are examples of evaluation measures formulated within the context of such situations.

A difficulty which arises in connection with economic evaluation relates to the specification of a decision maker's utilities. This problem is particularly apparent in situations in which many different decision makers are of concern. In such situations, measures have been proposed which are assumed to consist of "representative" utilities for the collection of decision makers. Such measures, which we refer to as "psuedo-value" measures, have been formulated and/or discussed by Glahn (1974), Gringorten (1965), and Thompson (1972) among others.

Finally, the relationships (if any) between measures concerned with scientific evaluation and measures concerned with economic evaluation are of considerable interest. Gringorten (1958) is believed to be the first meteorologist to explicitly investigate such relationships. Murphy (1972d) recently demonstrated that ordinal relationships exist between measures of "accuracy" and measures of

"value" in certain special situations. The work of Epstein (1969), Murphy (1966, 1969a, 1970, 1971), and Stael von Holstein (1970b) also relates to this question.

The evaluation process and many of the measures mentioned in this section will be considered in greater detail in subsequent chapters.

1.2 Probability Forecasting in Meteorology

While probability forecasting in meteorology can quite properly be said to have received its initial impetus almost seventy years ago (see below), most of the research in this area has been conducted by a relatively small group of meteorologists and statisticians during the last twenty-five years. Still, as in the case of forecast evaluation, the literature in this area is quite extensive, and we can only briefly describe some of the most significant developments in this section. For collections of papers and/or reviews relating to the work in this area, we refer the reader to American Meteorological Society (1968, 1971, 1973), Glahn (1965), Murphy (1972e), and World Meteorological Organization (1966). For a comprehensive bibliography on probability forecasting, refer to Murphy and Allen (1970).

For the purposes of this survey, we shall distinguish between "objective" and subjective methods of probability forecasting. "Objective" methods involve the use of probabilistic and/or statistical procedures either by themselves or in combination with dynamical and/

or physical models, and are "objective" in the sense that, for a particular method and set of relevant data, the forecast produced does not depend upon a meteorologist's judgment. Subjective methods, on the other hand, are methods in which the formulation of the forecasts is based at least in part upon the judgment of a particular meteorologist. According to these definitions, then, the weather forecasts issued to the general public today are subjective forecasts, even though they may often be based largely upon guidance forecasts produced by "objective" procedures. In this section we also briefly discuss the reasons and/or need for probability forecasts.

The need to quantify the uncertainty inherent in weather forecasts was first recognized by Cooke (1906a), who assigned numerical weights, or "confidence factors," to the statements contained in public forecasts for Western Australia. In defense of his procedure, Cooke (1906b) concluded with a statement which, even today, meteorologists and others would do well to heed:

> 'Now this is the point that I wish to make clear. Those forecasts which were marked "doubtful" were the <u>best I could frame</u> under the circumstances. I could see no way of improving them at the time, and they would not have been expressed differently whether I weighted them or not. If I make no distinction between these and others, I degrade the whole. But if, on the other hand, I attach a figure which practically says "I'm sorry, but this is the best

I can do for you today - do not attach too much importance to it," I eliminate beforehand the adverse opinion which a great number of incorrect forecasts must produce, and I raise the bulk of the predictions to their true value.'

von Myrbach (1913), apparently without knowledge of Cooke's proposal, advocated that meteorologists express the "approximate degree of probability of success" of forecasts in order to increase their reliability, and his proposal was adopted by the Signal Corps Meteorological Service of the American Expeditionary Forces in France in World War I (U.S. Army Signal Corps, 1919). Hallenbeck (1920) described the results of forecasting precipitation in "percentages of probability" in the Pecos Valley in New Mexico. The subjective forecasting procedure, which was based upon a series of composite weather maps, was developed in response to requests by local farmers for "the chances of rain" and provided probabilities that corresponded quite closely to the observed relative frequencies.

The first "objective" probability forecasting procedures were based upon simple contingency tables and scatter diagrams and involved at most two predictors (see, for example, Besson, 1905; Blair, 1921, 1924; Rolf, 1917). Interestingly, in these early studies, the "probability forecasts" provided by the techniques were immediately translated into categorical forecasts by selecting, in each case, the event with the largest probability. The scatter diagram, or graphical regression, approach was extended to more than two variables

within a general framework by Brier (1946). This paper stimulated considerable interest among meteorologists in "objective" forecasting, and resulted in the application of the technique to a variety of variables in various geographic locations (e.g., Dickey, 1949; Gentry, 1950; Jorgenson, 1953; Penn, 1948; Schmidt, 1951; Thompson, 1950; Vernon, 1947). While this technique yielded probability forecasts, these forecasts were still generally transformed into categorical forecasts. Other studies during this period involving the formulation of "objective" probability forecasting techniques included the work of Gringorten (1950, 1952) and Lund (1955). For reviews of "objective" weather forecasting during this period, the reader is referred to Allen and Vernon (1951) and Gringorten (1955).

Several "objective" statistical procedures were developed and/or refined at the Travelers Research Center in the period between 1959 and 1969, including multiple discriminant analysis (MDA) (Miller, 1962) and regression estimation of event probabilities (REEP) (Miller, 1964). These techniques, both of which involved the use of a stepwise procedure to select predictors, were applied extensively to various meteorological variables, in particular ceiling height and/or visibility. The results of this research indicated that the forecasts produced by these two procedures were approximately comparable, and the computational advantages of the REEP technique have led to its becoming the favored procedure. Another technique developed during this period was the logit model (Brelsford and Jones, 1967; see also

Bryan, 1968), and this model is presently being used by the NWS to produce "objective" forecasts of the conditional probability of frozen precipitation (Bocchieri and Glahn, 1973).

More recently, "objective" probability forecasting techniques have primarily involved the combined use of statistical and numerical procedures. Two procedures which have received particular attention are the "perfect prog" and "imperfect prog," or model output statistics (MOS), techniques (Klein, 1970). The MOS procedure uses a statistical technique, generally REEP, to relate the forecasts produced by the numerical models to the "surface weather" predictands of interest. This procedure is presently being used to produce probability forecasts of precipitation occurrence (the probability of precipitation, or POP, forecasts), occurrence of frozen precipitation (see above), precipitation amount, cloud amount, ceiling and visibility, and thunderstorms and severe weather (Klein and Glahn, 1974). These forecasts are provided to the forecasters as guidance in the preparation of their subjective forecasts, which with the exception of forecasts of precipitation occurrence are still expressed in categorical terms.

The "modern" era in subjective probability forecasting in meteorology began with a paper by Brier (1944), in which (among other things) he presented a set of experimental precipitation probability forecasts. Subsequently, the use of confidence factors in connection with both temperature and precipitation forecasts was investigated experimentally

by Williams (1951). The results of these and other experiments during this period (1944 to 1953) indicated that forecasters could describe their uncertainty quantitatively, although the number of forecasts made in these experimental studies was often such that the forecasts were not particularly reliable for certain probability values.

In 1954, the Travelers Weather Service began to issue precipitation probability forecasts to the general public in the Hartford, Connecticut area on a regular basis. In 1956, the U.S. Weather Bureau Forecast Center in San Francisco, California added numerical rain probability statements to their public weather forecasts, and they were joined in this practice by the Forecast Center in Los Angeles, California in 1957 (see Root, 1962). Sanders (1958, 1963) described the results of several years of experimental probability forecasting in the Synoptic Laboratory of the Department of Meteorology at the Massachusetts Institute of Technology (MIT) and concluded that "forecasters are capable of skillful statements about the probability of occurrence of a wide variety of meteorological events over ranges up to seventy-two hours" (Sanders, 1963, p. 200). Root (1962), who analyzed four years of the Weather Bureau's public precipitation probability forecasts at San Francisco, reached a similar conclusion. Thus, even the results of these early experimental and operational programs revealed that forecasters could quantify the uncertainty inherent in their forecasts. During this

period (1944-1964), the need for and importance of probability forecasts was emphasized by Brier (1944, 1950, 1957), Gringorten (1959), Malone (1957), Thompson (1950, 1952), and Thompson and Brier (1955) among others.

After several years of additional experimental precipitation probability forecasting in several locations (see, for example, Hughes, 1965), the Weather Bureau initiated a nationwide program in 1965 in which precipitation probabilities were appended to all public weather forecasts on a regular basis. These forecasts have been subject to extensive evaluation both regionally (e.g., Hughes, 1967, 1968) and nationally (e.g., Derouin and Cobb, 1972; Sadowski and Cobb, 1973), and they have been found to be quite reliable.

More recently, the results of two experiments reveal that forecasters can express the uncertainty inherent in their forecasts of continuous variables such as temperature in terms of credible intervals (Murphy and Winkler, 1974b; Peterson, Snapper, and Murphy, 1972). Other recent subjective probability forecasting experiments have involved point and area precipitation probabilities and the effect of "objective" guidance forecasts upon the forecasters' subjective precipitation probability forecasts (Murphy and Winkler, 1974a). Sanders (1973) has recently summarized the results of six years of probability forecasting in the Synoptic Laboratory at MIT (minimum temperature and precipitation occurrence and amount). These forecasts appear to be quite reliable, but the results indicate

that little if any improvement in accuracy and/or skill has occurred over the six-year period.

1.3 Related Developments in Other Fields

As previously indicated, scientists in a number of fields have been (and are) conducting research which has important implications for forecast evaluation and probability forecasting in meteorology. Most of this research falls under the general heading of "subjective probability, utility, and decision making." The literature in this area is very extensive, and a general survey of the recent developments is beyond the scope of this study. Instead, we shall restrict our attention primarily to two topics, probability assessment and scoring rules, which we believe are particularly relevant to the problem of concern in this thesis.

The subjective view of probability, the view to which we subscribe in this study, prescribes that a probability represents a (rational) "degree of belief" (Winkler, 1972, pp. 15-18). This interpretation appears to be particularly appropriate in a meteorological context (see Epstein, 1966; Murphy and Winkler, 1971; Roberts, 1968). An axiomatic basis for the theory of subjective probability has been developed by de Finetti (1937) and Savage (1954). Until recently, however, little attention had been paid to the practical problem of assessing (i.e., quantifying) subjective probabilities. Individual subjective probabilities were assessed in connection with

oil drilling decisions (Grayson, 1960) and the outcomes of football matches (de Finetti, 1962). The first extensive study of the assessment of subjective probability distributions was conducted by Winkler (1967a). His paper describes several alternative assessment procedures and presents the results of an empirical study in which subjects used these techniques to assess distributions for selected characteristics of the student population at the University of Chicago. More recently, experiments involving the assessment of subjective probabilities have been conducted in several areas, including educational testing (e.g., Shuford, 1969; see also de Finetti, 1965), intelligence analysis (e.g., Kelly and Peterson, 1971), sporting events (e.g., Winkler, 1971), the stock market (e.g., Bartos, 1969; Stael von Holstein, 1970a), and general "almanac-type" questions (e.g., Alpert and Raiffa, 1969; Brown, 1973). In addition to these studies, psychologists have conducted many laboratory experiments related to the quantification of judgment and the revision of opinion, and these experiments have frequently involved the assessment of probabilities or probability distributions (see, for example, Peterson and Beach, 1967; Slovic and Lichtenstein, 1971; Rapaport and Wallsten, 1972; Tversky and Kahneman, 1973). While the results of these experiments indicate that different assessment procedures can lead to different probabilities (and different distributions) and that the subjects' probabilities and distributions often exhibit certain biases and inconsistencies, they also reveal that, in general, the subjects

are able to quantify their uncertainty in a reasonably satisfactory manner. In this regard, many of these experiments (particularly those experiments conducted by psychologists) have involved the use of artificial laboratory situations and naive subjects, and some evidence exists to support the belief that true experts in realistic situations in their fields of expertise will provide more consistent and unbiased assessments (see Winkler and Murphy, 1973). For a recent review of subjective probability assessment, refer to Hampton, Moore, and Thomas (1973). Finally, a forthcoming paper by Spetzler and Stael von Holstein (1974) contains an interesting discussion of probability assessment within the context of decision analysis.

According to the theory of subjective probability, probability assessments should meet certain (normative) requirements (e.g., Winkler, 1967b, p. 1106). One such requirement is that the assessments, or forecasts, should correspond with the assessor's judgments. While the degree of correspondence between assessments and judgments cannot be determined directly (since the latter exist only in the assessor's mind), "methods" are available to help the assessor make forecasts which are in accordance with his judgments. One such method involves the use of <u>scoring rules</u>. A scoring rule is simply a function of the assessor's (forecast) probabilities and the event or state that subsequently obtains. In order to encourage the assessor to make his forecasts correspond to his judgments, these scoring rules should be defined in such a way that the assessor can

maximize (or minimize, whichever is appropriate) his subjective expected score if and only if he makes his assessments correspond to his judgments. Such scoring rules will be referred to as strictly proper scoring rules in this study. A number of strictly proper scoring rules have been identified, including the quadratic score (de Finetti, 1962, 1965)[1], the logarithmic score (Good, 1952; van Naerssen, 1962), and the spherical score (Roby, 1965; Toda, 1963). The nature and properties of these and other scoring rules have been studied by a number of investigators, including Aczel and Pfanzagl (1966), Brown (1970), de Finetti (1965), Hendrickson and Buehler (1971), McCarthy (1956), Murphy and Winkler (1970), Raiffa (1969), Roberts (1968), Savage (1971), Shuford, Albert, and Massengill (1966), Stael von Holstein (1970a), and Winkler (1967b, 1969).

Strictly proper scoring rules have been used in connection with probability assessment tasks in a variety of contexts, including educational testing (e.g., Shuford and Massengill, 1967), intelligence analysis (e.g., Hoffman and Peterson, 1972), sporting events (e.g., Winkler, 1971), the stock market (e.g., Stael von Holstein, 1970a), and weather forecasting (e.g., Stael von Holstein, 1971). The results of these experiments indicate that the use of such scoring rules seems to lead assessors to make careful assessments. These

---

[1]The quadratic score is equivalent to the probability, or Brier, score (Brier, 1950) (see Chapter 6).

scoring rules, then, appear to play an important role in the assessment process. In this regard, the fact that hedging is a problem of some concern in meteorology (e.g., Murphy and Epstein, 1967b; Murphy and Winkler, 1971) suggests that strictly proper scoring rules, which are defined in such a way as to discourage hedging, will be of particular interest in this context.

Finally, researchers in several fields have been concerned with the formulation and application of models of decision making under uncertainty. The nature and properties of alternative models have been described by Ferguson (1967), Fishburn (1964), and Luce and Raiffa (1957) among others. In this study we are particularly concerned with the subjective expected-utility model, in which the decision maker is assumed to want to select that course of action which maximizes his expected utility with respect to a subjective probability distribution defined on the possible outcomes (see, for example, Edwards, 1954; Fishburn, 1964). This model is, in essence, equivalent to the Bayesian model for decision making. The nature and properties of the Bayesian model have been described by Aitchison (1970), Lindley (1971), Raiffa (1968), Schmitt (1969), and Winkler (1972) among others.

1.4 The Study

In this section we identify and briefly describe the problem of concern in this study. We also indicate the scope and content of the thesis.

1.4.1 The Problem

As the surveys in the previous sections of this chapter indicate, meteorologists as well as scientists in many other fields have undertaken numerous studies of (or related to) the evaluation of probability forecasts. However, these studies have been concerned primarily with the formulation and/or use of particular evaluation measures. The evaluation process itself has received little attention, and the need for a comprehensive framework for the evaluation of probability forecasts has become increasingly apparent. The basic objectives of this study are to formulate such a framework and to describe its basic elements (e.g., its measures of "goodness") in detail. We believe that such a study must necessarily include, at an early stage, the examination of the nature and purposes of forecast evaluation and the identification of desirable attributes of probability forecasts from both scientific and economic points of view. The framework itself should contain measures of both kinds of attributes and should be comprehensive in the sense that the framework should "admit" the existence of relationships between these two types of measures. Moreover, since the specification of the "values" associated with the use of such forecasts involves considerable

uncertainty from a meteorological point of view, the framework should also provide economic measures for situations in which knowledge of these "values" is incomplete.

1.4.2 Scope of Thesis

In this study we shall adopt the general point of view of an individual who is concerned with the problem of assessing the "goodness" of probability forecasts formulated by one or more meteorologists. We shall refer to this individual as the evaluator. The evaluator may be a meteorologist, a user or decision maker, or simply an interested observer. We may, on occasion, implicitly assume that the evaluator is a particular one of these individuals. For example, economic evaluation in situations in which the evaluator's knowledge of the decision maker's "values" is assumed to be incomplete would appear to be of particular interest to an evaluator who is a meteorologist.

Probability forecasts are, of course, simply probability distributions. As such, these forecasts may be continuous or discrete, depending upon the manner in which the variable of concern is defined. In meteorology, probability forecasting has generally involved variables that have been divided into two or more mutually exclusive and collectively exhaustive categories or states. For example, the probability of precipitation forecasts currently issued to the general public by the National Weather Service involve a dichotomous

variable (precipitation/no precipitation). Variables such as temperature have only recently been treated as continuous variables even on an experimental basis (Murphy and Winkler, 1974b; Peterson, Snapper, and Murphy, 1972), and these experiments have involved the use of credible intervals rather than continuous distributions. Thus, we shall assume that the variables of concern have been divided into N mutually exclusive and collectively exhaustive states and that a probability forecast consists of N probabilities, one assigned to each of the relevant states.

The evaluation frameworks to be described herein are general frameworks in that they are applicable to forecasts of N-state (N>2) as well as two-state (N=2) variables. However, as we shall discover, certain practical problems arise in connection with the use of these frameworks as N (i.e., their dimensionality) increases.[1] For example, the formulation of measures of "value" in situations involving three or more states is, at best, a difficult and time-consuming process. In addition, the representation of the forecasts within the proposed geometrical framework becomes difficult when the number of states exceeds three. Of greater importance, difficulties encountered in

---

[1] We have no reason to believe that these problems are uniquely related to the frameworks that we have adopted. In fact, the nature of the evaluation problem is such that most (if not all) approaches to this problem can be expected to encounter similar difficulties.

N-state situations have prevented the author from generalizing some of the results obtained for two-state and three-state situations. For these reasons, and because the two-state situation is of considerable interest from a meteorological point of view, this study will focus particular attention upon the results and relationships obtained in this special situation. Whenever possible, we shall either extend these results to N-state situations or speculate about the nature of the results and relationships which may exist in such situations.

As indicated in Section 1.1, some meteorologists (and others) have argued that the fact that the variable of concern is or is not an ordered variable is a relevent consideration in the evaluation process. Specifically, they believe that different frameworks and/or different measures of "goodness" are appropriate for ordered and unordered variables. As an active participant in this dialogue, the author recognizes that this point of view has some merit. However, the formulation of a general framework for the evaluation of probability forecasts of ordered variables was beyond the scope of this study. Thus, the frameworks to be described herein are implicitly based upon the assumption that order is <u>not</u> a relevant consideration (or that the variables of concern are unordered). On the other hand, the formulation of a framework for unordered variables can be considered to be the first step in the formulation of a framework for ordered variables. In this regard, we consider a

measure of "goodness" for ordered variables in Chapter 6, and we briefly describe a specific framework for the evaluation of ordered variables in Chapter 9.

1.4.3 Content of Thesis

This work contains eight chapters in addition to the introductory chapter, and we briefly describe the contents of these chapters in this subsection.

Chapter 2. Forecast Evaluation: Its Nature and Purposes

The evaluation process is described, with particular reference to the purposes of evaluation. Two forms of evaluation are identified, empirical evaluation and operational evaluation.

Chapter 3. Probability Forecasts: Their Nature and Attributes

The nature of probability forecasts is considered from both probabilistic and meteorological points of view, and a subjective interpretation of probability is adopted. Desirable attributes of probability forecasts from both normative and substantive viewpoints are identified and described, and certain implications of these attributes for evaluation measures are indicated.

Chapter 4. Evaluation Frameworks

Frameworks for empirical and operational evaluation are described in algebraic and geometric terms. The framework for operational evaluation, which is based in part upon a modified version of the subjective, expected-utility model of a decision situation, is shown to be a comprehensive framework for evaluation.

Chapter 5. The Frameworks: Specification of Primitive Elements

The evaluator's state of knowledge concerning the primitive elements which constitute the frameworks for evaluation is considered. His knowledge of all of the primitives except the decision maker's utilities is assumed to be complete. The evaluator's knowledge of the utilities may be complete or incomplete and, when incomplete, is assumed to be expressed in terms of either inequality or probability measures.

Chapter 6. Empirical Evaluation Measures

Measures of the attributes of concern in empirical evaluation are formulated within the framework described in Chapter 4. The nature and properties of these measures are investigated and the relationships between the measures and measures proposed heretofore are described.

Chapter 7. Operational Evaluation Measures

Measures of the attributes of concern in operational evaluation are formulated within the framework presented in Chapter 4 for situations in which the evaluator's knowledge of the decision maker's utilities is complete and incomplete. The properties of these measures are described.

Chapter 8. Empirical and Operational Evaluation Measures: Some Relationships

The relationships between empirical and operational evaluation measures are investigated within the comprehensive framework for

evaluation described in Chapter 4.

Chapter 9. Conclusion

The results of the study are summarized and a number of possible extensions are suggested.

CHAPTER TWO. FORECAST EVALUATION: ITS NATURE AND PURPOSES

This chapter consists of two sections in which we consider the nature and purposes of forecast evaluation, with particular reference to probability forecasts. In Section 2.1 forecast evaluation is described as a process which consists of three phases, a formulation phase, a comparison phase, and a utilization phase, and several steps in the formulation phase are identified. The primary purposes of forecast evaluation are identified and described in some detail in Section 2.2.

2.1 Nature of Forecast Evaluation

Evaluation is "the process of assessing the value of" (Webster's Third New International Dictionary, 1966). Thus, forecast evaluation is the process of assessing the "value" (or "goodness") of forecasts. We consider that this process consists of three phases: (a) a formulation phase, in which "measures of value," i.e., evaluation measures, are formulated; (b) a comparison phase, in which evaluation measures are compared and the relationships between such measures are investigated; and (c) a utilization phase, in which these measures are actually used to assess the "goodness" of forecasts. In Chapters 1, 2, and 3 of this dissertation we are concerned with certain basic considerations relative to the formulation phase. A framework within which evaluation measures can be formulated is described in Chapters 4 and 5 and measures are formulated and

compared within this framework in Chapters 6, 7, and 8. While we do not consider the utilization phase in detail, certain aspects of this phase of the evaluation process are discussed in Chapters 6, 7, 8, and 9.

The formulation phase consists of several steps. Note that the definition of the term "evaluation" implies that, in this phase, we are concerned with the formulation of "measures of value," each of which establishes a value scale. However, the term "value" has meaning only with reference to an individual with an objective (Fishburn, 1964, p. 3). Thus, we must identify (a) the individuals who are concerned with forecast evaluation and (b) the objectives of these individuals. Therefore, the first step in the formulation phase is to identify the individuals of concern and their objectives. We shall refer to this first step simply as the identification of the purposes of evaluation. Each purpose defines a distinct form of forecast evaluation (see Section 2.2).

An important consequence of this first step is the identification, in qualitiative terms, of "properties" of forecasts which are indicative of their "value" to the relevant individuals. In order to formulate measures of value we must, of course, describe these "desirable" properties, or attributes, in quantitative terms. Thus, the second step in the formulation phase is to describe, in quantitative terms, the attributes of the forecasts. We are often able, at this step, to establish an incomplete value scale by

identifying one or more "natural" points on the scale. We shall refer to this second step simply as the description of the attributes for each purpose.

The third step in the formulation phase is to formulate the measures for a particular attribute. A measure provides a means of determining the extent to which the forecasts possess a particular attribute. In particular, a measure establishes a complete, but to a certain extent arbitrary, or artificial, value scale. We shall refer to this third step simply as the formulation of the measures for the attributes.

2.2 Purposes of Forecast Evaluation

Meteorological forecasts are of concern to two individuals, the meteorologist and the decision maker. Thus, at least two general purposes for, and forms of, forecast evaluation can be identified. The nature of the meteorologist's and the decision maker's concern as well as the nature of the two forms of evaluation are considered in this section. These and related specific purposes for forecast evaluation have been discussed by Brier and Allen (1951, 1958), Johnson (1957), and Murphy and Epstein (1967a) among others.

2.2.1 Empirical Evaluation

The presence of uncertainty in the forecasting process implies that meteorological forecasts are, at present, imperfect. In such a situation, the meteorologist's immediate objective is to formulate

a forecasting system that produces forecasts (or, simply, to make forecasts) which possess a greater "degree of perfection." The meteorologist's ultimate objective is, of course, to make perfect forecasts. Thus, the degree of perfection is the general attribute of forecasts which is indicative of the "value" of the forecasts (or of the parent forecasting system) to the meteorologist. Note that the degree of perfection of a forecast is directly related to the degree of correspondence, or association, between the forecast and the observation. Hereafter, we shall refer to the general attribute of concern from a meteorological point of view as the "degree of association" (of forecasts and observations), and to the process of determining the absolute and/or relative degree of association of forecasts and observations as <u>empirical</u>, or <u>inferential</u>, evaluation.

At this stage, we are able to specify (only) one point of a degree-of-perfection scale. Note that a probability forecast is perfect if and only if the forecast assigns probability one to the "state of the atmosphere" which subsequently obtains, i.e., if and only if the forecast is categorical and correct. However, precise definitions of the term "degree of perfection" are required for the description of attributes, as well as for the formulation of measures, each of which provides a complete, but, to a certain extent, arbitrary, degree-of-perfection scale. We undertake the development of such definitions in Chapter 3.

2.2.2 Operational Evaluation

A decision maker is herein an individual in a "meteorological" decision situation. That is, the decision maker is an individual (1) whose activities, or operations, are affected, to some extent, by meteorological conditions, i.e., by the state of the atmosphere, and (2) who must select a course of action from a set of available courses of action. The framework for decision making to which we subscribe is a variant of the subjective expected-utility (SEU) model of the decision situation (Fishburn, 1964, pp. 50-56), in which selected elements of the SEU model are subject to certain conditions. The SEU model and the conditional SEU model are considered, in some detail, in Chapter 4. In this subsection we are concerned with the identification of the decision maker's objectives, which necessitates the consideration of certain elements of the conditional SEU model.

The decision maker's objective is to select the "best" course of action. His preferences among the consequences, which result when he selects a particular course of action and a particular state of the atmosphere obtains, are expressed in terms of the "operational value," or utility, of the consequences to the decision maker. The decision maker is, of course, uncertain, prior to the selection of a course of action, which state of the atmosphere will obtain. We shall assume that the decision maker accepts, as his personal probabilities, the meteorologist's probabilities, i.e., the probability

forecast defined on the states of the atmosphere. Then, the decision maker selects that course of action for which the expected utility is a maximum. Thus, utility is the attribute of the forecasts which is indicative of the "value" of the forecasts (or of the parent forecasting system) to the decision maker. This form of evaluation, i.e., the process of determining the utility of forecasts, is referred to, in this study, as operational, or decision-theoretic, evaluation. We shall not attempt to specify any points on the utility scale at this time.

CHAPTER THREE. PROBABILITY FORECASTS: THEIR NATURE AND ATTRIBUTES

This chapter consists of two sections in which we discuss the nature and the attributes, respectively, of probability forecasts. Certain basic considerations of a probabilistic and meteorological nature are discussed in Section 3.1. From a probabilistic point of view, we are primarily concerned with the interpretation of the probabilities themselves, and we describe the subjective view of probability, the interpretation to which we subscribe in this study. From a meteorological point of view, we are primarily concerned with the definition of certain terms and the nature of certain primitive elements of the evaluation frameworks, and we define and/or describe these terms and primitive elements in this section.

Several "desirable" properties, or attributes, of probability forecasts are described in Section 3.2. The fact that we subscribe to the subjective view of probability suggests that normative as well as substantive attributes will be of interest. The normative attributes relate to the assessment process, i.e., the process of formulating the forecasts, while the substantive attributes relate to the evaluation process. The substantive attributes such as "accuracy" and "value" will, of course, be of primary concern. However, since we want a meteorologist's forecasts to meet both the normative and substantive standards of "goodness," the normative attributes must also be considered. The normative attribute "honesty," in particular, has important implications for the

evaluation process.

3.1 The Nature of Probability Forecasts

3.1.1 Probabilistic Considerations

The primary consideration from a probabilistic point of view is simply the manner in which the probabilities which constitute the forecasts are interpreted. Probabilities are, of course, subject to several different interpretations (see, for example, Fishburn, 1964; Good, 1959; Winkler, 1972). We subscribe to the subjective, or personal, view of probability, a view which was developed by de Finetti (1937) and Savage (1954), in which "probability measures the confidence that an individual has in the truth of a particular proposition, for example, the proposition that it will rain tomorrow" (Savage, op. cit., p. 3). Subjective probabilities are not, of course, assessed by individual whim or fancy, but rather are subject to certain rules of consistency. We consider these rules and their implications for probability assessment and evaluation in Section 3.2.

The subjective interpretation of probability appears to be consistent with probability forecasting as currently practiced in meteorology (e.g., Murphy and Winkler, 1971; Sanders, 1973). In fact, we believe that the subjective view of probability is particularly appropriate in meteorological context (e.g. Epstein, 1966; Murphy and Winkler, 1971; Roberts, 1968). In this regard, the

subjective interpretation of probabilities places no particular restriction on the manner in which a meteorologist formulates his forecasts. That is, the meteorologist may utilize "objective" as well as subjective procedures and/or information in the process of formulating his forecasts. However, we shall assume that each forecast expresses the judgment of some meteorologist and, as a result, consists of subjective probabilities.

3.1.2 Meteorological Considerations

3.1.2.1 Meteorologists

In this study the term "meteorologist" will be taken to mean an experienced weather forecaster. We shall refer to such individuals as either meteorologists or forecasters.

3.1.2.2 Forecasting Procedures

The nature of the procedures, or techniques, used to formulate the forecasts will not be of particular concern in this study. In particular, these procedures may be either "objective" or subjective in a meteorological sense. However, in accordance with the subjective view of probability, we shall assume that each forecast expresses the judgment of some meteorologist or forecaster.

3.1.2.3 Predictands

The individual or composite variables for which a meteorologist formulates probability forecasts will be referred to as predictand-variables, or simply as variables. These variables, whether

continuous or discrete, will be assumed to have been subjected to an N-fold partition, i.e., to have been divided into N mutually exclusive and collectively exhaustive "states of the atmosphere." Let $s = \{s_1,\ldots,s_N\}$ denote this set of N states. The manner in which these states of the atmosphere are defined is relevant for evaluation. That is, the results of an evaluation will, in general, depend upon the manner in which the states are defined. However, as Goodman and Kruskal (1954) and others have indicated, this situation should be considered to be desirable since careful attention should be given to the definition of the states in light of the use for which the results are intended.

A discrete predictand-variable may be ordered or unordered (a continuous predictand-variable is, by definition, ordered). The presence or absence of order in the predictand-variable may be relevant for evaluation when N, the number of states, is greater than two. While we shall be primarily concerned in this study with situations in which the variable of concern is unordered or in which the presence or absence of order is not considered to be relevant, the implications of order in the predictand-variable for evaluation in general and for the formulation of evaluation measures for ordered predictand-variables in particular will be considered in Chapters 6, 8, and 9.

Meteorological forecasts are generally prepared for a particular predictand-element for a particular location (i.e., a point or an

area) for a particular period (i.e., an interval of time). A particular predictand-variable-location-period combination will be referred to as a predictand.

3.1.2.4 Forecasts

In this study a forecast will be a probability vector $\underline{r}$, where

$$\underline{r} = (r_1,\ldots,r_N),$$

in which

$$r_n \geq 0 \quad \text{for all } n \quad (n=1,\ldots,N)$$

and

$$\sum_{n=1}^{N} r_n = 1,$$

defined on the set $s = \{s_1,\ldots,s_N\}$ of N states of the atmosphere. Thus, r_n represents the meteorologist's (subjective) probability that state s_n obtains (n=1,...,N). We shall denote the set of N possible categorical forecasts by the vectors $\underline{r}_n$ (n=1,...,N), where $\underline{r}_1 = (1,0,0,\ldots,0)$, $\underline{r}_2 = (0,1,0,\ldots,0)$, ..., and $\underline{r}_N = (0,0,0,\ldots,1)$.

As indicated in Chapter 1, we are particularly concerned in this study with situations in which N, the number of states, equals two. In these two-state situations $\underline{r} = (r_1,r_2)$, where $r_2 = 1-r_1$, and

$0 \leq r_1, r_2 \leq 1.$

3.1.2.5 Observations

An observation will be a vector $\underline{d}$, where

$$\underline{d} = (d_1,\ldots,d_N),$$

defined on the set $s = \{s_1,\ldots,s_N\}$ of N states of the atmosphere, where d_n equals one if state s_n obtains and zero otherwise (n=1,...,N). We denote the set of N possible observations by the vectors $\underline{d}_n$ (n=1,...,N), where $\underline{d}_1 = (1,0,0,\ldots,0)$, $\underline{d}_2 = (0,1,0,\ldots,0)$, ..., and $\underline{d}_N = (0,0,0,\ldots,1)$. In two-state situations $\underline{d} = (d_1,d_2)$, where $d_2 = 1 - d_1$.

3.2 The Attributes of Probability Forecasts

3.2.1 Normative Attributes: Description and Measurement

The normative standards of "goodness" relate to the theory of subjective probability. According to this theory, an individual's (subjective) probabilities assessed in accordance with certain plausible postulates of coherence, or consistency, must conform (in a mathematical sense) to a probability measure (e.g., Winkler, 1967a). The theory of subjective probability, in addition, prescribes that an individual's statements should correspond (exactly) to his true beliefs or judgments, since the statements represent a quantification

of these judgments. Thus, a "perfect" assessor, in the normative sense, issues statements which (1) obey the postulates of consistency, i.e., are consistent, and (2) correspond to his judgments, i.e., are "honest" (cf. Winkler and Murphy, 1968). The "desirable" properties, or attributes, of the forecasts of concern from a normative point of view are, then, consistency and honesty.

3.2.1.1 Consistency

Consider a meteorologist who must make a probability forecast for a predictand with N mutually exclusive and collectively exhaustive states $\{s_1,\ldots,s_N\}$. Let the vector $\underline{p} = (p_1,\ldots,p_N)$ $(p_n>0, \Sigma p_n=1;$ $n=1,\ldots,N)$ denote the meteorologist's true beliefs, or judgments. Then, a meteorologist's probabilities are (completely) consistent if

$$p_n \geq 0 \quad \text{for all } n \quad \text{and} \quad \sum_n p_n = 1 \quad (n=1,\ldots,N) \qquad (3\text{-}1)$$

and

$$r_n \geq 0 \quad \text{for all } n \quad \text{and} \quad \sum_n r_n = 1. \quad (n=1,\ldots,N) \qquad (3\text{-}2)$$

That is, a meteorologist's judgment and forecast are consistent if the probabilities which constitute the judgment and forecast satisfy the familiar probability axioms. An evaluator who is interested in determining the consistency of a meteorologist's judgments and/or forecasts will be concerned with the nature and the extent of any

violations of the constraints in (3-1) and (3-2). However, the evaluator will not, in general, have access to the meteorologist's judgments. In this study, we shall assume that the probabilities which constitute a meteorologist's judgments and forecasts are always consistent. That is, we assume that a meteorologist who violates these constraints realizes his violation and revises his judgments and/or forecasts in order to satisfy the postulates of consistency.

3.2.1.2 Honesty and "Proper" and Improper Scoring Rules

A meteorologist's probabilities are completely honest if and only if

$$\underline{r} = \underline{p}, \tag{3-3}$$

i.e., if and only if the meteorologist's forecast, or statement, $\underline{r}$, corresponds (exactly) to his judgment, $\underline{p}$. Thus, an evaluator who is interested in determining the honesty of a meteorologist's statements will be concerned with the degree of association between his statements and his judgments. Since the evaluator will not, in general, have access to a meteorologist's judgments, he (i.e., the evaluator) cannot measure this degree of association directly. However, he can encourage the meteorologist to make his statements correspond to his judgments by rewarding (or penalizing) him in accordance with a "proper" scoring rule.

A scoring rule is, in general, and function of the meteorologist's (stated) probabilities, $(r_1,\ldots,r_N)$, and the state (of the atmosphere) that obtains, d_j (say), and, as such, is simply another name for an evaluation measure. Let $S_j(\underline{r})$ denote the score assigned by a scoring rule S to the forecast $\underline{r}$ when state s_j obtains (i.e., $d_j = 1$). Then, the meteorologist's subjective expected score is $S(\underline{r},\underline{p})$ where

$$S(\underline{r},\underline{p}) = \sum_{j=1}^{N} p_j S_j(\underline{r}). \tag{3-4}$$

Now, if

$$S(\underline{p},\underline{p}) > S(\underline{r},\underline{p}) \text{ for all } \underline{r} \neq \underline{p}, \tag{3-5}$$

then S is a strictly proper scoring rule, while if

$$S(\underline{p},\underline{p}) \geq S(\underline{r},\underline{p}) \text{ for all } \underline{r}, \tag{3-6}$$

then S is (only) a proper scoring rule. Thus, a strictly proper scoring rule encourages a meteorologist to be completely honest, i.e., to make his forecast correspond exactly to his judgment, since the meteorologist can maximize his expected score if and only if he sets $\underline{r}$ equal to $\underline{p}$.[1] The meteorologist also maximizes his expected score

---

[1] In this study we shall assume that the meteorologist's utility function is linearly related to the scoring rule of concern. The effects of non-linear utility functions in this context have been investigated by Winkler (1969) and Winkler and Murphy (1970) among others.

by setting $\underline{r}$ equal to $\underline{p}$ in the presence of a proper scoring rule. However, forecasts $\underline{r} \neq \underline{p}$ may also realize this same maximum expected score. We shall occasionally use the term "proper" (in quotation marks) to describe scoring rules that are either strictly proper or proper. A scoring rule which does not satisfy either (3-5) or (3-6) will be referred to as an improper scoring rule.

Thus, from the point of view of probability assessment, strictly proper and proper scoring rules are of particular interest. In this study we shall assume, unless otherwise noted, that meteorologists formulate their forecasts in the presence of strictly proper or proper scoring rules, and that their statements correspond (exactly) to their judgments. The use of these scoring rules in the assessment process has certain implications for the measures (of the attributes) that are used in the evaluation process, and we consider these implications in Chapter 6.

3.2.2 Substantive Attributes: Description

The substantive standards of "goodness," as indicated in Chapter 2, relate to the contexts in which the forecasts are formulated and utilized. Specifically, an evaluator will be concerned with the "goodness" of forecasts from two points of view: (1) a meteorological, or inferential, point of view and (2) an operational, or decision theoretic, point of view. Evaluation from a meteorological point of view will be referred to as empirical, or inferential evaluation, while evaluation from an operational point of view will be referred

to as operational, or decision theoretic, evaluation. The general attributes of concern in empirical and operational evaluation are "accuracy" and "value," respectively, and we examine these attributes in some detail in this subsection. The measurement of these attributes is considered in Chapters 6 and 7, respectively.

3.2.2.1 Empirical Evaluation

We have heretofore used the term "accuracy" to describe the general attribute (of the forecasts) of concern in empirical evaluation, namely the degree of association between forecasts and observations. However, in the meteorological literature, this term has been, and still is, generally associated with the "goodness" of categorical forecasts. In order to emphasize the fact that we are concerned with the "goodness" of probability forecasts (as well as for other reasons which will become apparent shortly), we will hereafter use the term "validity" to describe this general attribute.

The general attribute validity appears to possess two particularly relevant interpretations from a meteorological point of view. These interpretations lead to the identification of two specific attributes: (1) primary validity, an attribute concerned with the association between forecasts and observations on an individual basis and (2) secondary validity, an attribute concerned with the association between forecasts and observations on a collective basis. Within the context of empirical evaluation, we shall be primarily concerned with these two attributes of the forecasts, and

we shall refer to these attributes as the principal attributes.

We also consider several attributes suggested heretofore by other investigators and describe the relationships between these attributes and the principal attributes. In addition, we discuss the relevance of some of these supplemental attributes for empirical, or inferential, evaluation.

(1) Principal Attributes. As indicated above, the attributes of primary concern in this study are the principal attributes, primary and secondary validity, and we describe these attributes in some detail here.

(a) Primary validity. The degree of association between an individual forecast $\underline{r}$ and the relevant observation $\underline{d}$ is complete if and only if

$$r_n - d_n = 0 \quad \text{for all } n, \quad (n=1,\ldots,N) \tag{3-7}$$

i.e., if and only if the values assumed by the components of the vectors $\underline{r}$ and $\underline{d}$ are identical. A forecast $\underline{r}$ for which (3-7) holds will be referred to as a completely valid forecast in a primary sense. Such a forecast is categorical and correct (one possibility). Conversely, a forecast $\underline{r}$ for which

$$|r_n - d_n| = 1 \quad \text{for some } n, \quad (n=1,\ldots,N) \tag{3-8}$$

will be referred to as a completely invalid forecast in a primary

sense. Such a forecast is categorical and incorrect (N-1 possibilities). A probability forecast is, in general, neither completely valid nor completely invalid in a primary sense, i.e., in general,

$$0 < |r_n - d_n| < 1 \quad \text{for some } n. \quad (n=1,\ldots,N) \qquad (3\text{-}9)$$

A forecast $\underline{r}$ for which (3-9) holds will be referred to as a partially valid forecast in a primary sense.

Since one and N-1 possibilities exist, in general, for a forecast to be completely valid and invalid, respectively, only one reference point can be assigned to the primary validity scale. In two-state (N=2) situations, however, only one possibility exists for a forecast to be completely invalid, and, as a result, two reference points can be assigned to the primary validity scale in such situations.

(b) Secondary validity. The degree of association between a collection of forecasts and the relevant collection of observations depends upon the "statistic" of these collections which is selected as the basis for the comparison. Several such statistics can, of course, be identified. In this study we shall consider only the mean, a familiar member of the class of statistics concerned with location. Consideration of the mean leads to the identification of a general attribute of the forecasts which appears to possess a "reasonable" interpretation in a meteorological context. Thus, the association between forecasts and observations on a collective basis

is defined in terms of the association between the average forecast probabilities and the observed relative frequencies. Further consideration of the association between such average values leads to the identification of two specific attributes which are, in reality, simply two aspects of the familiar concept of bias.

(i) Bias in-the-large. Consider a collection of K forecasts $\underline{r}_k = (r_{1k},\ldots,r_{Nk})$, where $r_{nk} \geq 0$ and $\sum_n r_{nk} = 1$ and in which r_{nk} is the forecast probability associated with state s_n on occasion k, and the collection of the K relevant observations $\underline{d}_k = (d_{1k},\ldots,d_{Nk})$, where d_{nk} equals one if state s_n obtains on occasion k and zero otherwise (n=1,...,N;k=1,...,K). The degree of association between the forecasts in such a collection and the relevant observations will be considered to be complete, on a collective basis, for state s_n (n=1,...,N) if

$$\bar{r}_n - \bar{d}_n = 0, \tag{3-10}$$

where

$$\bar{r}_n = (1/K) \sum_{k=1}^{K} r_{nk} \tag{3-11}$$

and

$$\bar{d}_n = (1/K) \sum_{k=1}^{K} d_{nk}. \tag{3-12}$$

That is, the association is complete if the average forecast probability, $\bar{r}_n$, and the observed relative frequency, $\bar{d}_n$, are equal. A collection of forecasts for which (3-10) holds will be referred to as a completely unbiased collection in-the-large for state s_n. A collection for which (3-10) holds for all N states will be referred to as a completely unbiased collection in-the-large. Of course, a collection of forecasts will seldom if ever be completely unbiased in-the-large even for one state, since equality in (3-10) will seldom if ever be realized for such a collection. Thus, a collection of forecasts will be considered to be unbiased in-the-large for state s_n (n=1,...,N) if

$$0 \leq |\bar{r}_n - \bar{d}_n| \leq \varepsilon, \qquad (3\text{-}13)$$

where ε is a small positive number. In order to be meaningful, the value of ε should be related to the standard error of the quantity $|\bar{r}_n - \bar{d}_n|$. A collection of forecasts for which (3-13) holds for all N states will be referred to as an unbiased collection in-the-large (for that value of ε). The smaller ε, of course, the less likely a particular collection of forecasts is to be unbiased in-the-large for one state or for all N states. A collection of forecasts for which (3-13) does not hold for all N states will be referred to as a biased collection in-the-large (for that value of ε).

In the two-state (N=2) situation, a collection of forecasts which

is completely unbiased in-the-large for one state is completely unbiased in-the-large for both states and for the collection itself. Further, a collection which, from (3-13), is unbiased (biased) in-the-large for one state is unbiased (biased) in-the-large for both states (and for the collection itself for that value of ϵ).

(ii) Bias in-the-small. Suppose that the probabilities r_{nk} can assume only a finite set of S distinct values r^s, i.e., $r_{nk} = r^s$ for all n and k and for some s (n=1,...,N;k=1,...,K;s=1,...,S). Then, we can identify a finite set of T distinct forecasts $\underline{r}^t$, where $\underline{r}^t = (r_1^t,...,r_N^t)$ and for which $r_n^t = r^s$ for all n and t and for some s (n=1,...,N;t=1,...,T;s=1,...,S). If we assume that the values of r^s are equally spaced on the closed unit interval [0,1] and that they include the end points of that interval, then

$$T = \sum_{s=1}^{S} \binom{N+s-4}{s-1} (S-s+1), \qquad (3\text{-}14)$$

where $\binom{x}{y} = x!/[y!(x-y)!]$ for $0 \leq y \leq x$, $\binom{x}{y} = 1$ for $x = -1$ and $y = 0$, and $\binom{x}{y} = 0$ otherwise. Thus, we can divide the collection of K forecasts $\underline{r}_k$ (k=1,...,K) into T subcollections, where subcollection t contains the K^t forecasts for which $\underline{r}_k = \underline{r}^t$ ($k=1,...,K^t; \sum_t K^t = K$; t=1,...,T).

The degree of association between the forecast $\underline{r}^t$ associated with subcollection t and the K^t relevant observations $\underline{d}_k^t$ will be considered to be complete, on a collective basis, for state s_n (n=1,...,N)

if

$$r_n^t - \bar{d}_n^t = 0, \tag{3-15}$$

where

$$\bar{d}_n^t = (1/K^t) \sum_{k=1}^{K^t} d_{nk}^t, \tag{3-16}$$

in which d_{nk}^t equals one if state s_n obtains on occasion k (actually on the kth of the K^t occasions for which $\underline{r}_k = \underline{r}^t$) and zero otherwise (k=1,...,K;t=1,...,T). A subcollection of forecasts for which (3-15) holds will be referred to as completely unbiased in-the-small for state s_n, while a subcollection for which (3-15) holds for all N states will be referred to as a completely unbiased subcollection in-the-small. A collection of forecasts for which (3-15) holds for all T subcollections will be referred to as completely unbiased in-the-small.

A subcollection of forecasts will seldom if ever be completely unbiased in-the-small even for one state, since equality in (3-15) will seldom if ever be realized for such a subcollection. In this study, a subcollection of forecasts will be considered to be unbiased in-the-small for state s_n (n=1,...,N) if

$$0 \leq |r_n^t - \bar{d}_n^t| \leq (1/2)\Delta \tag{3-17}$$

where Δ is the difference between adjacent values of r^s (s=1,...,S). A subcollection for which (3-17) holds for all N states will be referred to as an unbiased subcollection in-the-small. A collection of forecasts for which (3-17) holds for all T subcollections will be referred to as unbiased in-the-small. A subcollection (collection) of forecasts for which (3-17) does not hold for all N states (T subcollections) will be referred to as a biased subcollection (collection) in-the-small.

In the two-state (N=2) situation, a subcollection of forecasts which is completely unbiased in-the-small for one state is completely unbiased in-the-small for both states and for the subcollection itself. Further, a subcollection which, from (3-17), is unbiased (biased) in-the-small for one state is unbiased (biased) in-the-small for both states (and for the subcollection itself).

(2) Supplemental Attributes. Substantive attributes of probability forecasts have also been described by Bross (1953) and Sanders (1958, 1963) among others. We briefly examine the relationships between these attributes and the principal attributes and consider the relevance of these attributes for empirical, or inferential, evaluation.

(a) "Validity." The attribute "validity," as described by Bross (op. cit. pp. 47-48), is equivalent to the attribute bias in-the-small, as described above, except that Bross treats each probability as a distinct forecast. That is, Bross considers forecasts to be scalars, while we consider forecasts to be vectors. We shall

defer further discussion of this attribute until we discuss the attribute reliability.

(b) Sharpness. A forecast is sharp, according to Bross (op. cit., pp. 48-52), if the forecast "discriminates" among the relevant states. Specifically, a forecast $\underline{r}$ is considered to be completely sharp if

$$r_n = 0 \text{ or } 1 \quad \text{for all } n, \quad (n=1,\ldots,N) \tag{3-18}$$

i.e., if the forecast assigns a probability of unity to one of the N states, while such a forecast is completely lacking in sharpness if

$$r_n = 1/N \quad \text{for all } n, \quad (n=1,\ldots,N) \tag{3-19}$$

i.e., if the forecast assigns equal probability to each of the N states. Equations (3-18) and (3-19) suggest that the sharpness of a forecast may be related to the amount of information contained in the probabilities which constitute the forecast (see, for example, Khinchin, 1957; Shannon and Weaver, 1949). These equations also indicate that the attribute sharpness is concerned exclusively with the probabilities which constitute a forecast. As a result, sharpness (by itself) does not appear to be a particularly appropriate attribute for empirical evaluation, the purpose of which is to

determine the degree of association between forecasts and observations. However, this attribute could be relevant for empirical evaluation if considered in combination with other attributes.

(c) Reliability. The attribute reliability (Sanders, 1958) is equivalent to the attribute "validity" (Bross, op. cit.). Sanders, who also treats each probability as a distinct forecast, divides the collection of K vector, or NK scalar, forecasts into S subcollections, where subcollection s consists of those $(NK)^s$ (scalar) forecasts for which $r_{nk} = r^s$ for all n and k and some s ($n=1,\ldots,N; k=1,\ldots K; \sum_s (NK)^s = NK; s=1,\ldots,S$). A subcollection s of (scalar) forecasts is considered to be completely reliable (or "valid" in Bross' terminology) if

$$r^s - \bar{d}^s = 0, \tag{3-20}$$

where

$$\bar{d}^s = [1/(NK)^s] \sum_{k=1}^{K} \sum_{n=1}^{N} d_{nk}^s, \tag{3-21}$$

in which d_{nk}^s equals one if state s_n obtains on the kth occasion that $r_{nk} = r^s$. Reliable and unreliable subcollections and collections of forecasts can, of course, be identified in a manner similar to that which we have employed for the attribute bias in-the-small. However, since different subcollections are formed if forecasts are considered to be scalars rather than vectors, a one-to-one relationship does not exist between the attributes reliability and bias in-the-small. In this regard, Murphy (1972a,b) has recently

described the relationships between particular measures of these attributes (see Chapter 6). Since we believe that forecasts should be treated as vectors rather than scalars, we shall be concerned in this study with the (principal) attribute bias in-the-small.

(d) Resolution. The attribute resolution (Sanders, 1958) is concerned primarily with the observed relative frequencies. Specifically, the subcollection s of $(KN)^S$ scalar forecasts is completely resolved if

$$\bar{d}^s = 0 \text{ or } 1, \tag{3-22}$$

and the entire collection of KN forecasts is completely resolved if

$$\bar{d}^s = 0 \text{ or } 1 \quad \text{for all } s. \quad (s=1,\ldots,S) \tag{3-23}$$

Subsequently, Sanders (1963) used the term sharpness to describe this attribute of the forecasts. Clearly, however, this attribute is not equivalent (or even similar) to the attribute sharpness as defined by Bross (op. cit.). The attribute resolution can also be defined for vector forecasts (Murphy, 1972a,b, 1973b). Specifically, a subcollection t of K^t vector forecasts is completely resolved if

$$\bar{d}^t_n = 0 \text{ or } 1 \quad \text{for all } n, \quad (n=1,\ldots,N) \tag{3-24}$$

and the entire collection of K forecasts is completely resolved if

$$\bar{d}_n^t = 0 \text{ or } 1 \quad \text{for all n and t.} \quad (n=1,\ldots,N;t=1,\ldots,T) \quad (3\text{-}25)$$

We can, of course, formulate definitions of resolved and unresolved subcollections and collections of both scalar and vector forecasts. As in the case of the attributes bias-in-the-small and reliability, a one-to-one-relationship does not exist between the definition of the attribute resolution for scalar forecasts and the definition of this attribute for vector forecasts. We believe that the fact that the attribute resolution is concerned only indirectly with the forecasts (through the specification of the subcollections) suggests that this attribute is not a particularly appropriate attribute for empirical evaluation.

3.2.2.2 Operational Evaluation and Utility

As indicated in Chapter 2, the general attribute of the forecasts of concern in operational evaluation is their "value." Specifically, we are concerned with the "value" of forecasts to decision makers, i.e., to individuals who use these forecasts within the context of decision situations. The framework for decision making to which we subscribe in this study is the subjective expected-utility (SEU) model of a decision situation (see, for example, Fishburn, 1964; the SEU model is described in Chapter 4). Within this framework, the "values" of the consequences, each of which is associated with

a particular course of action and a particular state of nature, to the decision maker are expressed in terms of their utility. The utilities, then, describe the decision maker's preferences for the consequences, and the decision maker is assumed to want to choose that course of action which maximizes his expected utility, where the expectation is taken with respect to the probabilities defined on the states. Since the meteorologist's forecast provides the decision maker with this probabilistic information, utility would appear to be a desirable attribute of the forecasts. This attribute, in reality, describes the extent to which the meteorologist's forecasts satisfy the decision maker's own criteria. In this study, then, the specific attribute of the forecasts of concern in operational evaluation will be their utility.

3.2.3 The Normative and Substantive Attributes of Primary Concern

In subsections 3.2.1 and 3.2.2 we have identified and described a number of desirable attributes of probability forecasts. Four of these attributes appear to be particularly important within the context of the evaluation problem as formulated in this study, and hereafter we shall focus our attention upon these attributes. From a normative point of view, the attribute of primary concern is honesty. As a result, evaluation measures which are strictly proper or proper scoring rules will be of particular interest. From a substantive point of view, the principal attributes primary validity and bias-in-the-small are of primary concern in empirical

evaluation. Hereafter, we shall refer to these attributes simply as validity and bias (or reliability), respectively. In operational evaluation, the attribute utility is, of course, the only relevant attribute.

CHAPTER FOUR. EVALUATION FRAMEWORKS

This chapter contains descriptions, in both algebraic and geometric terms, of frameworks for the representation and evaluation of probability forecasts. The framework for the representation of such forecasts, which serves in part as a foundation for the evaluation frameworks, is presented in Section 4.1, while frameworks for empirical and operational evaluation are described in Sections 4.2 and 4.3, respectively. The framework for operational evaluation, which is also based in part upon a modified version of the subjective expected-utility model for decision making, is shown to represent a comprehensive framework for evaluation. This framework, then, provides a basis for studies of the relationships between empirical and operational evaluation.

4.1 A Framework for the Representation of Probability Forecasts

Let R^N denote the set of all possible forecasts $\underline{r} = (r_1,\ldots,r_N)$. Then,

$$R^N = \{(r_1,\ldots,r_N) \mid r_n \geq 0, \textstyle\sum_n r_n = 1; n=1,\ldots,N\}. \qquad (4\text{-}1)$$

Hereafter, $\underline{r} \in R^N$ will denote that the forecast $\underline{r}$ is a member of the set R^N. The set R^N represents a regular (N-1)-dimensional simplex, in which each point corresponds to a particular forecast (and vice versa). A regular simplex is a line segment in the two-state

(N=2) situation, an equilateral triangle in the three-state (N=3) situation, and a regular tetrahedron in the four-state (N=4) situation (see Pontryagin, 1952). The sets R^2 and R^3 are depicted in a rectangular coordinate system in Figure 4-1. The end points of the line segment R^2 and the vertices of the triangle R^3 represent the two and three possible categorical forecasts in the two-state and three-state situations, respectively. In this coordinate system, the (Euclidean) distance between two points (or forecasts) $\underline{r}$ and $\underline{r}^*$ is $D(\underline{r},\underline{r}^*)$, where

$$D(\underline{r},\underline{r}^*) = (\underline{r}-\underline{r}^*)(\underline{r}-\underline{r}^*)' = [\sum_{n=1}^{N}(r_n - r_n^*)^2]^{1/2}. \quad (4\text{-}2)$$

The set R^N can also be depicted within a "content" coordinate system, in which the distances between the point which represents a forecast and the "faces" of the (N-1)-dimensional simplex are the N components of the forecast. The sets R^2 and R^3 are depicted in "length" and "area" coordinate systems, respectively, in Figure 4-2. In this coordinate system, R^2 is a unit line segment and R^3 is an equilateral triangle with unit height. As a result, the (Euclidean) distance between two points $\underline{r}$ and $\underline{r}^*$ in this coordinate system is $[(N-1)/N]^{1/2}D(\underline{r},\underline{r}^*)$.

The primitive elements of the framework for representation are, then, the meteorologist, or forecaster, who formulates the forecasts, the set $s = \{s_1,\ldots,s_N\}$ of N mutually exclusive and collectively exhaustive states of the atmosphere, and the forecast vector

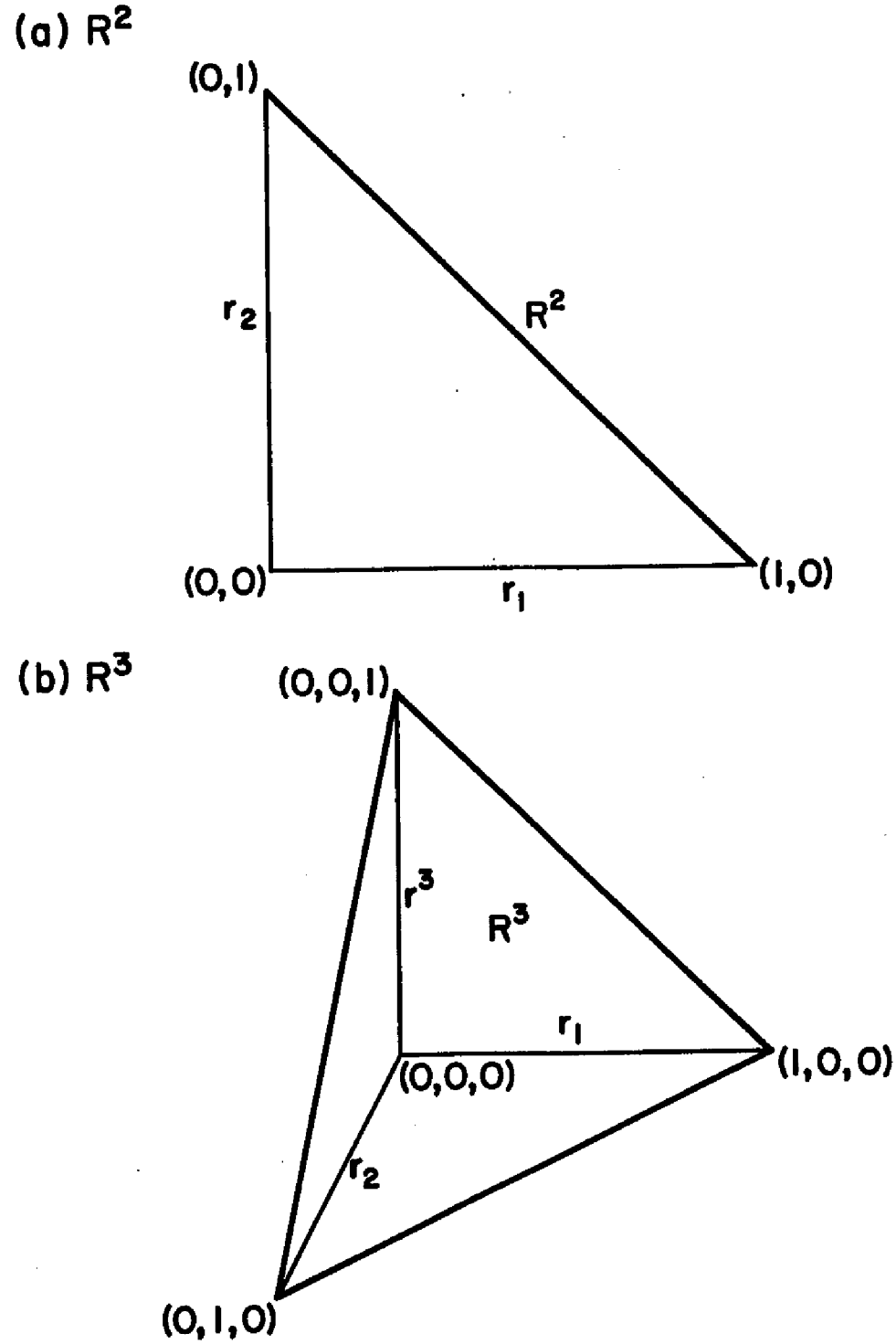


Figure 4-1. The sets (a) R^2 and (b) R^3 in a rectangular coordinate system.

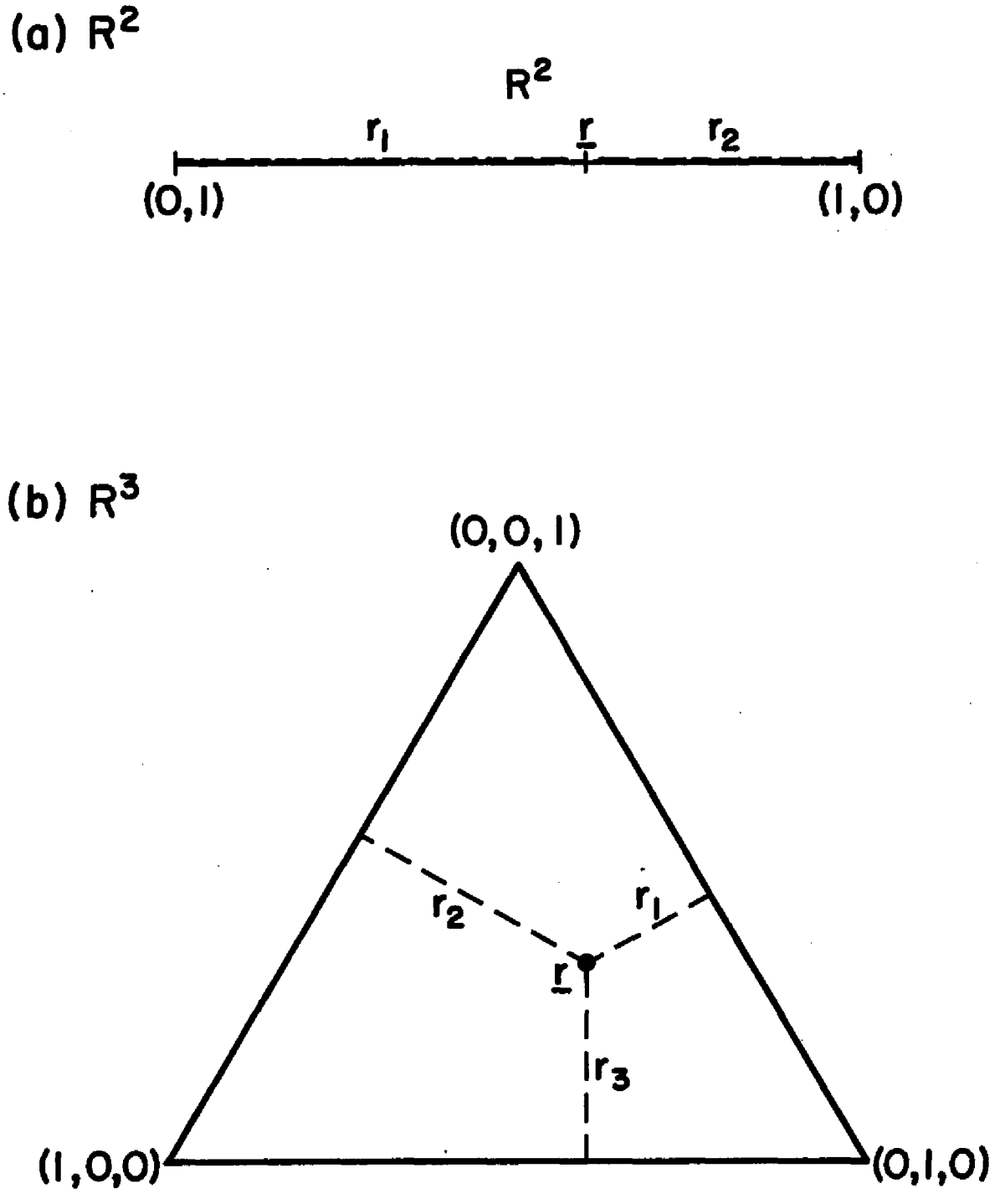


Figure 4-2. The sets (a) R^2 and (b) R^3 in length and area coordinate systems, respectively.

$\underline{r} = (r_1,\ldots,r_N)$ $(r_n \geq 0, \Sigma_n r_n = 1; n=1,\ldots,N)$. In general, we shall depict this framework in a content coordinate system.[1]

4.2 A Framework for Empirical Evaluation

Empirical evaluation, as described in Chapters 2 and 3, is concerned with the degree of association between forecasts and observations. Now, an observation is a vector $\underline{d} = (d_1,\ldots,d_N)$, where d_n equals one if state s_n obtains and zero otherwise, and the vector $\underline{d}_n$ denotes the observation for which $d_n = 1$ $(n=1,\ldots,N)$. Thus, the "vertices" of the (N-1)-dimensional simplex, which constitutes a framework for the representation of probability forecasts, represent the N possible observations $\underline{d}_n$ as well as the N possible categorical forecasts $\underline{r}_n$ $(n=1,\ldots,N)$. This "augmented" framework, then, is a framework for empirical evaluation and is depicted in Figure 4-3 for the two-state (N=2) and three-state (N=3) situations in length and area coordinate systems, respectively. Therefore, the primitive elements of the framework for empirical evaluation are the forecaster, the set $s = \{s_1,\ldots,s_N\}$ of N mutually exclusive and collectively exhaustive states of the atmosphere, the forecast vector $\underline{r} = (r_1,\ldots,r_N)$ $(r_n \geq 0, \Sigma_n r_n = 1; n=1,\ldots,N)$, and the observation vector $\underline{d} = (d_1,\ldots,d_N)$ (where $d_n = 1$ if state s_n obtains and zero otherwise).

---

[1] This coordinate system is often referred to as a barycentric coordinate system (e.g., Pontryagin, 1952, p. 10).

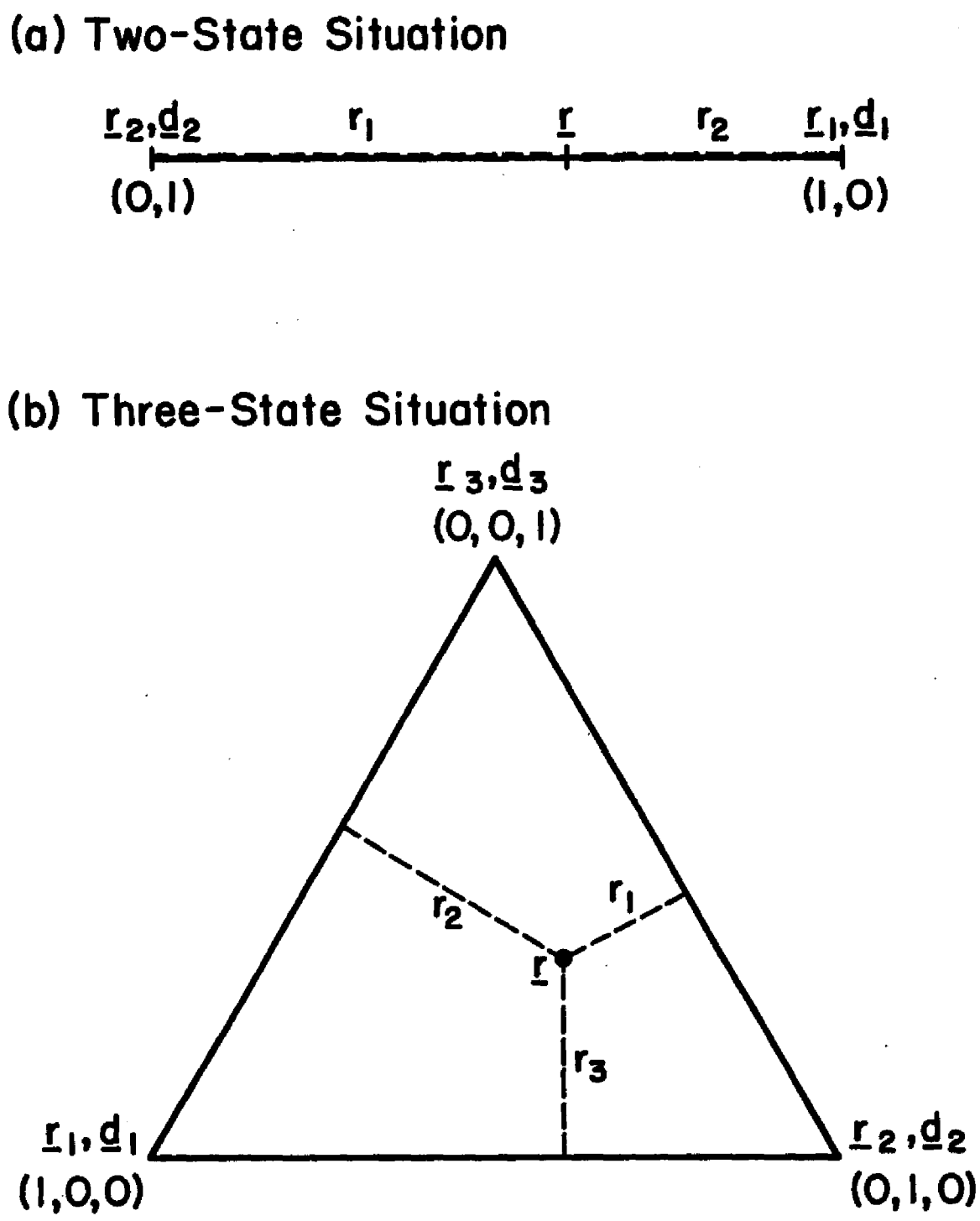


Figure 4-3. The framework for empirical evaluation in the (a) two-state and (b) three-state situations.

4.3 A Framework for Operational Evaluation

As indicated in Chapters 2 and 3, the decision maker, i.e., the individual who uses the forecasts, becomes an important element in the evaluation process when we are concerned with operational evaluation. In particular, any framework for operational evaluation must necessarily be based, at least in part, upon a model for decision making. The decision-making model, upon which our framework for operational evaluation is based, is described in subsection 4.3.1, and the framework itself is presented in subsection 4.3.2.

4.3.1 A Model for Decision Making

The model for decision making, upon which the framework for operational evaluation is based, is a modified version of the subjective expected-utility (SEU) model of a decision situation (see, for example, Fishburn, 1964, pp. 50-56). This subsection contains (1) a description of the basic SEU model in terms of the primitive elements which constitute the model, (2) a brief critique of the SEU model, and (3) a description of the modified SEU model in terms of the assumptions which we have made relative to certain elements in the basic SEU model.

4.3.1.1 The SEU Model: A Description

The primitive elements of the SEU model of a decision situation are a decision maker, a set of actions, a set of states of nature, a set of consequences, a matrix of utilities, a vector of probabilities,

and a decision criterion. The nature of these elements may be described briefly as follows:

(1) The decision maker is an individual in a decision situation who must select an action from a set of (available) actions in order to obtain a consequence from a set of consequences. The situation is such that, prior to the selection of an action, the decision maker is uncertain as to which action best serves his objectives, and he wishes to select the action which presents him with the best opportunity to realize these objectives.

(2) The set of actions consists of all of the courses of action, or alternatives, available to the decision maker in the pursuit of his objectives. We shall assume that the set of actions available to the decision maker can be partitioned into two mutually exclusive subsets, the subset $a = \{a_1,\ldots,a_M\}$ of M admissible actions and the subset of inadmissible actions. An available action is admissible if and only if the decision maker believes that the action would be selected under some conditions (a more precise definition is provided below). An (available) action that is not admissible is inadmissible. Hereafter, we shall refer to the subset $a = \{a_1,\ldots,a_M\}$

as the set of admissible actions.

(3) The set of states of nature is a set $e = \{e_1,\ldots,e_{N'}\}$ of N' mutually exclusive and collectively exhaustive states of nature, or events, which describe the relevant aspects of the "environment" within which the decision of concern is to be made.

(4) The set of consequences is a set $O = \{o_1,\ldots,o_{1N'},\ldots, o_{M1},\ldots,o_{MN'}\}$ of MN' consequences, or outcomes, where $o_{mn'}$ is the consequence that results when the decision maker selects action a_m and state $e_{n'}$ obtains ($m=1,\ldots, M;n'=1,\ldots,N'$).

(5) The matrix of utilities is a matrix $\underline{U} = (u_{mn'})$ of MN' utilities, or values, defined on the consequences, where $u_{mn'}$ is the utility of consequence $o_{mn'}$, ($m=1,\ldots, M;n'=1,\ldots,N'$). In the SEU model, the utilities express the decision maker's preferences among the consequences.

(6) The vector of probabilities is a vector $\underline{q} = (q_1,\ldots,q_{N'})$ of N' probabilities, defined on the states of nature, where $q_{n'}$ is the probability of state $e_{n'}$, ($q_{n'}\geq 0$, $\sum_{n'} q_{n'}=1;n'=1,\ldots,N'$). In the SEU model, a probability is a subjective, or personal, probability and, as such, expresses the decision maker's "degree of belief" concerning the occurrence of the state (or event) in question.

(7) The decision criterion, in the SEU model, is the maximization of (subjective) expected utility, and the expected utility of action a_m is U_m, where

$$U_m = \sum_{n'=1}^{N'} q_{n'} u_{mn'}. \quad (m=1,\ldots,M) \tag{4-3}$$

According to this criterion, the decision maker selects the action a_{m^*} for which $U_{m^*} \geq U_m$ for all $m \neq m^*$ $(m,m^*=1,\ldots,M)$.

A precise definition of an admissible action can be formulated in terms of the expected utility of an action [see (4-3)] and the SEU decision criterion. Specifically, an action a_m, is admissible if and only if, for some $\underline{q} \in R^N$,

$$U_{m'} > U_m \text{ for all } m \neq m'. \quad (m,m'=1,\ldots,M,M+1,\ldots)$$

The SEU model of a M-action, N'-state decision situation is depicted in Figure 4-4.

4.3.1.2 The SEU Model: A Brief Critique

The SEU model is, of course, only one of several possible models for decision making (c.f., Fishburn, 1964, pp. 36-76; Luce and Raiffa, 1957, pp. 275-326). Although such models differ, in general, in a number of respects, perhaps the most important difference is that these models are based upon different decision

Utility Matrix $\underline{U}$

		States						Expected Utilities
		e_1	$\cdots$	$e_{n'}$	$\cdots$	$e_{N'}$		
	a_1	u_{11}	$\cdots$	$u_{1n'}$	$\cdots$	$u_{1N'}$	$U_1 =$	$\sum_{n'=1}^{N'} q_{n'} u_{1n'}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$		
Actions	a_m	u_{m1}	$\cdots$	$u_{mn'}$	$\cdots$	$u_{mN'}$	$U_m =$	$\sum_{n'=1}^{N'} q_{n'} u_{mn'}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$		
	a_M	u_{M1}	$\cdots$	$u_{Mn'}$	$\cdots$	$u_{MN'}$	$U_M =$	$\sum_{n'=1}^{N'} q_{n'} u_{Mn'}$
Probability Vector	$\underline{q}$	q_1	$\cdots$	$q_{n'}$	$\cdots$	$q_{N'}$		

Figure 4-4. The SEU model of a M-action, N'-state decision situation.

criteria. The use of a particular criterion is, in turn, related to the decision maker's interpretation of the probabilities. For example, if the decision maker subscribes to the relative frequency interpretation of probability, then he may utilize a minimax or some other classical decision criterion (at least in many situations). On the other hand, if the decision maker subscribes to the subjective or personal interpretation of probability, then he will adopt the SEU, or maximization of expected utility, decision criterion (see, for example, Fishburn, 1964, p. 66). Since the subjective interpretation of probability appears to be more appropriate in a meteorological context (see Chapter 3), and since we subscribe to this interpretation, the SEU decision criterion will be used in this study.

While the framework and measures for operational evaluation to be described herein (see subsection 4.3.2 and Chapter 7, respectively) are formulated within the context of the SEU model, we believe that many (if not all) of the results of this study are applicable in other contexts as well. Of course, the interpretations placed upon the results may differ from context to context.

4.3.1.3 A Modified SEU Model

We are concerned, in this chapter, with the formulation of a framework for the evaluation of the probabilities which constitute a meteorologist's forecast. In order to evaluate these probabilities within a framework for operational evaluation which is based upon the SEU (or any other) decision-making model, the meteorologist's

probability forecast must be a primitive element of this model. The assumptions which we will make relative to the basic SEU model in order to obtain a suitable framework for operational evaluation can be described as follows:

(1) the decision maker's set $e = \{e_1,\ldots,e_N,\}$ of states of nature and the meteorologist's set $s = \{s_1,\ldots,s_N\}$ of states of the atmosphere will be assumed to be equivalent and will be referred to simply as the set $s = \{s_1,\ldots,s_N\}$ of states; and

(2) the decision maker is assumed to accept the probabilities which constitute the meteorologist's forecast as his subjective probabilities on the states, i.e., we assume that $q_n = r_n$ for all n (n=1,...,N), and we shall denote these probabilities by the vector $\underline{r} = (r_1,\ldots,r_N)$ $(r_n \geq 0, \Sigma r_n = 1; n=1,\ldots,N)$.[1]

Thus, the primitive elements of the modified SEU model are a meteorologist, a decision maker, a set a of (admissible) actions, a set

---

[1] Alternatively, we might assume that the decision maker assesses his own subjective probabilities on the states and then revises these probabilities in the light of meteorologist's forecast. However, since we are concerned in this study with the evaluation of the meteorologist's forecasts, we shall restrict our attention to the hypothetical, but hopefully not irrational, decision maker who accepts the forecaster's probabilities.

s of states, a set O of consequences, a matrix $\underline{U}$ of (the decision maker's) utilities defined on the consequences, a vector $\underline{r}$ of (the meteorologist's) probabilities defined on the states, and the maximum expected-utility decision criterion. The modified SEU model, the decision-making model upon which the framework for operational evaluation is based, is depicted in Figure 4-5.

4.3.2 The Framework for Operational Evaluation

The modified SEU model, in its entirety, can be depicted within the framework of the simplex R^N. The forecast $\underline{r}$ corresponds to a point in R^N and the vertices of R^N represent both the N possible categorical forecasts $\underline{r}_n$ and the N possible observations $\underline{d}_n$ (n=1,...,N). Further, the observation $\underline{d}_n$ defines the occurrence of the state s_n, since $d_n = 1$ if and only if state s_n obtains (n=1,...,N). Let R_m^N denote the set of all forecasts for which the decision maker prefers action a_m to the other M-1 actions. Then, the set R^N can be divided into M convex subsets R_m^N, where

$$R_m^N = \{(r_1,\ldots,r_N) \mid r_n \geq 0, \Sigma_n r_n = 1, U_m > U_{m'} \text{ for all } m' \neq m; \quad m,m' = 1,\ldots,M; n = 1,\ldots,N\}. \tag{4-4}$$

Thus, if $\underline{r} \in R_m^N$, then $U_m > U_{m'}$ for all $m' \neq m$ $(m,m' = 1,\ldots,M)$. The sets R_m^N and $R_{m'}^N$, $(m' \neq m)$ are separated by the (N-2)-dimensional indifference hyperplane $R_{mm'}^N$, where

Utility Matrix $\underline{U}$

		States				
		s_1	$\cdots$	s_n	$\cdots$	s_N	Expected Utilities
Actions	a_1	u_{11}	$\cdots$	u_{1n}	$\cdots$	u_{1N}	$U_1 = \sum_{n=1}^{N} r_n u_{1n}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	a_m	u_{m1}	$\cdots$	u_{mn}	$\cdots$	u_{mN}	$U_m = \sum_{n=1}^{N} r_n u_{mn}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	a_M	u_{M1}	$\cdots$	u_{Mn}	$\cdots$	u_{MN}	$U_M = \sum_{n=1}^{N} r_n u_{Mn}$
Probability Vector	$\underline{r}$	r_1	$\cdots$	r_n	$\cdots$	r_N	

Figure 4-5. The modified SEU model of a M-action, N-state decision situation.

$$R^N_{mm'} = \{(r_1,\ldots,r_N) \mid r_n \geq 0, \Sigma_n r_n = 1, U_m = U_{m'} \text{ for } m \neq m';$$

$$m,m' = 1,\ldots,M; n = 1,\ldots,N\}. \qquad (4\text{-}5)$$

This hyperplane represents the set of forecasts for which the decision maker is indifferent between actions a_m and $a_{m'}$, $(m,m' = 1,\ldots,M; m \neq m')$. Sets of forecasts for which the decision maker is indifferent among three or more actions can be defined in a similar manner. If we let $\bar{R}^N_m$ denote the set of forecasts for which none of the other M-1 actions is preferred to the action a_m, then

$$\bar{R}^N_m = \{(r_1,\ldots,r_N) \mid r_n \geq 0, \Sigma_n r_n = 1, U_m \geq U_{m'} \text{ for all } m' \neq m$$

$$m,m' = 1,\ldots,M; n = 1,\ldots,N\}. \qquad (4\text{-}6)$$

The set $\bar{R}^N_m$ represents the closure of the set R^N_m (R^N_m is an open set).

For illustrative purposes, we briefly consider the representation of two specific decision situations within the framework of the simplex R^N. A three-action, two-state decision situation is presented in Figure 4-6, and this situation is depicted within the framework of the simplex R^2 in a length coordinate system in Figure 4-7. The set R^2_m $(m = 1,2,3)$ denotes the set of forecasts for which the decision maker prefers action a_m to the other two actions. The indifference hyperplanes are points in the two-state situation.

Utility Matrix $\underline{U}$

		States		
		s_1	s_2	Expected Utilities
	a_1	4	-2	$U_1 = 4r_1 - 2r_2$
Actions	a_2	2	0	$U_2 = 2r_1$
	a_3	-2	2	$U_3 = -2r_1 + 2r_2$
Probability Vector	$\underline{r}$	r_1	r_2	

Figure 4-6. A specific three-action, two-state decision situation.

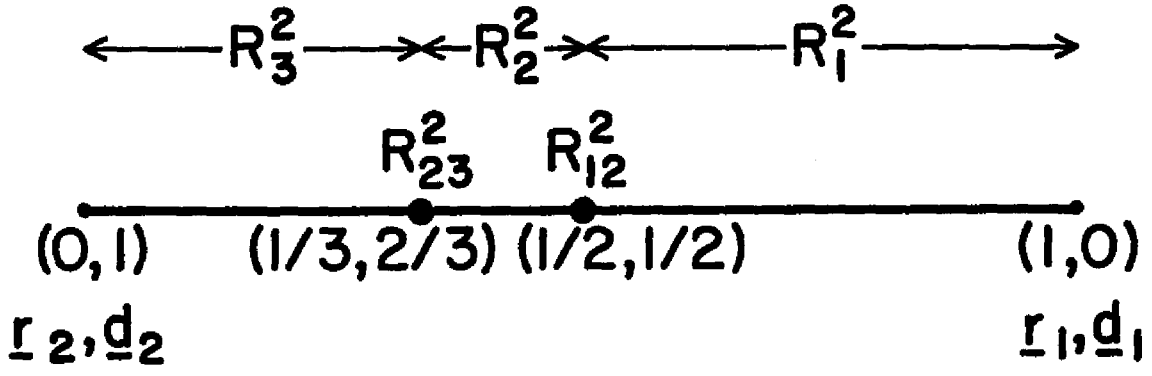


Figure 4-7. The framework for operational evaluation for the three-action, two-state decision situation described in Figure 4-6.

Specifically, the points R^2_{12} ($r_1 = 1/2$) and R^2_{23} ($r_1 = 1/3$) represent the forecasts for which the decision maker is indifferent between actions a_1 and a_2 and a_2 and a_3, respectively. Since $U_2 > U_m$ ($m=1,3$) for $1/3 < r_1 < 1/2$, the point R^2_{13} does not serve as a boundary point for the sets $\bar{R}^2_m$ ($m=1,2,3$). Note also that the end points of the unit line segment R^2 represent the categorical forecasts $\underline{r}_n$ [i.e., $\underline{r}_1 = (1,0)$ and $\underline{r}_2 = (0,1)$] and the observations $\underline{d}_n$ [i.e., $\underline{d}_1 = (1,0)$ and $\underline{d}_2 = (0,1)$] ($n=1,2$).

A four-action, three-state decision situation is presented in Figure 4-8, and this situation is depicted within the framework of R^3 in an area coordinate system in Figure 4-9. As in the two-state situation, R^3_m ($m=1,2,3,4$) denotes the set of forecasts for which the decision maker prefers action a_m to the other three actions. In a three-state situation, the indifference hyperplanes are the lines $R^3_{mm'}$, ($m,m'=1,2,3,4; m'>m$). Since U_m ($m=2,4$) > $U_{m'}$, ($m'=1,3$) for all $\underline{r} \in R^3_{13}$, the line R^3_{13} does not serve as boundary for the sets $\bar{R}^3_m$ ($m=1,2,3,4$). The points R^3_{124} and R^3_{234} represent forecasts for which the decision maker is indifferent among three actions (namely, a_1, a_2, and a_4 and a_2, a_3, and a_4, respectively). In a three-state situation, the vertices of the triangle represent the three categorical forecasts $\underline{r}_n$ and the three observations $\underline{d}_n$ ($n=1,2,3$).

The primitive elements of the framework for operational evaluation are, then, the primitive elements of the modified SEU model for decision making (see paragraph 4.3.1.3), together with the observation

Utility Matrix $\underline{U}$

		States			
		s_1	s_2	s_3	Expected Utilities
Actions	a_1	4	-2	0	$U_1 = 4r_1 - 2r_2$
	a_2	2	3	0	$U_2 = 2r_1 + 3r_2$
	a_3	0	2	1	$U_3 = 2r_2 + r_3$
	a_4	-1	0	2	$U_4 = -r_1 + 2r_3$
Probability Vector	$\underline{r}$	r_1	r_2	r_3	

Figure 4-8. A specific four-action, three-state decision situation.

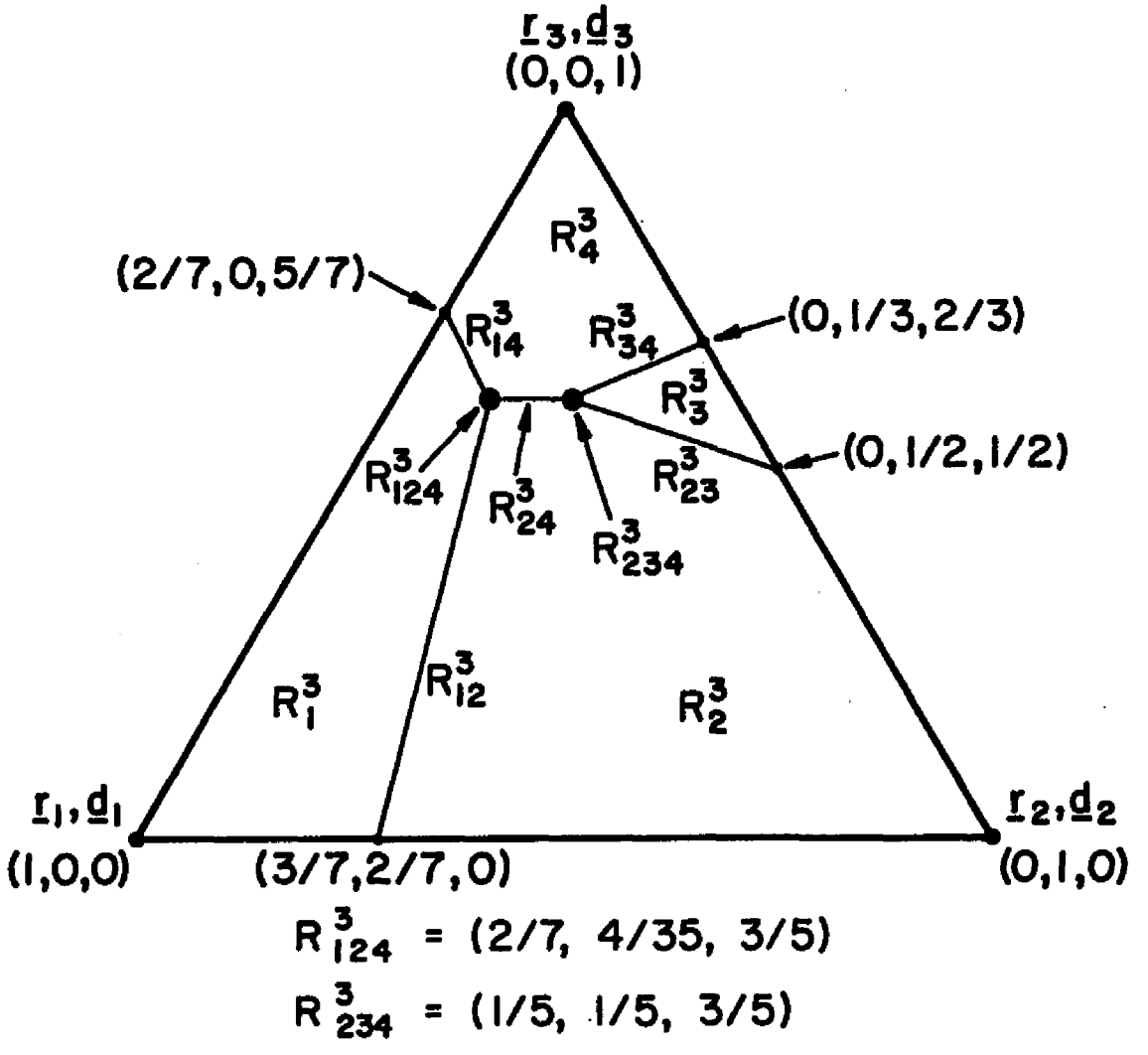


Figure 4-9. The framework for operational evaluation for the four-action, three-state decision situation described in Figure 4-8.

vector $\underline{d} = (d_1,\ldots,d_N)$, for which $d_n = 1$ if state s_n obtains and zero otherwise.

4.4 A Comprehensive Framework for Evaluation

The set of primitive elements of the framework for empirical evaluation is a subset of the set of primitives of the framework for operational evaluation. The latter, then, can be considered to represent a comprehensive framework for evaluation. The existence of such a framework provides a basis for studies of the relationships between empirical and operational evaluation, and some results of such studies are presented in Chapter 8. The problem of the specification (or measurement) of the primitive elements in the frameworks for empirical and operational evaluation is considered in Chapter 5.

CHAPTER FIVE. THE FRAMEWORKS: SPECIFICATION OF PRIMITIVE ELEMENTS

A comprehensive framework for evaluation such as that described in Chapter 4 is of little practical interest unless the evaluator, who may be a meteorologist, a decision maker, or some other individual (see Chapter 1), is able to specify the "values" of the primitive elements which constitute the framework. The form of this specification will depend upon the nature of the primitives themselves as well as upon the evaluator's "state of knowledge." In this chapter, we first classify the primitive elements, and then we examine those primitives which, for the purposes of evaluation, involve specification on a quantitative (i.e., numerical) scale. Finally, we describe the evaluator's assumed state of knowledge with respect to these primitives and present several classes of measure which we will use (as examples) in situations in which this knowledge is incomplete or is expressed in probabilistic terms.

5.1 The Primitives: Classification

The primitive elements which constitute the comprehensive framework for evaluation may be classified as either qualitative or quantitative. The class of qualitative primitives consists, in general, of all those primitives which involve specification on at most a nominal scale. The meteorologist, the decision maker, the set a of actions, the set s of states, the set O of consequences, and the decision criterion are members of this set. The class of

quantitative primitives consists, in general, of those primitives which involve specification on at least an ordinal scale, namely the forecast vector $\underline{r}$, the observation vector $\underline{d}$, and the utility matrix $\underline{U}$.[1] We examine the quantitative primitives in Section 5.2, with particular reference to the number and range of the "parameters" which require specification for each such primitive element. The evaluator's assumed state of knowledge with respect to both the qualitative and quantitative primitives is described in Section 5.3.

5.2 The Quantitative Primitives: Description

5.2.1 The Forecast (or Probabilities)

The meteorologist's forecast $\underline{r} = (r_1,\ldots,r_N)$ $(r_n \geq 0, \sum_n r_n = 1;$ $n=1,\ldots,N)$ is a probability vector defined on the set $s = \{s_1,\ldots,s_N\}$ of N states. Since the sum of the probabilities must be equal to one, the specification of the forecast, or the probabilities, involves specifying N-1 parameters, the range of each of which is the closed unit interval [0,1].

---

[1] In reality, the observation $\underline{d}$ involves specification on only a nominal scale. However, since the specification of the observation involves the use of a "numerical" scale, we consider this primitive element together with the other primitives involving numerical specification.

5.2.2 The Observation

The observation $\underline{d} = (d_1,\ldots,d_N)$ is a vector defined on the set $s = \{s_1,\ldots,s_N\}$ of N states, where d_n equals one if state s_n obtains and zero otherwise. Since $d_n = 1$ for some n and $d_{n'} = 0$ for all $n' \neq n$ $(n,n'=1,\ldots,N)$, the specification of the observation involves specifying N-1 parameters, the values of at most N-1 and one of which are zero and one, respectively.

5.2.3 The Utilities

The decision maker's utility matrix $\underline{U} = (u_{mn})$, as defined in Chapter 4, is a matrix of MN utilities defined on the set $O = \{o_{11},\ldots,o_{1N},\ldots,o_{M1},\ldots,o_{MN}\}$ of MN consequences. We shall assume that the evaluator, in each situation of concern, is able to identify the most and least preferred consequences, o^M and o^L, respectively. If we let u^M and u^L denote the utilities of the consequences o^M and o^L, respectively, then

$$u^L \leq u_{mn} \leq u^M. \quad (m=1,\ldots,M;n=1,\ldots,N) \qquad (5\text{-}1)$$

The utilities, as defined herein, are unique only up to a positive linear transformation, i.e., the origin and unit of measure on the utility scale are arbitrary (see, for example, Fishburn, 1964). Thus, from (5-1), the utility matrix $\underline{U} = (u_{mn})$ can be transformed into an equivalent <u>standard</u> utility matrix $\underline{U}' = (u'_{mn})$, where

$$u'_{mn} = \begin{cases} 1 & \text{if } u_{mn} = u^M, \\ 0 & \text{if } u_{mn} = u^L, \quad (m=1,\ldots,M;n=1,\ldots,N) \\ (u_{mn} - u^L)/(u^M - u^L) & \text{otherwise.} \end{cases} \tag{5-2}$$

Note that

$$0 \leq u'_{mn} \leq 1. \quad (m=1,\ldots,M;n=1,\ldots,N)$$

Thus, after the most and least preferred consequences are identified, the specification of the utilities, in general, involves specifying MN-2 parameters, the range of each of which is the closed unit interval [0,1]. Hereafter, we assume the utility matrix $\underline{U} = (u_{mn})$ is in standard form. In this regard, the standard utility matrices corresponding to the three-action, two-state and four-action, three-state decision situations depicted in Figures 4-6 and 4-8 are presented in Figures 5-1(a) and 5-1(b), respectively.

We shall denote the MN-2 utilities requiring specification in M-action, N-state decision situations by $x_1,\ldots,x_{MN-2}$ $(0 \leq x_1,\ldots, x_{MN-2} \leq 1)$ and the complete set of relevant utilities in such situations by u, where $u = \{1,x_1,\ldots,x_{MN-2},0\}$. Now, recall from Chapter 1, that we are concerned in particular with two-action (M=2), two-state (N=2) decision situations. The specification of the utilities in such situations involves specifying only two parameters and, the set of relevant utilities in such situations is

(a)

Utility Matrix $\underline{U}$

		States	
		s_1	s_2
Actions	a_1	1	0
	a_2	2/3	1/3
	a_3	0	2/3

(b)

Utility Matrix $\underline{U}$

		States		
		s_1	s_2	s_3
Actions	a_1	1	0	1/3
	a_2	2/3	5/6	1/3
	a_3	1/3	2/3	1/2
	a_4	1/6	1/3	2/3

Figure 5-1. The standard utility matrices corresponding to the decision situations depicted in (a) Figure 4-6 and (b) Figure 4-8.

$u = \{1, x_1, x_2, 0\}$. In Figure 5-2, we depict the twenty-four (4!) distinct utility matrices which can be formulated using the members of the set u in terms of six key utility matrices. The three other matrices associated with each key matrix (which together form a group of matrices) can be obtained by interchanging the rows and/or the columns in the key matrix. Since the other matrices in each group can be obtained by means of operations whose effects (if any) on decision making (and evaluation) can be readily determined, we shall hereafter be primarily concerned with the key matrix in each group. In this regard, since $0 \leq x_1, x_2 \leq 1$, action a_2 is inadmissible in the decision situations associated with key matrices (a) and (c) (see Figure 5-2). Thus, these two groups of utilities matrices will be of no further interest.

A special set of two-action, two-state decision situations of particular interest arises when the decision maker is assumed to be indifferent between the two consequences neither of which is the most or least preferred consequence, i.e., when x_1 and x_2 are assumed to be equal. In these situations $x_1 = x_2 = x$ $(0 \leq x \leq 1)$ and $u = \{1, x, x, 0\}$. The twelve (4!/2!) distinct utility matrices which can be formulated using the members of the set u are depicted in Figure 5-3 in terms of three key utility matrices. Since $0 \leq x \leq 1$, action a_2 is inadmissible in the decision situations associated with key matrices (a) and (b). Thus, these groups of matrices will be of no further interest.

(a)

$$\begin{pmatrix} 1 & x_1 \\ x_2 & 0 \end{pmatrix}$$

(b)

$$\begin{pmatrix} 1 & x_1 \\ 0 & x_2 \end{pmatrix}$$

(c)

$$\begin{pmatrix} 1 & x_2 \\ x_1 & 0 \end{pmatrix}$$

(d)

$$\begin{pmatrix} 1 & x_2 \\ 0 & x_1 \end{pmatrix}$$

(e)

$$\begin{pmatrix} 1 & 0 \\ x_1 & x_2 \end{pmatrix}$$

(f)

$$\begin{pmatrix} 1 & 0 \\ x_2 & x_1 \end{pmatrix}$$

Figure 5-2. The six key utility matrices which represent the twenty-four possible utility matrices in the general two-action, two-state decision situation.

(a)

$$\begin{pmatrix} 1 & x \\ x & 0 \end{pmatrix}$$

(b)

$$\begin{pmatrix} 1 & x \\ 0 & x \end{pmatrix}$$

(c)

$$\begin{pmatrix} 1 & 0 \\ x & x \end{pmatrix}$$

Figure 5-3. The three key utility matrices which represent the twelve possible utility matrices in the special two-action, two-state decision situation.

The special two-action, two-state decision situations discussed in the previous paragraph are of particular interest because the situations described by the key utility matrix (c) in Figure 5-3 are equivalent, in a decision-making sense, to the familiar cost-loss ratio decision situation (see, for example, Thompson, 1952; Thompson and Brier, 1955; Murphy, 1966). The latter situation is presented in Figure 5-4, in which C is the cost of protection and L is the loss if protective action is not taken and adverse weather subsequently occurs (C<L). We can transform the cost-loss matrix in Figure 5-4 into a "standardized" cost-loss matrix by dividing each element in the original cost-loss matrix by L. If, in addition, we assume that the decision maker's utility function is linear in expenses (i.e., in costs and losses), then the standard utility matrix corresponding to the original cost-loss matrix consists of the set u of utilities, where $u = \{1-(C/L), 1-(C/L), 0, 1\}$. Finally, if $x = 1 - (C/L)$, then this utility matrix is (decision) equivalent to the key utility matrix (c) in Figure 5-3. Thus, the discussion in Chapter 7 regarding operational evaluation measures associated with this key utility matrix also applies to the cost-loss ratio decision situation.

5.3 The Primitives: Specification

In this section we describe the evaluator's assumed state of knowledge with respect to both the qualitative and quantitative primitives. If the evaluator's knowledge is complete, then he is able to specify the primitive in a certain, or categorical, mode.

Cost-Loss Matrix

		State		
		Adverse Weather (W)	No Adverse Weather (NW)	Expected Expense (e)
Action	Protect (P)	C	C	$e_P = r_W C + r_{NW} C = C$
	Do Not Protect (DNP)	L	0	$e_{DNP} = r_W L$
	Probabilities	r_W	r_{NW}	

Figure 5-4. The cost-loss ratio decision situation.

On the other hand, if his knowledge is incomplete, then he is only able to provide an incomplete specification of the primitive or to specify the primitive in an uncertain, or probabilistic, mode.

5.3.1 Qualitative Primitives

We assume herein that the evaluator's state of knowledge with respect to the qualitative primitives is (always) complete. In essence, then, we are assuming that the evaluator is able to identify and describe the relevant members of the sets of actions, states, and consequences on each occasion. These assumptions should not be taken to mean that the process of specification, with respect to these primitives, is easy and devoid of uncertainty. On the contrary, the specification of the actions, states, and consequences is, in general, a difficult, time-consuming process often involving considerable uncertainty. However, this assumption helps to make the problem of concern in this thesis more tractable and, as a result, we are able to focus our attention upon measures of knowledge of the quantitative primitives, in particular the utilities.

5.3.2 Quantitative Primitives

5.3.2.1 The Forecast (or Probabilities)

Recall, from subsection 5.2.1, that the specification of the probabilities involves specifying N-1 parameters. When we refer to these probabilities as parameters (rather than as probabilities), we may be led to consider the possibility that the evaluator's

knowledge of these parameters may be incomplete. If, in addition, we assume that this knowledge is expressed in probabilistic terms, we are, in essence, concerned with second-order probabilities (i.e., with probabilities of probabilities). We have resisted the introduction of second-order probabilities into this study for several reasons:

(1) If an individual adopts the subjective expected-utility model of a decision situation, then the (first-order) probabilities are all that are required (where these probabilities are assumed to represent expectations over any higher-order distributions).

(2) As Savage (1954, pp. 57-60) and others (e.g., Winkler, 1972, pp. 97-98) have indicated, when second-order probabilities are considered the introduction of an endless hierarchy of higher-order probabilities seems inescapable. Such a hierarchy appears to be very difficult to interpret and, at best, to make the theory (of subjective probability) more difficult to interpret. Further, from a practical point of view, consideration of these second-order (and higher-order) probabilities is unnecessary, in terms of their impact upon the decision-making process.

(3) Since we are concerned herein with the evaluation of a meteorologist's probability forecasts (rather than with decision making <u>per se</u>), we believe that the assumption that these

forecasts are available to, and adopted by, both the decision maker and the evaluator is not unreasonable.

Thus, we shall assume herein that the evaluator's knowledge of the probabilities is complete.

5.3.2.2 The Observation

The specification of the observation involves (at most) specifying N-1 parameters. Now, observations of meteorological variables, in general, involve a degree of uncertainty (e.g., as to whether state s_n or state $s_{n'}$ actually obtained on a particular occasion). However, we believe that these uncertainties are of a different (considerably smaller) order of magnitude than the uncertainties inherent in, for example, the specification of the utilities. Thus, for the purposes of this study, we assume that the evaluator's knowledge of the observation is complete and, as a result, that he can specify the observation in categorical terms.

5.3.2.3 The Utilities

The specification of the utilities involves specifying (at most) MN-2 parameters $x_1,\ldots,x_{MN-2}$, where $0 \leq x_1,\ldots,x_{MN-2} \leq 1$ (after the most and least preferred consequences have been identified; see subsection 5.3.2). If the evaluator's knowledge is complete, then he is able to specify the values of the relevant utilities in a precise, or categorical, manner. On the other hand, if the evaluator's knowledge is incomplete, he is assumed to express

his knowledge in an imprecise and/or probabilistic manner. Specifically, we identify two general classes of measures of knowledge (or "knowledge measures"): (1) inequality measures (Class I) and (2) probability measures (Class II). These classes of measures, together with specific members of each class, are described below. The specific measures will be used in Chapters 7 and 8 to illustrate the formulation of operational evaluation measures in situations in which the evaluator's knowledge of the utilities is incomplete.

(1) Inequality measures (Class I). Inequality measures, which bound the values of the utilities, have been described in detail by Fishburn (1964). In this study we consider only one such measure, the bounded interval measure, which specifies upper and lower bounds on the parameters of concern. If, for example, the evaluator's knowledge consists of a bounded interval measure in the general two-action, two-state situation, then $\alpha_1 \leq x_1 \leq \beta_1$ and $\alpha_2 \leq x_2 \leq \beta_2$, where $0 \leq \alpha_i \leq \beta_i \leq 1$ (i=1,2). In this study the bounds on one parameter are assumed to be independent of the bounds on the other parameters. The inequality measures do not include a specification of the (probability) distribution of the utilities within these bounds.

(2) Probability measures (Class II). The probability measures define probability distributions on the relevant utilities. Such distributions can be given different interpretations depending upon whether one decision maker or a collection of decision makers is of concern. (a) One decision maker: The evaluator's knowledge of

the decision maker's utilities is incomplete (which may be due either to a lack of knowledge on the part of the evaluator concerning the decision maker's utilities even though the latter may be able to specify his utilities precisely or to a basic lack of knowledge on the part of the decision maker himself), and the probability measure describes his state of knowledge. (b) A collection of decision makers: The evaluator's knowledge may or may not be complete, but the probability measure describes the distribution of the relevant utilities over the collection of decision makers.

For the purposes of this study, we consider two subclasses of this class of measures: (a) non-negative probability measures (Class IIa) and (b) positive probability measures (Class IIb). The Class IIa measures are assumed to define an independent, non-negative (discrete or continuous) probability distribution $f(x_i)$ (≥ 0) on each utility x_i ($i=1,\ldots,MN-2$). The cumulative distribution functions $F(x_i)$ corresponding to these measures are non-decreasing functions. The Class IIb measures, on the other hand, are assumed to define an independent, positive (discrete or continuous) probability distribution $f(x_i)$ (>0) on each utility x_i ($i=1,\ldots,MN-2$). The corresponding cumulative distribution functions $F(x_i)$ are non-decreasing functions if $f(x_i)$ is discrete and increasing functions if $f(x_i)$ is continuous.[1]

As examples of these probability measures, we consider two specific families of distributions, the inequality-uniform measures and the beta measures. An inequality-uniform measure on the

---

[1] Hereafter, we shall use the terms "increasing" and "strictly increasing," respectively, to describe these properties of functions (see Chapters 7 and 8).

utilities defines a uniform probability distribution over the region defined by the corresponding inequality measure. Thus, when the evaluator's knowledge in the general two-action, two-state situation is expressed in terms of the inequality-uniform measure for which $\alpha_1 \leq x_1 \leq \beta_1$ and $\alpha_2 \leq x_2 \leq \beta_2$ $[0 \leq \alpha_i \leq \beta_i \leq 1\ (i=1,2)]$, the measure of concern is $f(x_1,x_2) = f_1(x_1)f_2(x_2)$, where

$$f_i(x_i) = \begin{cases} 1/(\beta_i - \alpha_i) & \alpha_i \leq x_i \leq \beta_i, \\ 0 & \text{otherwise} \end{cases} \tag{5-3}$$

$(i=1,2)$.

A beta measure on the utilities defines (independent) beta distributions on each of the relevant utilities. Thus, when the evaluator's knowledge of the utilities in the general two-action, two-state situation consists of a beta measure, the measure of concern is $f(x_1,x_2) = f_1(x_1)f_2(x_2)$, where

$$f_i(x_i) = \begin{cases} [(a_i + b_i - 1)!/(a_i - 1)!(b_i - 1)!]x_i^{a_i-1}(1-x_i)^{b_i-1} & \\ \qquad 0 \leq x_i \leq 1 \text{ and } a_i, b_i > 0, & \\ 0 \qquad \text{otherwise} & \end{cases} \tag{5-4}$$

$(i=1,2)$. We shall assume, in general, that the parameters a_i and b_i $(i=1,\ldots,MN-2)$ are positive integers, in which case definite

integrals involving these beta distributions can be expressed in terms of finite summations (see, for example, Raiffa and Schlaifer, 1961, pp. 216-220). Note that when the parameters a and b are both equal to one, the beta distribution in (5-4) reduces to the uniform distribution. That is, $\beta(1,1) = U(0,1)$.

Specific members of these classes of measures will be used for illustrative purposes in Chapters 7 and 8 to provide examples of utility and expected-utility (evaluation) measures.

CHAPTER SIX. EMPIRICAL EVALUATION MEASURES

Measures of the attributes of concern in empirical evaluation are described and compared in this chapter, with particular reference to measures of the principal attributes, namely validity and bias (see Chapter 3). Several basic concepts are considered in Section 6.1, including "scales of goodness," "proper" and improper measures, equivalent measures, and absolute and relative measures. In Section 6.2, measures of the principal (and several supplementary) attributes are formulated within the framework for empirical evaluation described in Chapter 4. These "distance" measures are the basic empirical evaluation measures of concern in this study. Other measures, proposed heretofore, are described and compared with the relevant distance measures in Section 6.3.

6.1 Some Basic Concepts

6.1.1 Scales of Goodness

Recall, from Chapters 2 and 3, that the processes of identifying both the purposes for evaluation and the desirable attributes of probability forecasts led to the consideration of "scales of goodness." Further, recall that these scales remain largely arbitrary in the case of empirical evaluation, even though the meteorologist is able to specify at least one point on most (if not all) of the relevant scales. On the other hand, the decision maker's utilities define

the appropriate scale of goodness (at least up to a positive linear transformation) in the case of operational evaluation. Since we are concerned with measures for empirical evaluation in this chapter, we must recognize that, in the absence of generally acceptable scales of goodness, the scores assigned by these measures involve a certain degree of arbitrariness. Specifically, each measure implicitly specifies a particular scale of goodness and different measures necessarily impose different scales. This arbitrariness can only be eliminated by the general acceptance, by all meteorologists, of a universal scale of goodness for each attribute of concern in empirical evaluation, an unlikely prospect at this time or in the immediate future. In the absence of such a universal scale, we will simply consider and compare the scales associated with different measures.

6.1.2 "Proper" and Improper Measures

The concept of "proper" scoring rules was defined and described in Chapter 3. This concept relates to the assessment of probabilities (i.e., to the formulation of probability forecasts), and strictly proper scoring rules reward the assessor (i.e., the forecaster) in such a way as to encourage honesty and discourage hedging. Now, in this chapter we are concerned with the evaluation of probability forecasts, an apparently separate and distinct process. However, as Winkler (1969) has indicated, game-theoretic considerations can arise if the forecaster becomes aware that he is being rewarded according

to one function and evaluated with yet another function. This difficulty can be overcome if the same function, or measure, is used in both the assessment and evaluation processes. Therefore, in formulating and comparing evaluation measures in this study, we shall be particularly concerned with whether or not the measures are strictly proper, proper, or improper. Specifically, the fact that a particular scoring rule is improper will be sufficient grounds upon which to eliminate this function from further consideration as a suitable evaluation measure.

6.1.3 Equivalent Measures

In this study two evaluation measures will be said to be equivalent if they are linearly related. That is, the measure $S'(\underline{r},\underline{d})$ is equivalent to the measure $S(\underline{r},\underline{d})$ if

$$S'(\underline{r},\underline{d}) = aS(\underline{r},\underline{d}) + b,$$

where a and b are constants.

Winkler and Murphy (1968) and others have shown that a scoring rule that is linearly related to a strictly proper scoring rule is itself strictly proper. Thus, an evaluation measure that is equivalent to a strictly proper measure is itself strictly proper.

We shall use the concept of equivalent measures to reduce the degree of arbitrariness in the measures of concern and to make these measures more readily comparable. Specifically, we shall, whenever

possible, transform each measure in such a way that its range of values is the closed unit interval [0,1]. Further, we shall define the measures in such a way that they possess a positive rather than a negative orientation. An evaluation measure is said to have a positive (negative) orientation if a larger (smaller) score is better.

6.1.4 Absolute and Relative Measures

As indicated in Chapter 2, an evaluator is often interested in the relative as well as the absolute "goodness" of forecasts. For example, the evaluator may be concerned with the relative validity of forecasts prepared by two (or more) forecasters or with the validity of a particular meteorologist's forecasts relative to the validity of forecasts produced by some "standard" forecasting system. In this regard, evaluators are often concerned with the <u>skill</u> of forecasts, where the term "skill" refers to the validity (or some other attribute) of the forecasts relative to the validity of forecasts based upon the relevant climatological probabilities. We shall adopt the climatological probabilities as the standard of comparison in this study, and the relative measures considered herein will, in general, be defined in terms of this standard. We shall denote a climatological "forecast" by a vector $\underline{\pi}$, where $\underline{\pi} = (\pi_1,\ldots,\pi_N)$ $(\pi_n \geq 0, \sum_n \pi_n = 1; n=1,\ldots,N)$.

The climatological probabilities considered in the previous paragraph are the "long-term" relative frequencies of occurrence of the relevant events or states for a period, in general, prior to

that for which the forecasts of interest have been prepared. When we consider a collection of forecasts, the sample relative frequencies, $\bar{d}$, represent "estimates" of the climatological probabilities, π, and the former can be used in place of the latter as the "climatological" standard of comparison. As we shall discover in Section 6.2, the relative frequencies, $\bar{d}$, are used instead of the climatological probabilities, π, in the formulation of the basic relative measure of (primary) validity. For a discussion of the relative merits of using climatological probabilities and sample relative frequencies in this context, refer to Murphy (1974).

The relative measures considered in this work will, in general, be formulated indirectly in terms of absolute measures rather than directly in terms of the relevant primitive elements (the set of primitive elements now includes the climatological probabilities). Further, we shall consider members of two classes of relative measures: (1) difference measures and (2) ratio measures. As we shall discover, the former are in general strictly proper while the latter are in general improper. The relative merits of difference and ratio skill scores are also discussed by Murphy (1974).

Finally, the relative "goodness" of forecasts is generally of interest for a collection of forecasts (rather than for an individual forecast). Thus, the discussion of relative measures in this chapter will relate primarily to collections of forecasts. We can, of course, formulate relative measures for an individual forecast

using the climatological probabilities, $\underline{\pi}$. On the other hand, the sample relative frequencies, $\underline{\bar{d}}$, for an individual forecast or for a small collection of forecasts will not, in general, represent "good" estimates of the climatological probabilities, $\underline{\pi}$.

6.2 Distance Measures

The basic measures of the principal attributes, as well as measures of several supplemental attributes, are formulated in this section. These measures are "distance" measures formulated within the (geometrical) framework for empirical evaluation described in Chapter 4.

6.2.1 A Family of Distance Measures

Within the framework for empirical evaluation described in Chapter 4, a forecast $\underline{r}$ corresponds to a point in the simplex R^N and the relevant observation $\underline{d}$ corresponds to a vertex of R^N. Thus, the forecast $\underline{r}$ is completely valid, in the primary sense, if and only if the points corresponding to $\underline{r}$ and $\underline{d}$ are coincident. Thus, some function of the (Euclidean) distance $\Delta(\underline{r},\underline{d})$ (see Chapter 4) between the points $\underline{r}$ and $\underline{d}$, where

$$\Delta(\underline{r},\underline{d}) = [\sum_{n=1}^{N}(r_n - d_n)^2]^{1/2}, \qquad (6\text{-}1)$$

is a natural measure of the validity of the forecast $\underline{r}$. Perhaps the simplest class of such functions is the class of power functions

of the form $\Delta^k(\underline{r},\underline{d})$, where

$$\Delta^k(\underline{r},\underline{d}) = [\sum_{n=1}^{N}(r_n - d_n)^2]^{k/2}, \qquad (6\text{-}2)$$

in which k is a positive number. We shall restrict our attention to this class of functions in this study. In Appendix A, we show that the only member of this class of measures which is a strictly proper scoring rule is the squared distance measure $\Delta^2(\underline{r},\underline{d})$, where

$$\Delta^2(\underline{r},\underline{d}) = (\underline{r} - \underline{d})(\underline{r} - \underline{d})' = \sum_{n=1}^{N}(r_n - d_n)^2 \qquad (6\text{-}3)$$

(a prime denotes a column vector). Hereafter, we shall refer to the measure $\Delta^2(\underline{r},\underline{d})$ as the <u>primitive</u> absolute distance measure. The range of the measure $\Delta^2(\underline{r},\underline{d})$ is the closed interval [0,2] and $\Delta^2(\underline{r},\underline{d})$ has a negative orientation (i.e., a smaller score is better).

6.2.2 The Absolute Distance Measure

Since the concept of distance is a fundamental concept in this work, the primitive distance measure $\Delta^2(\underline{r},\underline{d})$ defined in subsection 6.2.1 represents a basic measure of (primary) validity. However, as indicated in subsection 6.1.1, we would prefer (for comparative purposes) to define evaluation measures in such a way that, if possible, they possess a unit interval range and a positive orientation. If we let $D(\underline{r},\underline{d})$ denote a measure that is equivalent to the primitive distance measure $\Delta^2(\underline{r},\underline{d})$ and that possesses these desirable properties,

then

$$D(\underline{r},\underline{d}) = 1 - (1/2)\Delta^2(\underline{r},\underline{d}), \tag{6-4}$$

or, from (6-3),

$$D(\underline{r},\underline{d}) = 1 - (1/2)(\underline{r} - \underline{d})(\underline{r} - \underline{d})'. \tag{6-5}$$

The measure $D(\underline{r},\underline{d})$ will serve as the basic absolute validity measure (or, more accurately, as the basic absolute measure of primary validity) in this work. Hereafter, we shall refer to $D(\underline{r},\underline{d})$ as the absolute distance measure.

When state s_j obtains (i.e., when $d_j = 1$), $D(\underline{r},\underline{d})$ in (6-5) becomes $D_j(\underline{r})$, where

$$D_j(\underline{r}) = 1 - (1/2)(\underline{r} - \underline{d}_j)(\underline{r} - \underline{d}_j)', \tag{6-6}$$

or

$$D_j(\underline{r}) = (1/2)(1 + 2r_j - \sum_{n=1}^{N} r_n^2). \tag{6-7}$$

Thus, the range of the measure $D(\underline{r},\underline{d})$ is the closed unit interval [0,1] and $D(\underline{r},\underline{d})$ has a positive orientation (i.e., a larger score is better). Specifically, the measure $D(\underline{r},\underline{d})$ assumes its maximum

value of one when the forecast $\underline{r} = \underline{r}_j$ (i.e., when the forecast is categorical and correct), while $D(\underline{r},\underline{d})$ assumes its minimum value of zero when $\underline{r} = \underline{r}_i$ $(i=1,\ldots,N;i\neq j)$ (i.e., when the forecast is categorical and incorrect). As indicated in Chapter 3, the former "event" can occur in only one way, while the latter event can occur in N-1 ways.

The distance measure $D(\underline{r},\underline{d})$ in the two-state (N=2) situation is depicted in Figure 6-1. The solid (dashed) curve represents $D_j(\underline{r})$ when $j = 1(2)$. Note that $D_j(\underline{r})$ is a strictly increasing function of r_j $(j=1,2)$.

6.2.3 A Partition of the Distance Measures

In this subsection we describe a partition of the distance measures considered in subsections 6.2.1 and 6.2.2. We shall formulate the partition in terms of the primitive measure $\Delta^2(\underline{r},\underline{d})$. However, since the measures $\Delta^2(\underline{r},\underline{d})$ and $D(\underline{r},\underline{d})$ are equivalent, this partition is a partition of both of these measures. This particular partition is of interest for two reasons: (1) the elements of the partition are themselves measures of certain attributes of the forecasts; and (2) a relative distance measure (i.e., a relative measure of primary validity) can be formulated in terms of these elements. The partition itself and the elements thereof are considered in this subsection, while the relative distance measure is described in subsection 6.2.4.

The partitions of concern in this study are partitions of the

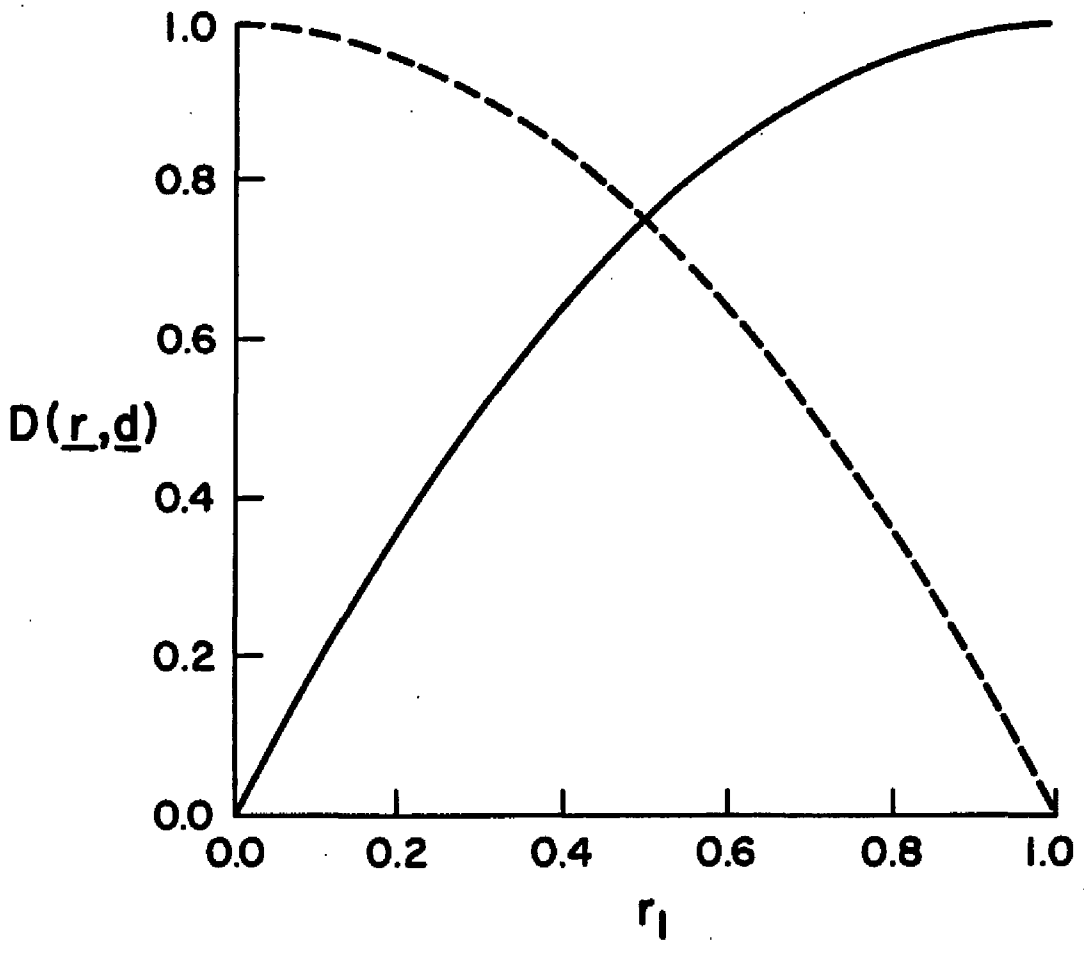


Figure 6-1. The absolute distance measure $D(\underline{r},\underline{d})$ in the two-state situation. The solid curve denotes $D_1(\underline{r})$ and the dashed curve denotes $D_2(\underline{r})$.

average score for a collection of K forecasts $\underline{r}_k$ and the K relevant observations $\underline{d}_k$ (k=1,...,K). When the measure of concern is the primitive distance measure $\Delta^2(\underline{r},\underline{d})$, the average score is $\overline{\Delta^2}(\underline{r},\underline{d})$, where

$$\overline{\Delta^2}(\underline{r},\underline{d}) = (1/K)\sum_{k=1}^{K}(\underline{r}_k - \underline{d}_k)(\underline{r}_k - \underline{d}_k)'. \qquad (6\text{-}8)$$

The measure $\overline{\Delta^2}(\underline{r},\underline{d})$ is the average squared distance between the forecasts $\underline{r}_k$ and the relevant observations $\underline{d}_k$ (k=1,...,K). These partitions are also based upon the assumption that the probabilities which constitute the forecasts can assume only a finite set of values. This assumption leads to the identification of T subcollections of the collection of K (vector) forecasts, each of which corresponds to a distinct forecast $\underline{r}^t$ (t=1,...,T) (see Chapter 3). Murphy (1973b) has shown that a partition of $\overline{\Delta^2}(\underline{r},\underline{d})$ can be formulated in terms of these subcollections and that this partition can be expressed as follows:

$$\overline{\Delta^2}(\underline{r},\underline{d}) = (1/K)\sum_{k=1}^{K}(\bar{\underline{d}} - \underline{d}_k)(\bar{\underline{d}} - \underline{d}_k)' + (1/K)\sum_{t=1}^{T}K^t(\underline{r}^t - \bar{\underline{d}}^t)(\underline{r}^t - \bar{\underline{d}}^t)' - (1/K)\sum_{t=1}^{T}K^t(\bar{\underline{d}}^t - \bar{\underline{d}})(\bar{\underline{d}}^t - \bar{\underline{d}})'. \qquad (6\text{-}9)$$

The first term on the right-hand side (RHS) of (6-9) is simply the average score, according to the measure $\Delta^2(\underline{r},\underline{d})$, if the forecasts, $\underline{r}_k$ (k=1,...,K), are replaced by the sample relative frequencies for the collection of K forecasts, $\bar{\underline{d}}$. That is,

$$\overline{\Delta^2}(\bar{\underline{d}},\underline{d}) = (1/K)\sum_{k=1}^{K}(\bar{\underline{d}} - \underline{d}_k)(\bar{\underline{d}} - \underline{d}_k)'. \qquad (6\text{-}10)$$

This term, the average squared distance between the individual observations $\underline{d}_k$ (k=1,...,K) and the average observation $\bar{\underline{d}}$, represents the variance of the observations $\underline{d}_k$ (k=1,...,K) and, as such, is a measure of the uncertainty inherent in the N events, or states, on the K occasions of concern. As this measure of uncertainty increases (i.e., as the uncertainty itself increases), the average score $\overline{\Delta^2}(\underline{r},\underline{d})$ increases (i.e., becomes poorer).

The second term on the RHS of (6-9) is a measure of the degree to which the T distinct forecasts $\underline{r}^t$ differ from the relevant sample relative frequencies $\bar{\underline{d}}^t$ (t=1,...,T). As such, this term is a measure of the secondary validity, or, more precisely, of the bias-in-the-small (or reliability), of the forecasts. As this measure of bias-in-the-small increases (i.e., as the bias-in-the-small itself increases) the average score $\overline{\Delta^2}(\underline{r},\underline{d})$ increases (i.e., becomes poorer). This term represents the weighted average of the squared distances between the T distinct forecasts $\underline{r}^t$ and the respective sample relative frequencies $\bar{\underline{d}}^t$ (t=1,...,T). We shall denote this term by $\overline{\Delta^2}(\underline{r}^t,\bar{\underline{d}}^t)$, i.e.,

$$\overline{\Delta^2}(\underline{r}^t,\bar{\underline{d}}^t) = (1/K)\sum_{t=1}^{T}K^t(\underline{r}^t - \bar{\underline{d}}^t)(\underline{r}^t - \bar{\underline{d}}^t)'. \qquad (6\text{-}11)$$

The measure $\overline{\Delta^2}(\underline{r}^t,\bar{\underline{d}}^t)$ will serve as the basic (absolute) measure of

the attribute bias-in-the-small in this work.

The third term on the RHS of (6-9), the weighted average of the squared distances between the T sample relative frequencies $\underline{\bar{d}}^t$ and the overall relative frequency $\underline{\bar{d}}$, is a measure of the degree to which the sample relative frequencies corresponding to the T distinct forecasts, $\underline{\bar{d}}^t$ (t=1,...,T), differ from the sample relative frequencies for the entire collection of forecasts, $\underline{\bar{d}}$. We shall refer to this term as a measure of the attribute resolution, although this definition of the attribute differs from the definition of this attribute presented in Chapter 3 (see discussion in Section 6.3). As this measure of resolution increases (i.e., as the resolution itself increases) the average score $\overline{\Delta^2}(\underline{r},\underline{d})$ decreases (i.e., becomes better). We shall denote this term by $\overline{\Delta^2}(\underline{\bar{d}}^t,\underline{\bar{d}})$, i.e.,

$$\overline{\Delta^2}(\underline{\bar{d}}^t,\underline{\bar{d}}) = (1/K)\sum_{t=1}^{T} K^t(\underline{\bar{d}}^t - \underline{\bar{d}})(\underline{\bar{d}}^t - \underline{\bar{d}})'. \qquad (6\text{-}12)$$

The measure $\overline{\Delta^2}(\underline{\bar{d}}^t,\underline{\bar{d}})$ will serve as the basic (absolute) measure of the attribute resolution in this work.

Thus, the partition of the average (primitive) distance measure $\overline{\Delta^2}(\underline{r},\underline{d})$ yields three terms, each of which is a measure of an attribute of the forecasts (or the observations). Moreover, these measures are themselves distance measures and, as such, are natural measures of the respective attributes within the framework for empirical evaluation described in Chapter 4.

A sample collection of ten (K=10) forecasts in the three-state (N=3) situation is presented in Table 6-1. The partition of the primitive distance measure $\overline{\Delta^2}(\underline{r},\underline{d})$ for these forecasts is presented in Table 6-2. The sample relative frequencies for this collection of forecasts, $\bar{\underline{d}}$, appears under the column headed "Sample Relative Frequency" in the row headed "Average." Thus, $\bar{\underline{d}} = (0.2,0.4,0.4)$. Now, the measure of uncertainty $\overline{\Delta^2}(\bar{\underline{d}},\underline{d})$ in (6-10) can also be expressed as follows:

$$\overline{\Delta^2}(\bar{\underline{d}},\underline{d}) = \bar{\underline{d}}(\underline{u} - \bar{\underline{d}})', \tag{6-13}$$

where $\underline{u} = (1,\ldots,1)$. Therefore, $\overline{\Delta^2}(\bar{\underline{d}},\underline{d})$ for this collection of forecasts is equal to 0.640. The values of the bias-in-the-small measure, $\overline{\Delta^2}(\underline{r}^t,\bar{\underline{d}}^t)$, and the resolution measure, $\overline{\Delta^2}(\bar{\underline{d}}^t,\bar{\underline{d}})$, are 0.292 and 0.440, respectively. Thus, from (6-9), $\overline{\Delta^2}(\underline{r},\underline{d}) = 0.640 + 0.292 - 0.440 = 0.492$. The partition of $\overline{\Delta^2}(\underline{r},\underline{d})$ for a subset of this collection of forecasts, namely the subset consisting of the subcollections for which t = 4, 5, 6, and 7, is depicted within the framework of the simplex R^3 in Figure 6-2. The contributions to each term in the partition of the measure $\overline{\Delta^2}(\underline{r},\underline{d})$ are indicated in this figure (for the subcollections of concern).

6.2.4 The Relative Distance Measure

Consider the partition of the primitive distance measure $\overline{\Delta^2}(\underline{r},\underline{d})$ in (6-9), and recall that the first term on the RHS of (6-9)

Forecast/ Observation Number k	Forecast $\underline{r}_k$	Observation $\underline{d}_k$
1	(0.1,0.3,0.6)	(0,0,1)
2	(0.1,0.7,0.2)	(0,1,0)
3	(0.3,0.5,0.2)	(0,1,0)
4	(0.5,0.4,0.1)	(0,1,0)
5	(0.7,0.3,0.0)	(1,0,0)
6	(0.6,0.1,0.3)	(0,0,1)
7	(0.5,0.4,0.1)	(1,0,0)
8	(0.1,0.8,0.1)	(0,1,0)
9	(0.1,0.6,0.3)	(0,0,1)
10	(0.1,0.7,0.2)	(0,0,1)

Table 6-1. A sample collection of forecasts and the relevant observations in the three-state situation.

Subcollection Number t	Forecast $\underline{r}^t$	Number of Forecasts K^t	Sample Relative Frequency $\bar{\underline{d}}^t$	Subcollection Bias-in-the-Small $K^t(\underline{r}^t-\bar{\underline{d}}^t)(\underline{r}^t-\bar{\underline{d}}^t)'$	Subcollection Resolution $K^t(\bar{\underline{d}}^t-\bar{\underline{d}})(\bar{\underline{d}}^t-\bar{\underline{d}})'$
1	(0.1,0.3,0.6)	1	(0.0,0.0,1.0)	0.26	0.56
2	(0.1,0.6,0.3)	1	(0.0,0.0,1.0)	0.86	0.56
3	(0.1,0.7,0.2)	2	(0.0,0.5,0.5)	0.28	0.12
4	(0.1,0.8,0.1)	1	(0.0,1.0,0.0)	0.06	0.56
5	(0.3,0.5,0.2)	1	(0.0,1.0,0.0)	0.38	0.56
6	(0.5,0.4,0.1)	2	(0.5,0.5,0.0)	0.04	0.52
7	(0.6,0.1,0.3)	1	(0.0,0.0,1.0)	0.86	0.56
8	(0.7,0.3,0.0)	1	(1.0,0.0,0.0)	0.18	0.96
Total		10		2.92	4.40
Average			(0.2,0.4,0.4)	0.292	0.440

Table 6-2. The partition of the distance measure $\overline{\Delta^2}(\underline{r},\underline{d})$ for the sample collection of forecasts presented in Table 6-1.

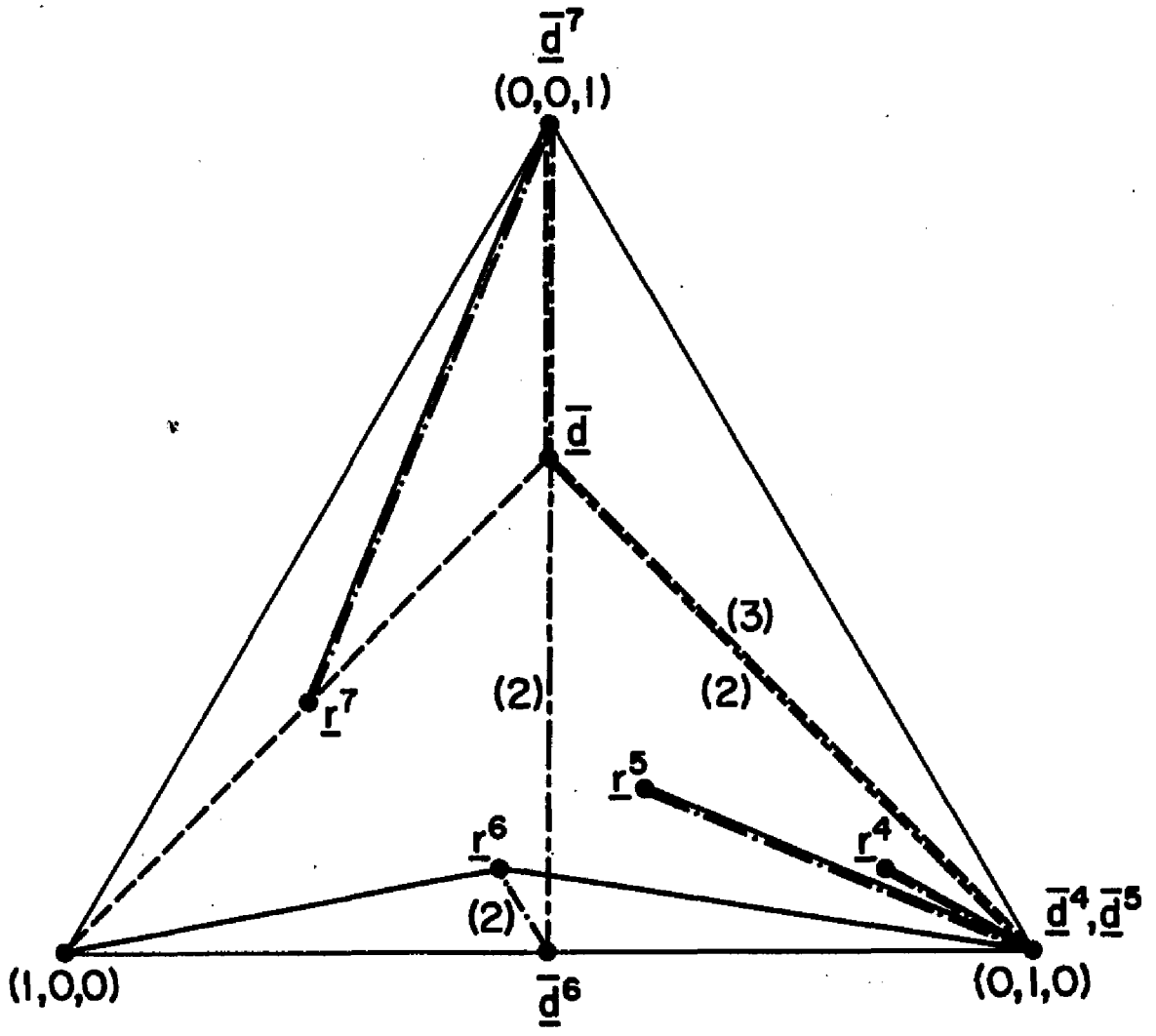


Figure 6-2. The representation of the terms in the partition of the measure $\Delta^2(\underline{r},\underline{d})$ within the simplex R^3 for a subset of the collection of ten forecasts presented in Table 6-1. The five (K=5) forecasts in the four subcollections for which t = 4, 5, 6, and 7 are depicted (see Table 6-2). The contributions to the terms $\Delta^2(\underline{r},\underline{d})$, $\Delta^2(\bar{\underline{d}},\underline{d})$, $\Delta^2(\underline{r}^t,\bar{\underline{d}}^t)$, and $\Delta^2(\bar{\underline{d}}^t,\bar{\underline{d}})$ are represented by solid (——), dashed (– – –), dashed-dotted (– · –), and dashed-long-and-short (—— – ——) lines, respectively. The numbers in parentheses adjacent to certain lines indicate the number of times that this particular distance occurs for this subset of forecasts (the absence of a number indicates that this distance occurs only once).

is the (average) distance measure for the sample relative frequencies, i.e., $\overline{\Delta^2}(\bar{\underline{d}},\underline{d})$. Since the sample relative frequencies, $\bar{\underline{d}}$, represent a suitable standard of comparison (see subsection 6.1.4), we can formulate a difference distance measure of (primary) validity in terms of the elements of this partition. In particular, we can rewrite (6-9) as follows:

$$\overline{\Delta^2}(\bar{\underline{d}},\underline{d}) - \overline{\Delta^2}(\underline{r},\underline{d}) = (1/K)\sum_{t=1}^{T} K^t(\bar{\underline{d}}^t - \bar{\underline{d}})(\bar{\underline{d}}^t - \bar{\underline{d}})' - (1/K)\sum_{t=1}^{T} K^t(\underline{r}^t - \bar{\underline{d}}^t)(\underline{r}^t - \bar{\underline{d}}^t)', \tag{6-14}$$

or, if we let $\overline{\Delta^2}(\bar{\underline{d}},\underline{r},\underline{d}) = \overline{\Delta^2}(\bar{\underline{d}},\underline{d}) - \overline{\Delta^2}(\underline{r},\underline{d})$,

$$\overline{\Delta^2}(\bar{\underline{d}},\underline{r},\underline{d}) = (1/K)\sum_{t=1}^{T} K^t(\bar{\underline{d}}^t - \bar{\underline{d}})(\bar{\underline{d}}^t - \bar{\underline{d}})' - (1/K)\sum_{t=1}^{T} K^t(\underline{r}^t - \bar{\underline{d}}^t)(\underline{r}^t - \bar{\underline{d}}^t)'. \tag{6-15}$$

Hereafter, we shall refer to the measure $\overline{\Delta^2}(\bar{\underline{d}},\underline{r},\underline{d})$ as the <u>primitive</u> relative distance measure.

Since $\overline{\Delta^2}(\bar{\underline{d}},\underline{r},\underline{d}) = \overline{\Delta^2}(\bar{\underline{d}},\underline{d}) - \overline{\Delta^2}(\underline{r},\underline{d})$ and $\overline{\Delta^2}(\bar{\underline{d}},\underline{d})$ is a constant (with respect to $\underline{r}$), the measure $\overline{\Delta^2}(\bar{\underline{d}},\underline{r},\underline{d})$ is a strictly proper scoring rule. The range of $\overline{\Delta^2}(\bar{\underline{d}},\underline{r},\underline{d})$ is the closed interval $[-2,(N-1)/N]$ and $\overline{\Delta^2}(\bar{\underline{d}},\underline{r},\underline{d})$ has a positive orientation. If we let $\bar{D}(\bar{\underline{d}},\underline{r},\underline{d})$ denote a relative distance measure with a unit interval range and a positive orientation which is equivalent to the measure

$\overline{\Delta^2}(\bar{\underline{d}},\underline{r},\underline{d})$, then

$$\bar{D}(\bar{\underline{d}},\underline{r},\underline{d}) = [N/(3N-1)][2 + (1/K)\sum_{t=1}^{T} K^t(\bar{\underline{d}}^t - \bar{\underline{d}})(\bar{\underline{d}}^t - \bar{\underline{d}})' - (1/K)\sum_{t=1}^{T} K^t(\underline{r}^t - \bar{\underline{d}}^t)(\underline{r}^t - \bar{\underline{d}}^t)'] . \quad (6\text{-}16)$$

The measure $\bar{D}(\bar{\underline{d}},\underline{r},\underline{d})$ will serve as the basic relative validity measure (or, more accurately, as the basic relative measure of primary validity) in this work. Hereafter, we shall refer to $\bar{D}(\bar{\underline{d}},\underline{r},\underline{d})$ as the relative distance measure.

As defined in (6-16), the range of the measure $\bar{D}(\bar{\underline{d}},\underline{r},\underline{d})$ is the closed unit interval [0,1] and $\bar{D}(\bar{\underline{d}},\underline{r},\underline{d})$ has a positive orientation. Specifically, the measure $\bar{D}(\bar{\underline{d}},\underline{r},\underline{d})$ assumes its maximum value of one when $\underline{r}_k = \underline{d}_k$ for all k (k=1,...,K) (i.e., when the K forecasts are all categorical and correct) and $\bar{\underline{d}} = (1/N,...,1/N)$ (i.e., when the sample relative frequencies are all equal), while $\bar{D}(\bar{\underline{d}},\underline{r},\underline{d})$ assumes its minimum value of zero when $r_{nk} = 1$ for some n for all k and $\bar{d}_m = 1$ for some m (m,n=1,...,N;m≠n;k=1,...,K) (i.e., when the K forecasts are all categorical and incorrect and the same event, or state, obtains on all K occasions). Note that the measure $\bar{D}(\bar{\underline{d}},\underline{r},\underline{d})$ will seldom attain its maximum and minimum values. In this regard, the value of $\bar{D}(\bar{\underline{d}},\underline{r},\underline{d})$ for the collection of forecasts presented in Table 6-1 is 0.8055.

6.3 Distance Measures and Other Measures: Some Comparisons

In this section we consider empirical evaluation measures proposed heretofore and compare these measures with the distance measures described in Section 6.2. We shall be primarily concerned with measures of the principal attributes, namely (primary) validity and bias (in-the-small).

6.3.1 Measures of Validity

In this subsection we examine several measures of validity and compare these measures with the relevant distance measures. Absolute measures are considered in paragraph 6.3.1.1, while relative measures are considered in paragraph 6.3.1.2.

6.3.1.1 Absolute Measures

The following absolute measures of primary validity are described in this paragraph: (1) the linear score; (2) the quadratic, or probability, score; (3) the logarithmic score; (4) the spherical score; and (5) the ranked probability score. First we examine each measure separately and then we compare these measures with the absolute distance measure $D(\underline{r},\underline{d})$.

(1) The linear score. The linear, or absolute, score $A(\underline{r},\underline{d})$, proposed by Woodall (1967), is defined as follows:

$$A(\underline{r},\underline{d}) = \sum_{n=1}^{N} |r_n - d_n| . \qquad (6\text{-}17)$$

When state s_j obtains (i.e., when $d_j = 1$), $A(\underline{r},\underline{d})$ becomes $A_j(\underline{r})$, where

$$A_j(\underline{r}) = 2(1 - r_j). \tag{6-18}$$

The range of $A(\underline{r},\underline{d})$ is the closed unit interval [0,1] and $A(\underline{r},\underline{d})$ has a negative orientation.

The measure $A(\underline{r},\underline{d})$ is an <u>improper</u> scoring rule (e.g., Murphy and Epstein, 1967b). Specifically, if we let $\underline{p} = (p_1,\ldots,p_N)$ denote the forecaster's judgment and let p_i denote the largest of the p_n, then the forecaster will minimize his (subjective) expected score $A(\underline{r},\underline{p})$, where

$$A(\underline{r},\underline{p}) = \sum_{j=1}^{N} p_j A_j(\underline{r}),$$

or

$$A(\underline{r},\underline{p}) = 2(1 - \sum_{j=1}^{N} p_j r_j),$$

by setting r_i equal to one and by setting r_n equal to zero for all $n \neq i$ $(n=1,\ldots,N)$. The linear measure $A(\underline{r},\underline{d})$, then, encourages a considerable amount of hedging, since the forecaster will minimize his (expected) score by making a categorical forecast on each occasion. As a result, this measure will be of no further concern in this study.

(2) <u>The quadratic score</u>. The quadratic, or probability, score

$Q(\underline{r},\underline{d})$, originally formulated by Brier (1950) (see also de Finetti, 1962), is defined as follows:

$$Q(\underline{r},\underline{d}) = (\underline{r} - \underline{d})(\underline{r} - \underline{d})' = \sum_{n=1}^{N}(r_n - d_n)^2. \qquad (6\text{-}19)$$

When $d_j = 1$, $Q(\underline{r},\underline{d})$ becomes $Q_j(\underline{r})$, where

$$Q_j(\underline{r}) = 1 - 2r_j + \sum_{n=1}^{N} r_n^2. \qquad (6\text{-}20)$$

As defined in (6-19), the range of $Q(\underline{r},\underline{d})$ is the closed interval [0,2] and $Q(\underline{r},\underline{d})$ has a negative orientation.

A comparison of (6-3) and (6-19) reveals that the quadratic measure $Q(\underline{r},\underline{d})$ is identical to the primitive distance measure $\Delta^2(\underline{r},\underline{d})$. Thus, the measure $Q(\underline{r},\underline{d})$ is (1) equal to the square of the distance between the forecast $\underline{r}$ and the relevant observation $\underline{d}$ (cf. Epstein and Murphy, 1965); (2) equivalent to the absolute distance measure $D(\underline{r},\underline{d})$; and (3) a strictly proper scoring rule (cf. Murphy and Epstein, 1967b; Winkler and Murphy, 1968).

If, for comparative purposes, we define a measure $Q'(\underline{r},\underline{d})$ to be equivalent to the measure $Q(\underline{r},\underline{d})$ and to possess a unit interval range and a positive orientation, then

$$Q'(\underline{r},\underline{d}) = 1 - (1/2)\sum_{n=1}^{N}(r_n - d_n)^2. \qquad (6\text{-}21)$$

The measure $Q'(\underline{r},\underline{d})$ is, of course, identical to the absolute distance measure $D(\underline{r},\underline{d})$ [see (6-5)].

(3) The logarithmic score. The logarithmic score $L(\underline{r},\underline{d})$, proposed by Good (1952) (see also van Naerssen, 1962; Winkler and Murphy, 1968), is defined as follows:

$$L(\underline{r},\underline{d}) = \ln(\underline{r}\underline{d}') = \ln(\sum_{n=1}^{N} r_n d_n) \tag{6-22}$$

(ln = natural logarithm). When $d_j = 1$, $L(\underline{r},\underline{d})$ becomes $L_j(\underline{r})$, where

$$L_j(\underline{r}) = \ln(r_j). \tag{6-23}$$

As defined in (6-22), the range of $L(\underline{r},\underline{d})$ is the half-open interval $(-\infty,0]$ and $L(\underline{r},\underline{d})$ has a positive orientation.

The measure $L(\underline{r},\underline{d})$ possesses several interesting properties. First, this measure is a strictly proper scoring rule (cf. Winkler and Murphy, 1968). Second $L_j(\underline{r}) = 0$ when $r_j = 0$. From the point of view of probability assessment, this particular property of the logarithmic score can be considered to be a desirable property, since such a measure would encourage forecasters to be very careful in assigning a probability of zero to any event or state. On the other hand, from the point of view of evaluation, this property presents certain obvious practical problems. In this regard, Shuford, Albert, and Massengill (1966) have suggested that a truncated

version of the logarithmic score, $L^*(\underline{r},\underline{d})$, be used in place of the original score, where

$$L_j^*(\underline{r}) = \begin{cases} \ln(r_j^*) & \text{when } r_j \leq r_j^*, \\ \ln(r_j) & \text{otherwise.} \end{cases}$$

The probability r_j^* can be taken to be some suitably small value, e.g., 0.01 or 0.001. Note that $L_j^*(\underline{r})$ is not a strictly proper scoring rule. Other properties of the measure $L(\underline{r},\underline{d})$ will be discussed when the measures of absolute validity are compared.

For comparative purposes, we shall define a measure $L'(\underline{r},\underline{d})$ which is equivalent to $L(\underline{r},\underline{d})$, where

$$L'(\underline{r},\underline{d}) = 1 + \ln(\sum_{n=1}^{N} r_n d_n). \tag{6-24}$$

The range of $L'(\underline{r},\underline{d})$ is the half-open interval $(-\infty,1]$ and $L'(\underline{r},\underline{d})$ has a positive orientation.

(4) The spherical score. The spherical score $S(\underline{r},\underline{d})$, described by Roby (1965) (see also Winkler and Murphy, 1968), is defined as follows:

$$S(\underline{r},\underline{d}) = (\underline{r}\underline{d}')/(\underline{r}\underline{r}')^{1/2} = (\sum_{n=1}^{N} r_n d_n)/(\sum_{n=1}^{N} r_n^2)^{1/2}. \tag{6-25}$$

When $d_j = 1$, $S(\underline{r},\underline{d})$ becomes $S_j(\underline{r})$, where

$$S_j(\underline{r}) = r_j / \left(\sum_{n=1}^{N} r_n^2\right)^{1/2}. \tag{6-26}$$

The range of $S(\underline{r},\underline{d})$ is the closed unit interval [0,1] and $S(\underline{r},\underline{d})$ has a positive orientation.

The measure $S(\underline{r},\underline{d})$ is also a strictly proper scoring rule (cf. Winkler and Murphy, 1968).

(5) The ranked probability score. The ranked probability score $RPS(\underline{r},\underline{d})$, formulated by Epstein (1969), can be expressed as follows:

$$RPS(\underline{r},\underline{d}) = 1 - [1/(N-1)] \sum_{n=1}^{N} \left(\sum_{m=1}^{n} r_m - \sum_{m=1}^{n} d_m\right)^2. \tag{6-27}$$

(see Murphy, 1971). When $d_j = 1$, $RPS(\underline{r},\underline{d})$ becomes $RPS_j(\underline{r})$, where

$$RPS_j(\underline{r}) = 1 - [1/(N-1)]\left[\sum_{n=1}^{j-1} \left(\sum_{m=1}^{n} r_m\right)^2 + \sum_{n=j}^{N} \left(\sum_{m=1}^{n} r_m - 1\right)^2\right],$$

or

$$RPS_j(\underline{r}) = 1 - [1/(N-1)]\left[(N - j + 1) - 2\sum_{n=j}^{N} \sum_{m=1}^{n} r_m + \sum_{n=1}^{N} \left(\sum_{m=1}^{n} r_m\right)^2\right]. \tag{6-28}$$

As defined in (6-27), the range of $RPS(\underline{r},\underline{d})$ is the closed unit interval [0,1] and $RPS(\underline{r},\underline{d})$ has a positive orientation.

The measure RPS($\underline{r}$,$\underline{d}$) is a strictly proper scoring rule (Murphy, 1969b). This measure is also "sensitive to distance" (Stael von Holstein, 1970b; see also Murphy, 1970). In order to illustrate this concept, consider the forecasts $\underline{r}^A$ and $\underline{r}^B$ prepared by forecasters A and B, respectively, where $\underline{r}^A$= (0.6,0.3,0.1) and $\underline{r}^B$= (0.6,0.1,0.3). We shall assume that $\underline{d}$ = (1,0,0). The validity measures considered heretofore would assign A's and B's forecasts the same score. For example, the quadratic score for these forecasts would be Q($\underline{r}$,$\underline{d}$), where $Q(\underline{r},\underline{d}) = Q^A(\underline{r},\underline{d}) = Q^B(\underline{r},\underline{d}) = 0.26$. Specifically, the scores assigned by these measures when state s_j obtains (d_j=1) are invariant with respect to changes in the indices associated with the other N-1 states s_i ($i=1,\ldots,N;i\neq j$). The ranked probability score, on the other hand, is "sensitive" to such changes and assigns higher (i.e., "better") scores to those forecasts for which the weighted sums of the probabilities ($p_i;i\neq j$) is smaller, where the weights are a function of the differences in indices (see below). In particular, the ranked probability score RPS($\underline{r}$,$\underline{d}$) would assign forecaster A a score of 0.915 and forecaster B a score of 0.875. Murphy (1970) has shown that the ranked probability score assigns weights to the probabilities p_i ($i\neq j$) which increase linearly as the difference in indices increases, while the quadratic score assigns equal weights to all such probabilities. Since the basic distance measure formulated in Section 6.2 is equivalent to the quadratic score, the framework for empirical evaluation which we have adopted implicitly assumes that such weights are equal. By analogy, the ranked probability score

would appear to correspond to a weighted distance measure and, as such, to a different framework for empirical evaluation. We briefly examine this framework in Chapter 9.

In the two-state (N=2) situation RPS($\underline{r}$,$\underline{d}$) in (6-27) can be expressed as follows:

$$RPS(\underline{r},\underline{d}) = 1 - (r_1 - d_1)^2.$$

Thus, from (6-21),

$$RPS(\underline{r},\underline{d}) = Q'(\underline{r},\underline{d}). \qquad (6\text{-}29)$$

That is, the ranked probability score and the quadratic, or probability, score are identical in the two-state situation.

(6) Comparison of absolute validity measures. As indicated in (2) above, the quadratic score $Q'(\underline{r},\underline{d})$ is identical to the distance measure $D(\underline{r},\underline{d})$. Thus, we are concerned with the relative merits of the (absolute) distance, logarithmic, spherical, and ranked probability measures. Unfortunately, very few definitive criteria exist upon which to base the choice of a particular measure in this situation. All of these measures are strictly proper scoring rules. With regard to their suitability as measures of the attribute validity, only the logarithmic score $L'(\underline{r},\underline{d})$ appears to possess any possible deficiency. This measure depends only

upon the probability assigned to the event that obtains (r_j) and, as such, could be considered to represent a partial rather than a complete measure of validity [see (3-7)]. However, this property of the logarithmic score has been considered to be desirable by other investigators (e.g., Raiffa, 1969; see also Winkler, 1969). In any case, each of the other three measures depends upon all of the N probabilities r_n (n=1,...,N). For the purposes of this study, we shall adopt the distance measure $D(\underline{r},\underline{d})$ as the "standard" measure of validity, since this measure has a natural interpretation within the context of the framework for empirical evaluation. As we shall discover, the relationships between empirical and operational evaluation measures to be described in Chapter 8 do not depend upon the choice of a particular measure of validity (as long as this measure is a strictly proper scoring rule).

For comparative purposes, we depict the measures $D(\underline{r},\underline{d})$ [$=Q'(\underline{r},\underline{d})=RPS(\underline{r},\underline{d})$], $L'(\underline{r},\underline{d})$, and $S(\underline{r},\underline{d})$ in Figure 6-3 for the two-state (N=2) situation. Note that $D_n(\underline{r})$, $L'_n(\underline{r})$, and $S_n(\underline{r})$ are all strictly increasing functions of r_n (n=1,2). That is, these scores increase as r_n increases. In N-state ($N\geq 2$) situations, all of these measures [including $RPS_n(\underline{r})$] are strictly increasing functions of r_n along <u>directed</u> line segments through the vertex $\underline{d}_n$ (n=1,...,N). A directed line segment is defined by the existence of a particular relationship among the ratios of the components of two (or more) forecasts $\underline{r}$ and r^* (say). In a three-state situation in which $d_j = 1$, this relationship can be expressed as follows: $r_i/r_i^* = r_k/r_k^* = (1-r_j)/(1-r_j^*)$

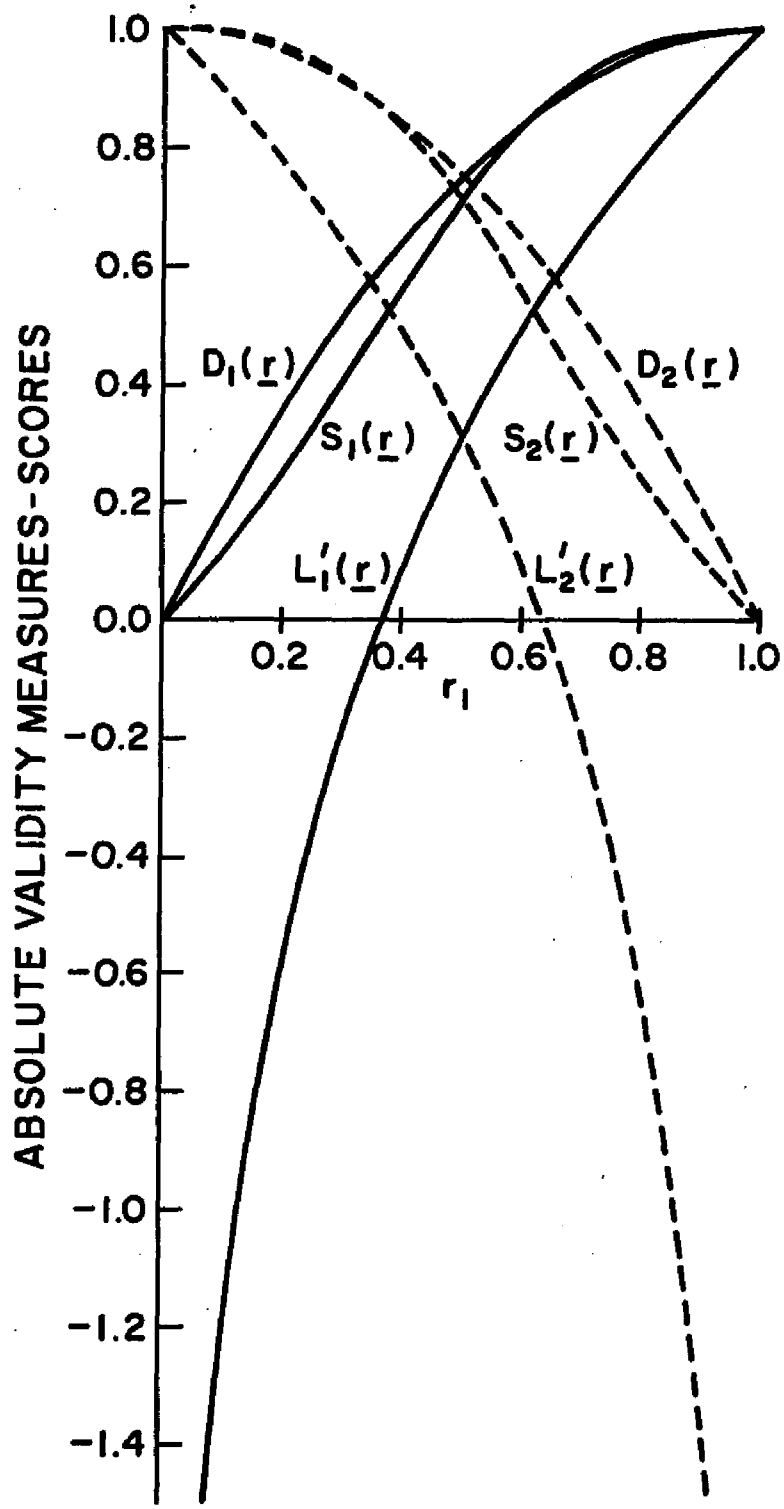


Figure 6-3. The absolute validity measures $D(\underline{r},\underline{d})$ [$=Q'(\underline{r},\underline{d})=RPS(\underline{r},\underline{d})$], $L'(\underline{r},\underline{d})$, and $S(\underline{r},\underline{d})$ in the two-state situation. The solid curves denote $D_1(\underline{r})$, $L_1'(\underline{r})$, and $S_1(\underline{r})$ and the dashed curves denote $D_2(\underline{r})$, $L_2'(\underline{r})$, and $S_2(\underline{r})$.

$(i,j,k=1,2,3;i\neq k;i,k\neq j)$. All line segments are directed line segments in the two-state situation, since $r_i/r_i^* = (1-r_j)/(1-r_j^*) = r_i/r_i^*$ $(i,j=1,2;i\neq j)$. In Figure 6-4, we depict the distance measure $D(\underline{r},\underline{d})$ when $d_j = 1$, i.e., $D_1(\underline{r})$. As indicated above, $D_1(\underline{r})$ increases as r_1 increases along any directed line segment through d_1. The other measures would be represented by lines $[L_1'(\underline{r})]$ or by curves with different shapes $[RPS_1(\underline{r})$ and $S_1(\underline{r})]$ in such a figure, but they would all possess this property. The measure $D(\underline{r},\underline{d})$ is, of course, also a strictly decreasing function of the distance $\Delta_n(\underline{r})$ [see (6-1)], but this property is <u>not</u> shared by the other measures.

6.3.1.2 Relative Measures

The following relative measures of primary validity are considered in this paragraph: (1) the ratio skill score; (2) the information ratio; and (3) the matrix skill score. These measures are first described and then compared with the (basic) relative distance measure $D(\bar{\underline{d}},\underline{r},\underline{d})$.

(1) <u>The ratio skill score</u>. The ratio skill score $RSS(\underline{\pi},\underline{r},\underline{d})$ (Sanders, 1958; Glahn and Jorgenson, 1970) is defined as follows:

$$RSS(\underline{\pi},\underline{r},\underline{d}) = [Q(\underline{\pi},\underline{d}) - Q(\underline{r},\underline{d})]/Q(\underline{\pi},\underline{d}),$$

or

$$RSS(\underline{\pi},\underline{r},\underline{d}) = 1 - [Q(\underline{r},\underline{d})/Q(\underline{\pi},\underline{d})], \qquad (6\text{-}30)$$

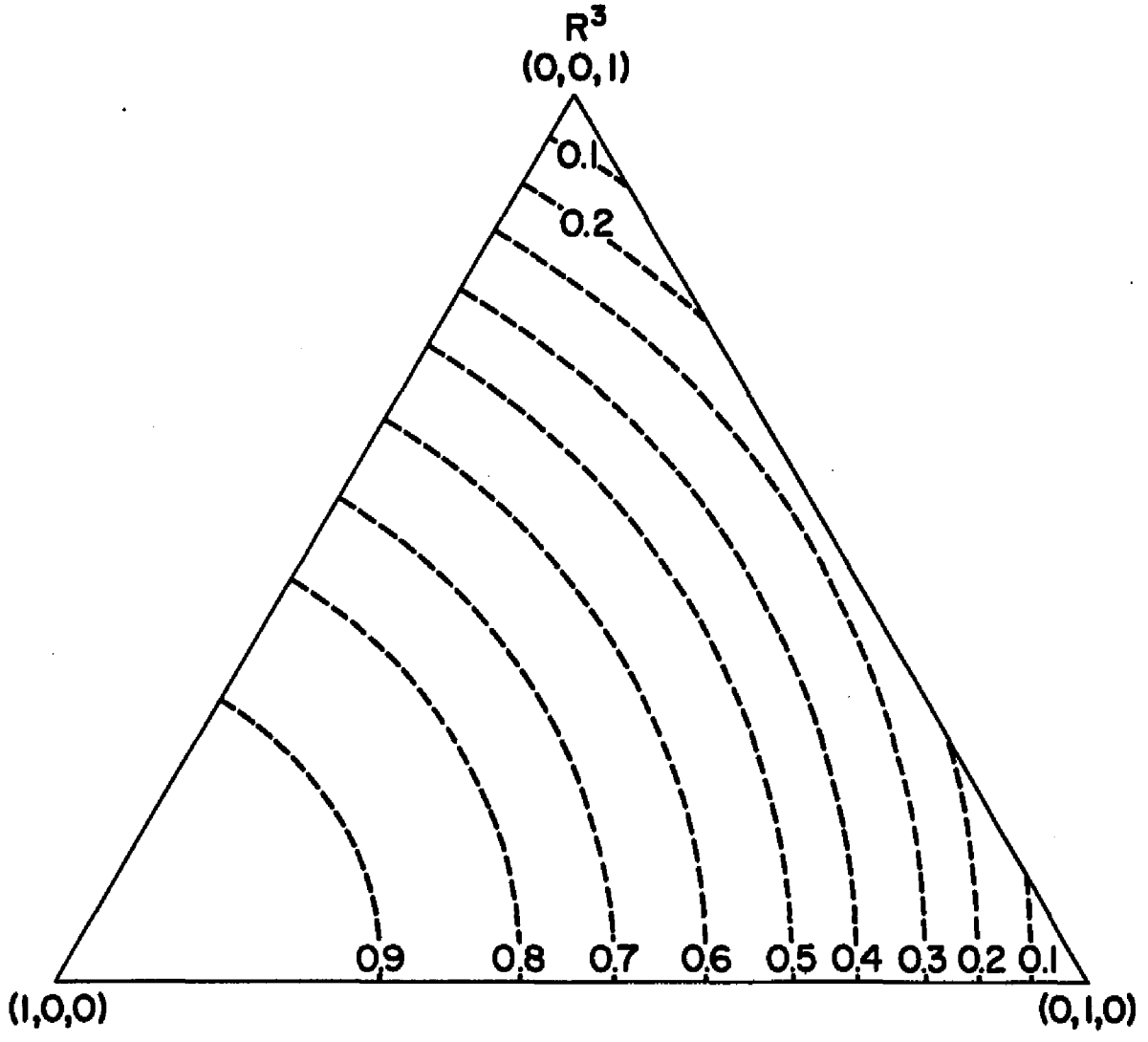


Figure 6-4. The validity measure $D(\underline{r},\underline{d})$ in the three-state situation when state s_1 obtains (i.e., when $d_1 = 1$).

where $Q(\underline{\pi},\underline{d})$ is the quadratic score for the climatological forecast $\underline{\pi}$. The range of possible values of the measure $RSS(\underline{\pi},\underline{r},\underline{d})$ is the half-open interval $(-\infty,1]$. However, unless $\pi_n = 1$ for some n and $d_n = 1$ $(n=1,\ldots,N)$, the actual range of $RSS(\underline{\pi},\underline{r},\underline{d})$ will be a closed interval determined by the climatological forecast $\underline{\pi}$ and the observation $\underline{d}$. As defined in (6-30), the measure $RSS(\underline{\pi},\underline{r},\underline{d})$ represents the fractional amount by which the quadratic score for the forecast $\underline{r}$ improves upon the quadratic score for the climatological forecast $\underline{\pi}$. Thus, this measure has a positive orientation.

Murphy (1973a) has shown that the measure $RSS(\underline{\pi},\underline{r},\underline{d})$ is an improper scoring rule. Specifically, the "optimal" forecast is $\underline{r}^* = (r_1^*,\ldots,r_N^*)$, where

$$r_m^* = [p_m/Q_m(\underline{\pi})]/\sum_{n=1}^{N}[p_n/Q_n(\underline{\pi})], \quad (m=1,\ldots,N) \tag{6-31}$$

in which $\underline{p} = (p_1,\ldots,p_N)$ is the forecaster's judgment and $Q_n(\underline{\pi})$ is the quadratic score for the climatological forecast $\underline{\pi}$ when state s_n obtains $(n=1,\ldots,N)$. Note that $r_m^* \neq p_m$ unless $\pi_n = 1/N$ for all n $(n=1,\ldots,N)$ or unless $p_m = 0$ or 1.

For a collection of K forecasts $\underline{r}_k$ $(k=1,\ldots,K)$, the average ratio skill score $\overline{RSS}(\underline{\pi},\underline{r},\underline{d})$ is defined as follows:

$$\overline{RSS}(\underline{\pi},\underline{r},\underline{d}) = 1 - [\overline{Q}(\underline{r},\underline{d})/\overline{Q}(\underline{\pi},\underline{d})], \tag{6-32}$$

where $\overline{Q}(\underline{r},\underline{d})$ is the average quadratic score for the forecasts $\underline{r}_k$ (k=1,...,K) and $\overline{Q}(\underline{\pi},\underline{d})$ is the average quadratic score for the climatological probabilities $\underline{\pi}$. The (actual) range of the measure $RSS(\underline{\pi},\underline{r},\underline{d})$ will be determined by the climatological probabilities $\underline{\pi}$ and the sample relative frequencies $\underline{\bar{d}}$. In order to determine whether the measure $RSS(\underline{\pi},\underline{r},\underline{d})$ is a "proper" or an improper scoring rule, we shall assume that the forecaster of concern has already made K forecasts $\underline{r}_k$ (k=1,...,K), that he wants to maximize his (average) expected score after the (K+1)st forecast $\underline{r}_{K+1}$, and that the (K+1)st forecast is independent of the previous K forecasts. Then, the "optimal" forecast is $\underline{r}^*_{K+1} = (r^*_{1,K+1},\ldots,r^*_{N,K+1})$, where

$$r^*_{m,K+1} = \{p_{m,K+1}/[K\overline{Q}(\underline{\pi},\underline{d}) + Q_m(\underline{\pi}_{K+1})]\}/ \sum_{n=1}^{N}\{p_{n,K+1}/[K\overline{Q}(\underline{\pi},\underline{d}) + Q_n(\underline{\pi}_{K+1})]\}, \quad (m=1,\ldots,N) \quad (6\text{-}33)$$

in which $\underline{p}_{K+1} = (p_{1,K+1},\ldots,p_{N,K+1})$ is the meteorologist's judgment on the (K+1)st occasion and $Q_n(\underline{\pi}_{K+1})$ is the quadratic score for a climatological forecast on the (K+1)st occasion if state s_n obtains (n=1,...,N) (Murphy, 1973a). Note that $r^*_{m,K+1} \neq p_{m,K+1}$ in (6-33) unless $\pi_{n,K+1} = 1/N$ for all n (n=1,...,N) or unless $p_{m,K+1} = 0$ or 1. Thus, the average ratio skill score $\overline{RSS}(\underline{\pi},\underline{r},\underline{d})$ is also an <u>improper</u> scoring rule. On the other hand, as the sample size K increases, the term $K\overline{Q}(\underline{\pi},\underline{d})$ in (6-33) increases and, for K large (e.g., $K \geq 100$), $K\overline{Q}(\underline{\pi},\underline{d}) >> Q_n(\underline{\pi}_{K+1})$ for all n (n=1,...,N). Then, (6-33) becomes

$$r^{*}_{m,K+1} \simeq [p_{m,K+1}/K\bar{Q}(\underline{\pi},\underline{d})]/\sum_{n=1}^{N}[p_{n,K+1}/K\bar{Q}(\underline{\pi},\underline{d})],$$

or, since $\sum_{n=1}^{N} p_{n,K+1} = 1$,

$$r_{m,K+1} \simeq p_{m,K+1}.$$

Thus, the measure $\overline{RSS}(\underline{\pi},\underline{r},\underline{d})$ is approximately strictly proper for large collections of forecasts.

(2) The information ratio. The information ratio $IR(\underline{\pi},\underline{r},\underline{d})$ (Holloway and Woodbury, 1955) can be defined as follows:

$$IR(\underline{\pi},\underline{r},\underline{d}) = \ln[(\underline{r}\underline{d}')/(\underline{\pi}\underline{d}')]/\ln(1/\underline{\pi}\underline{d}'),$$

or

$$IR(\underline{\pi},\underline{r},\underline{d}) = 1 - [\ln(\underline{r}\underline{d}')/\ln(\underline{\pi}\underline{d}')]. \qquad (6\text{-}34)$$

If $d_j = 1$, the measure $IR(\underline{\pi},\underline{r},\underline{d})$ in (6-34) becomes $IR_j(\underline{\pi},\underline{r})$, where

$$IR_j(\underline{\pi},\underline{r}) = 1 - (\ln r_j/\ln \pi_j).$$

The range of possible values of the measure $IR(\underline{\pi},\underline{r},\underline{d})$ is the open interval $(-\infty,\infty)$. However, as in the case of the ratio skill score $RSS(\underline{\pi},\underline{r},\underline{d})$, the actual range of $IR(\underline{\pi},\underline{r},\underline{d})$ is a closed interval with

a maximum value of one unless $r_j = 0$ or $\pi_j = 1$. The measure $IR(\underline{\pi},\underline{r},\underline{d})$ has a positive orientation.

The measure $IR(\underline{\pi},\underline{r},\underline{d})$ is an <u>improper</u> scoring rule. We prove this statement in Appendix B. Specifically, the optimal forecast is $\underline{r}^* = (r_1^*,\ldots,r_N^*)$, where

$$r_m^* = (p_m/\ln \pi_m)/\sum_{n=1}^{N}(p_n/\ln \pi_n). \quad (m=1,\ldots,N) \tag{6-35}$$

Note that $r_m^* \neq p_m$ in (6-35) unless $\pi_n = 1/N$ for all n (n=1,...,N) or unless $p_m = 0$ or 1. In Appendix B we also show that, under the same assumptions as those invoked for the ratio skill score, the average information ratio $\overline{IR}(\underline{\pi},\underline{r},\underline{d})$ is approximately strictly proper for large collections of forecasts.

(3) <u>The matrix skill score</u>. The matrix skill score $MSS(\underline{\pi},\underline{r},\underline{d})$ (Gringorten, 1965) can be defined as follows:

$$\begin{aligned} MSS(\underline{\pi},\underline{r},\underline{d}) = &\sum_{n=1}^{N} d_n \phi(r_n,\pi_n)(1/\pi_n) \\ &+ \sum_{n=1}^{N}(1 - d_n)\phi(\pi_n,r_n)[1/(1 - \pi_n)], \end{aligned} \tag{6-36}$$

where

$$\phi(x,y) = \begin{cases} 1 \text{ if } x > y, \\ 0 \text{ if } x < y. \end{cases}$$

The range of possible values of $MSS(\underline{\pi},\underline{r},\underline{d})$ is the half-open interval

$[0,\infty)$. However, unless $\pi_n = 0$ and $r_n > \pi_n$ for some n or unless $\pi_n = 1$ and $r_n < \pi_n$ for some n (n=1,...,N), the actual range of $MSS(\underline{\pi},\underline{r},\underline{d})$ will be a closed interval determined by the climatological probabilities $\underline{\pi} = (\pi_1,\ldots,\pi_N)$. This measure has a positive orientation.

The measure $MSS(\underline{\pi},\underline{r},\underline{d})$ is a proper, but not a strictly proper, scoring rule. We prove this statement in Appendix C.

(4) Comparison of relative validity measures. We are concerned with the relative merits of the distance measure $\overline{D}(\bar{\underline{d}},\underline{r},\underline{d})$, the ratio skill score $RSS(\underline{\pi},\underline{r},\underline{d})$, the information ratio $IR(\underline{\pi},\underline{r},\underline{d})$, and the matrix skill score $MSS(\underline{\pi},\underline{r},\underline{d})$. For individual forecasts, we have shown that the measures $RSS(\underline{\pi},\underline{r},\underline{d})$ and $IR(\underline{\pi},\underline{r},\underline{d})$ are improper scoring rules, while the measure $MSS(\underline{\pi},\underline{r},\underline{d})$ is only a proper scoring rule. As defined, the distance measure $\overline{D}(\bar{\underline{d}},\underline{r},\underline{d})$ is not an appropriate measure for individual forecasts or for very small collections of forecasts, since $\bar{\underline{d}}$ would not be a "good" estimate of $\underline{\pi}$ in such cases (see subsection 6.1.4). In these situations we shall replace $\bar{\underline{d}}$ with $\underline{\pi}$ in (6-16), and we shall denote the basic relative distance measure by $D(\underline{\pi},\underline{r},\underline{d})$. The measure $D(\underline{\pi},\underline{r},\underline{d})$ is, of course, a strictly proper scoring rule (see Murphy, 1973a). As the only scoring rule in this set of four measures which possesses this property, the distance measure $D(\underline{\pi},\underline{r},\underline{d})$ appears to be more appropriate than these other relative validity measures. Moreover, the measures $IR(\underline{\pi},\underline{r},\underline{d})$ and $MSS(\underline{\pi},\underline{r},\underline{d})$ both suffer from other possible deficiencies. The former depends only upon the forecast and climatological probabilities associated with the state that obtains, and thus represents only a

partial measure of relative validity. The latter, as a proper and thus discrete scoring rule, assigns only a relatively small number of distinct scores to individual forecasts (e.g., only three different scores can be assigned to a forecast in the two-state situation). Thus, the matrix skill score is often unable to distinguish between quite different forecasts.

For comparative purposes, we depict the measures $D(\underline{\pi},\underline{r},\underline{d})$, $RSS(\underline{\pi},\underline{r},\underline{d})$, $IR(\underline{\pi},\underline{r},\underline{d})$, and $MSS(\underline{\pi},\underline{r},\underline{d})$ in Figure 6-5 for an individual forecast in the two-state situation when (a) $\pi_1 = 0.2$ and (b) $\pi_1 = 0.5$. Since these measures are all symmetric in $\underline{r}$, the curves and lines corresponding to the situation in which $\pi_1 = 0.8$ are mirror images of the curves and lines in Figure 6-5a. Note that the measures $RSS(\underline{\pi},\underline{r},\underline{d})$ and $IR(\underline{\pi},\underline{r},\underline{d})$ can both assume very large negative values (in the case of the latter measure these values can actually be infinite) for some values of r_1, while the measure $MSS(\underline{\pi},\underline{r},\underline{d})$, as previously indicated, assumes (at most) three different values in this situation.

With regard to collections of forecasts, the distance measure $\overline{D}(\underline{\bar{d}},\underline{r},\underline{d})$ is strictly proper, while the matrix skill score $\overline{MSS}(\underline{\pi},\underline{r},\underline{d})$ is proper and the ratio skill score $\overline{RSS}(\underline{\pi},\underline{r},\underline{d})$ and the information ratio $\overline{IR}(\underline{\pi},\underline{r},\underline{d})$ are only approximately strictly proper for large collections of forecasts ($K \geq 100$). As a result, and in view of the comments with respect to deficiencies in these measures in the previous paragraphs, we believe that the distance measure $\overline{D}(\underline{\bar{d}},\underline{r},\underline{d})$ is the most appropriate relative measure of (primary) validity.

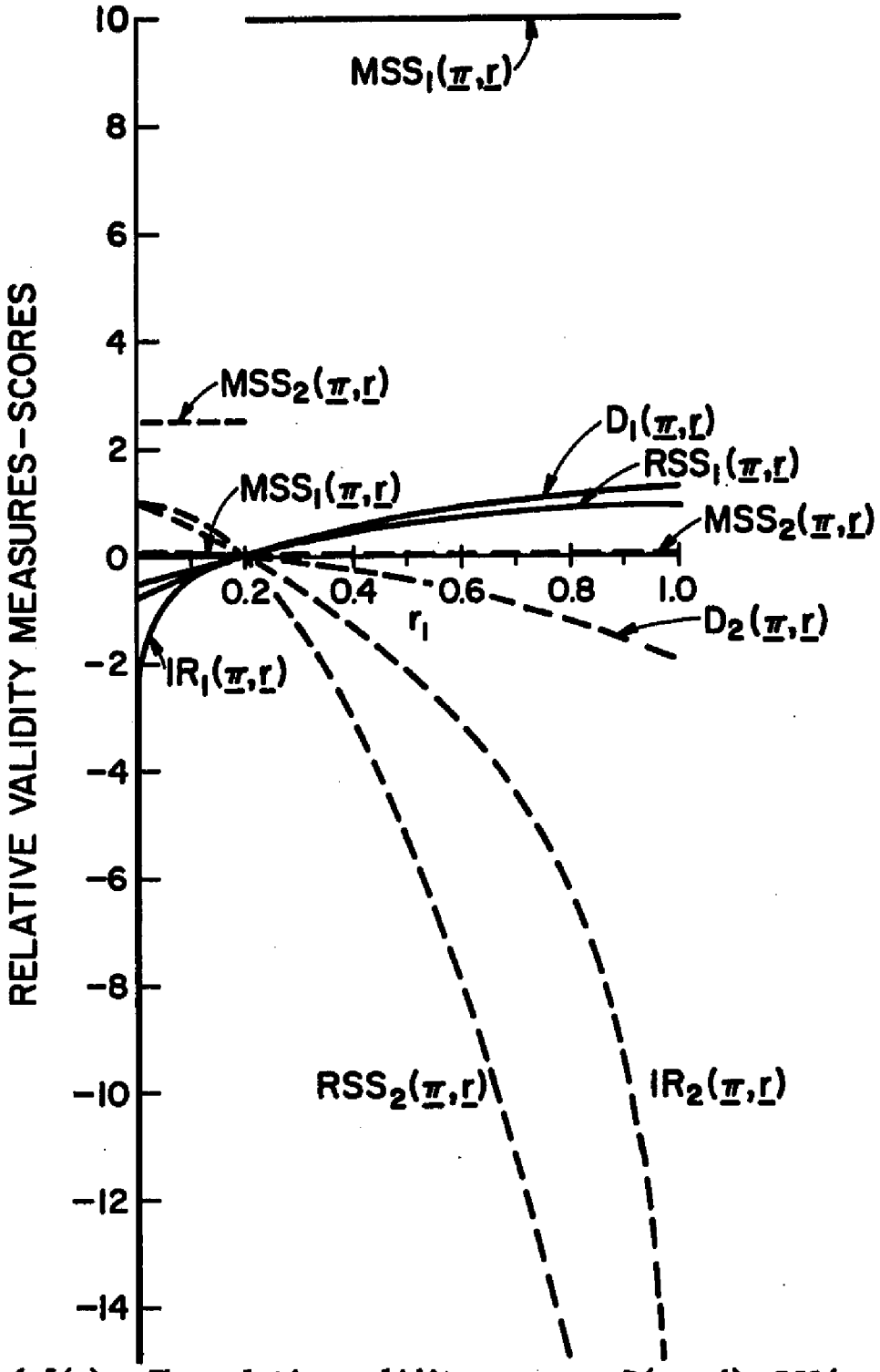


Figure 6-5(a). The relative validity measures $D(\underline{\pi},\underline{r},\underline{d})$, $RSS(\underline{\pi},\underline{r},\underline{d})$, $IR(\underline{\pi},\underline{r},\underline{d})$, and $MSS(\underline{\pi},\underline{r},\underline{d})$ in the two-state situation when $\pi_1 = 0.2$. The solid curves and lines denote $D_1(\underline{\pi},\underline{r})$, $RSS_1(\underline{\pi},\underline{r})$, $IR_1(\underline{\pi},\underline{r})$, and $MSS_1(\underline{\pi},\underline{r})$ and the dashed curves and lines denote $D_2(\underline{\pi},\underline{r})$, $RSS_2(\underline{\pi},\underline{r})$, $IR_2(\underline{\pi},\underline{r})$, and $MSS_2(\underline{\pi},\underline{r})$.

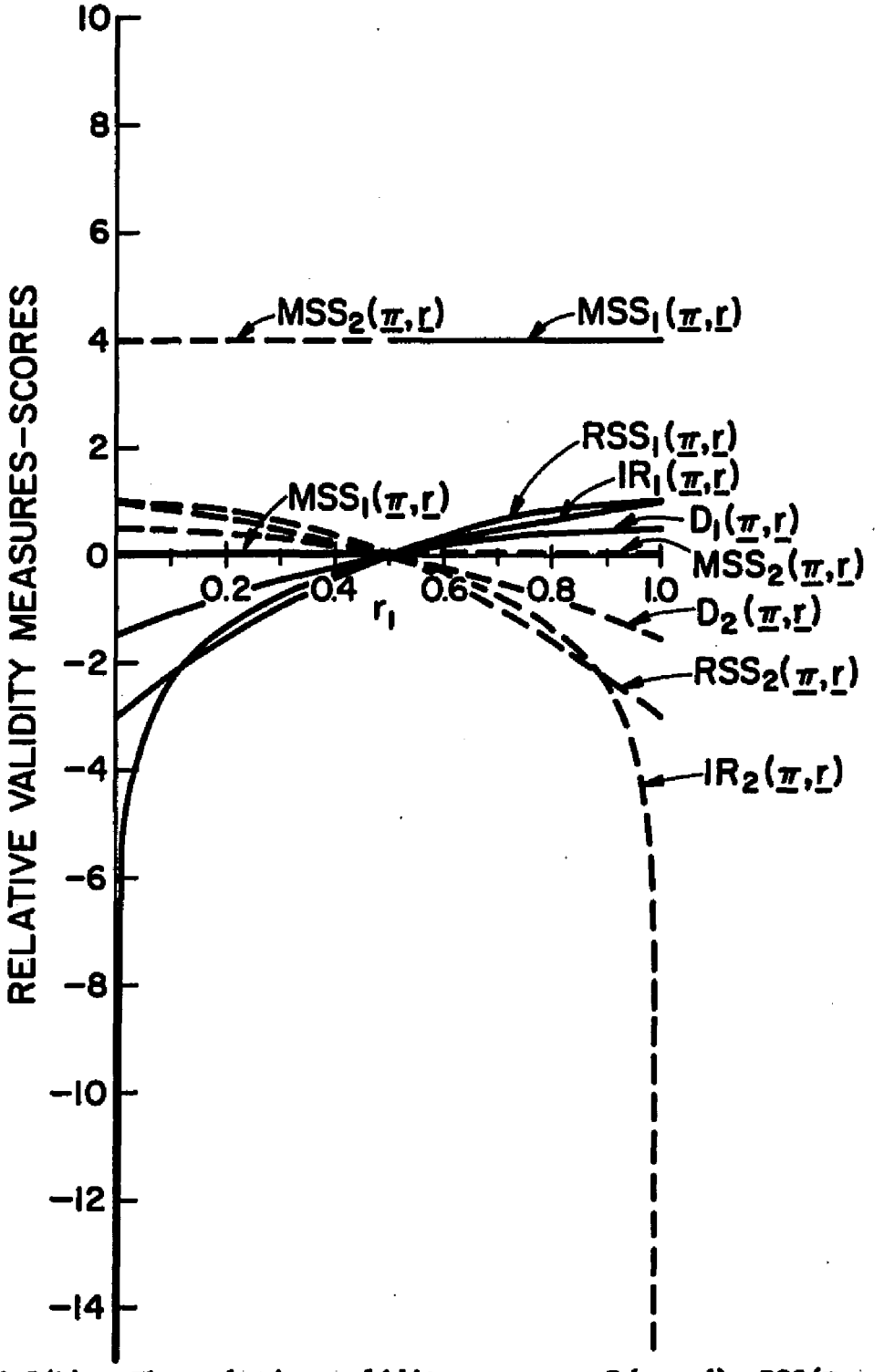


Figure 6-5(b). The relative validity measures $D(\underline{\pi},\underline{r},\underline{d})$, $RSS(\underline{\pi},\underline{r},\underline{d})$, $IR(\underline{\pi},\underline{r},\underline{d})$, and $MSS(\underline{\pi},\underline{r},\underline{d})$ in the two-state situation when $\pi_1 = 0.5$. The solid curves and lines denote $D_1(\underline{\pi},\underline{r})$, $RSS_1(\underline{\pi},\underline{r})$, $IR_1(\underline{\pi},\underline{r})$, and $MSS_1(\underline{\pi},\underline{r})$ and the dashed curves and lines denote $D_2(\underline{\pi},\underline{r})$, $RSS_2(\underline{\pi},\underline{r})$, $IR_2(\underline{\pi},\underline{r})$, and $MSS_2(\underline{\pi},\underline{r})$.

6.3.2 Measures of Bias

In this subsection we examine and compare two measures of the attribute bias (in-the-small) or reliability. The (basic) distance measure of bias, $\overline{\Delta^2}(\underline{r}^t,\underline{\bar{d}}^t)$ [see (6-11)], is based upon a partition of the primitive distance measure $\overline{\Delta^2}(\underline{r},\underline{d})$ in which each forecast in the collection of K "forecasts" is considered to be a vector consisting of N probabilities (see Section 6.2). This measure represents the square of the weighted distance between these vectors ($\underline{r}^t$) and the corresponding relative frequency vectors ($\underline{\bar{d}}^t$). We shall refer to this measure as the <u>vector</u> measure of bias.

A <u>scalar</u> measure of bias can be defined in an analogous manner if the collection of K (vector) forecasts is considered to consist of NK scalar forecasts (each probability is considered to be a distinct forecast; see Chapter 3). This measure of bias represents the square of the weighted distance between these scalar forecasts (r^s) and the corresponding scalar relative frequencies ($\bar{d}^s$). We shall denote this measure by $\overline{SMB}(r^s,\bar{d}^s)$, where

$$\overline{SMB}(r^s,\bar{d}^s) = (1/NK)\sum_{s=1}^{S}(NK)^s(r^s-\bar{d}^s)^2. \qquad (6\text{-}37)$$

Murphy (1972b) has investigated the relationships between these scalar and vector measures of bias in N-state ($N\geq 2$) situations and has shown that

$$\overline{\Delta^2}(\underline{r}^t, \underline{\bar{d}}^t) \geq \overline{SMB}(r^s, \bar{d}^s), \qquad (6\text{-}38)$$

with equality holding in (6-38) only if the observed relative frequencies, $\underline{\bar{d}}^t$ and $\bar{d}^s$, satisfy certain very restrictive conditions (Murphy, op. cit., p. 1186). Thus, in general, the bias of a collection of "forecasts" will be greater (i.e., poorer) when the collection is considered to consist of K vector forecasts than when the collection is considered to consist of NK scalar forecasts.

Murphy (op. cit., pp. 1188-1190) has discussed the relative merits of scalar and vector partitions and concludes that the latter appears to be more appropriate than the former from both scientific and economic points of view. With regard to scientific considerations, evaluators are, or should be, primarily concerned with the bias of forecasts rather than the bias of probabilities. In addition, the vector partition is, and the scalar partition is not, formulated in such a way that a one-to-one correspondence exists between the probabilities and the states for each subcollection of forecasts. With regard to economic considerations, the framework for operational evaluation (see Chapter 4) is such that forecasts, in order to be meaningful, must be represented as vectors rather than as scalars. The only apparent problem with the vector approach is that the size of the basic collection of forecasts and the relevant subcollections is reduced when the forecasts are considered to

be vectors rather than scalars (i.e., the collection consists of NK forecasts in the case of scalar forecasts and only K forecasts in the case of vector forecasts). However, this problem can be "solved" by combining subcollections of adjacent forecasts (within the vector framework). In view of the fact that vector forecasts are obviously more appropriate than scalar forecasts in a meteorological context, we believe that the appropriate measure of bias or reliability is the distance measure $\overline{\Delta^2}(\underline{r}^t,\underline{\bar{d}}^t)$ (or possibly a suitable linear transformation thereof).

6.3.3 Measures of Resolution

In this section we briefly examine and compare three measures of the attribute resolution. The (basic) distance measure of resolution, $\overline{\Delta^2}(\underline{\bar{d}},\underline{\bar{d}}^t)$ [see (6-12)], is based upon a new vector partition of the primitive distance measure $\overline{\Delta^2}(\underline{r},\underline{d})$ (see Murphy, 1973b). This measure represents the square of the weighted distance between the relative frequency vector for the entire collection of forecasts, $\underline{\bar{d}}$, and the relative frequency vectors corresponding to the T subcollections of forecasts, $\underline{\bar{d}}^t$. We shall refer to this measure as the new vector measure of resolution.

The original scalar and vector partitions of the primitive distance measure $\overline{\Delta^2}(\underline{r},\underline{d})$ [or, equivalently, of the quadratic or probability score $\overline{Q}(\underline{r},\underline{d})$], formulated by Murphy (1972a,b), provide scalar and vector measures of resolution. If we denote these measures by $\overline{SMR}(\bar{d}^s)$ and $\overline{VMR}(\underline{\bar{d}}^t)$, then

$$\overline{\mathrm{SMR}}(\bar{d}^s) = (1/NK)\sum_{s=1}^{S}(NK)^s\bar{d}^s(1 - \bar{d}^s), \qquad (6\text{-}39)$$

and

$$\overline{\mathrm{VMR}}(\underline{\bar{d}}^t) = (1/K)\sum_{t=1}^{T}\underline{\bar{d}}^t(\underline{u} - \underline{\bar{d}}^t)', \qquad (6\text{-}40)$$

respectively. The vector $\underline{u}$ is the unity vector, i.e., $\underline{u} = (1,\ldots,1)$. Murphy (op. cit.) has shown that

$$\overline{\mathrm{VMR}}(\underline{\bar{d}}^t) \leq \overline{\mathrm{SMR}}(\bar{d}^s), \qquad (6\text{-}41)$$

with equality holding in (6-41), as in the case of the scalar and vector measures of bias, only under very restrictive conditions on the relative frequencies $\underline{\bar{d}}^t$ and $\bar{d}^s$. Thus, in general, the resolution of a collection of "forecasts" will be smaller (i.e., better) when the collection is considered to consist of K vector forecasts than when the collection is considered to consist of NK scalar forecasts. The arguments in favor of the vector partition (vis-a-vis the scalar partition) presented in subsection 6.3.2 are equally applicable in this situation. Thus, the (original) vector measure of resolution appears to be more appropriate than the scalar measure of resolution.

With regard to the relationship between the two vector measures of resolution, Murphy (1973b) has shown that the measure of uncertainty $\overline{\Delta^2}(\underline{\bar{d}},\underline{d})$ [see (6-10)] is the sum of these two measures. That

is,

$$\overline{\Delta^2}(\bar{\underline{d}},\underline{d}) = \overline{VMR}(\bar{\underline{d}}^t) + \overline{\Delta^2}(\bar{\underline{d}},\bar{\underline{d}}^t). \qquad (6\text{-}42)$$

Thus, these resolution measures are simply two components of the basic uncertainty of the relevant observations. With regard to the relative merits of these two measures, no definitive criteria exist upon which to base the choice of a particular measure. We prefer the new vector measure of resolution, $\overline{\Delta^2}(\bar{\underline{d}},\bar{\underline{d}}^t)$, because (1) this measure arises as a result of a particularly useful partition of the basic distance measure $\overline{\Delta^2}(\underline{r},\underline{d})$ and (2) this measure can be interpreted directly in terms of a weighted distance [the original measure of resolution can only be interpreted as a distance measure indirectly via (6-42)].

6.3.4 Measures of Sharpness

A measure of the attribute sharpness was not defined in Section 6.2, in which the basic distance measures of the other attributes were described. However, a measure of sharpness can be readily formulated within the framework of concern in this study if we consider the reference points on the scale of sharpness (see Chapter 3). In this regard, if we let $D(\underline{r},\underline{e})$ denote the basic measure of sharpness, then

$$D(\underline{r},\underline{e}) = (\underline{r} - \underline{e})(\underline{r} - \underline{e})' = \sum_{n=1}^{N}(r_n - e_n)^2, \qquad (6\text{-}43)$$

where $\underline{e} = (e_1,\ldots,e_N) = (1/N,\ldots,1/N)$. This measure represents the square of the distance between the meteorologist's forecast $\underline{r}$ and the "uniform forecast" $\underline{e}$ (the latter is completely lacking in sharpness). The choice of a measure involving the square of the distance is, of course, largely an arbitrary decision in this situation (the concept of "proper" scoring rules is not applicable to measures of sharpness). We have chosen this particular measure because the basic measures of the other attributes are all squared distance measures. The range of the measure $D(\underline{r},\underline{e})$ is the closed interval $[0,(N-1)/N]$ and $D(\underline{r},\underline{e})$ has a positive orientation. If we let $D'(\underline{r},\underline{e})$ denote a measure that is linearly related to $D(\underline{r},\underline{e})$ and that has a unit interval range, then

$$D'(\underline{r},\underline{e}) = [N/(N-1)](\underline{r} - \underline{e})(\underline{r} - \underline{e})'. \qquad (6\text{-}44)$$

Another possible measure of sharpness is the information measure $IM(\underline{r})$ (see Murphy and Epstein, 1967a), where

$$IM(\underline{r}) = -\sum_{n=1}^{N} r_n \ln r_n. \qquad (6\text{-}45)$$

The range of the measure $IM(\underline{r})$ is the closed interval $[0,\ln N]$ and $IM(\underline{r})$ has a negative orientation. If we let $IM'(\underline{r})$ denote a measure that is linearly related to $IM(\underline{r})$ and that has a unit interval range and a positive orientation, then

$$IM'(\underline{r}) = 1 + (1/\ln N)\sum_{n=1}^{N} r_n \ln r_n. \qquad (6\text{-}46)$$

We depict the measures $D'(\underline{r},\underline{e})$ and $IM'(\underline{r})$ in Figure 6-6 for the two-state situation. The scores assigned by these two measures are obviously quite similar. Thus, while the distance measure $D'(\underline{r},\underline{e})$ is a "natural" measure of sharpness within the framework adopted in this study, no definitive criteria exist upon which to base the choice of a particular one of these measures.

6.4 Empirical Evaluation Measures: A Brief Summary

The basic distance measures of the principal attributes, (primary) validity and bias (in-the-small), and several secondary attributes were defined in Sections 6.2 and 6.3. These measures can be considered to be "natural" measures of the attributes within the framework for empirical evaluation described in Chapter 4. These distance measures were then compared with the measures of these attributes proposed heretofore by other investigators. Unfortunately, few definitive criteria exist in most cases upon which to base the choice of a particular measure. However, one important criterion is whether or not the measure is strictly proper, but this criterion only applies to measures of those attributes which are defined for individual forecasts and which depend upon the state that obtains. In this regard, the relevant distance measures are all strictly proper scoring rules.

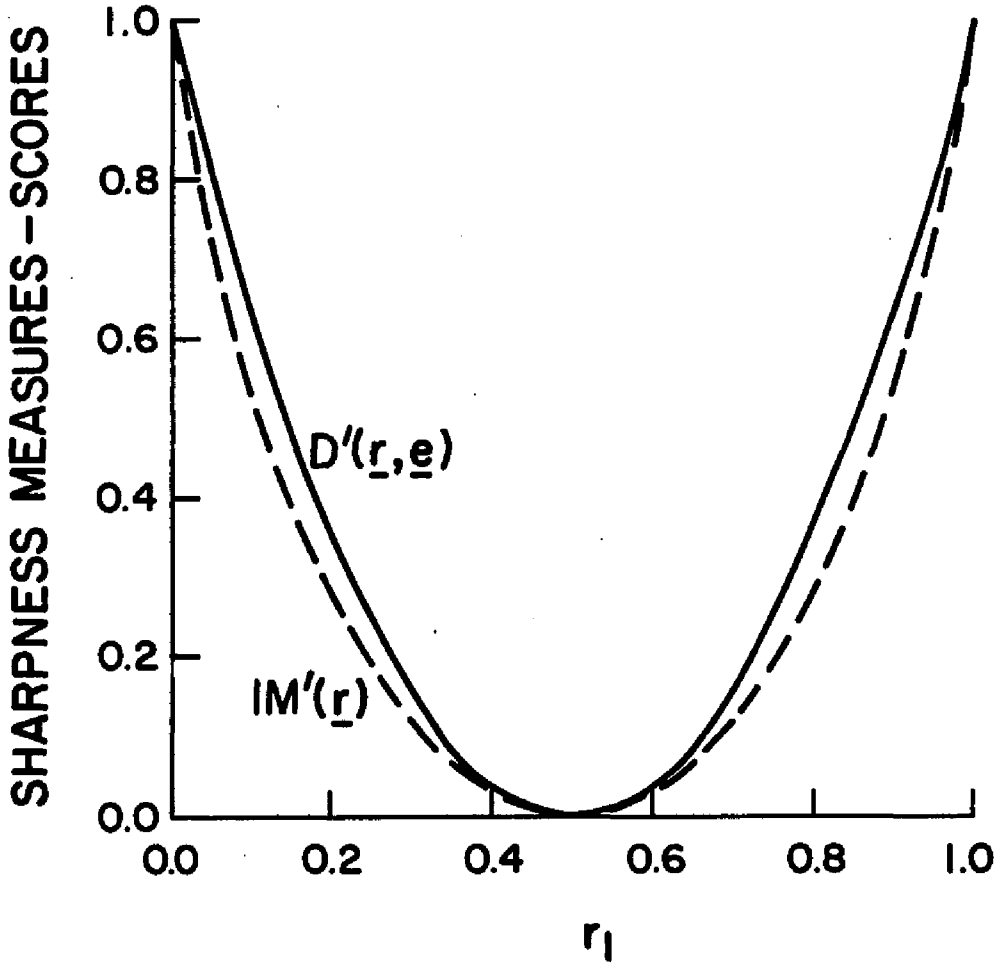


Figure 6-6. The sharpness measures D'(r,e) and IM'(r) in the two-state situation. The solid curve denotes D'(r,e) and the dashed curve denotes IM'(r).

With regard to absolute measures of validity, several strictly proper scoring rules have been proposed (see paragraph 6.3.1.1), and each measure seems to possess certain desirable properties (and to have its proponents). For the purposes of this study, in which we are concerned with the relationships between measures of validity and measures of utility and expected utility (see Chapter 8), any of these strictly proper measures can be used as the "standard" measure of validity. With regard to relative measures of validity, only the distance measure is strictly proper, although two other measures are approximately strictly proper for large collections of forecasts ($K \geq 100$). For summary purposes, we describe the form and properties of the absolute and relative measures of (primary) validity considered in this study in Table 6-3a.

Scalar and vector measures of the principal attribute bias (in-the-small) and the supplemental attribute resolution were described. The vector measures of both of these attributes are preferred, because we believe that forecasts should be considered to be vectors rather than scalars (see subsection 6.3.2). With regard to the resolution measures, we prefer the measure of this attribute which is obtained from the new vector partition of the basic distance measure of validity (see subsection 6.3.3). Finally, two measures of sharpness were briefly described and compared (see subsection 6.3.4), although this attribute does not appear to be particularly appropriate for empirical evaluation (see Chapter 3). To complete

Key to Table 6-3(a)

[1]P = Positive, N = Negative

[2]SP = Strictly Proper, P = Proper, I = Improper

[3]The absolute measures are presented in their original form.

[4]The relative measures are defined for individual forecasts. See subsection 6.3.1.2 for definitions of these measures for collections of forecasts.

[5]These measures are approximately strictly proper for large collections of forecasts ($K \geq 100$).

[6]$\phi(x,y) = 1$ if $x > y$ and $\phi(x,y) = 0$ if $x < y$.

Table 6-3(a). The name, form, and properties of the absolute and relative measures of (primary) validity considered in this study (see subsection 6.3.1).

Table 6-3(a)

Name	Form	Range	Orientation[1]	SP/P/I[2]
(1) **Absolute Measures**[3]				
Distance Measure	$D(\underline{r},\underline{d}) = 1 - (1/2)(\underline{r}-\underline{d})(\underline{r}-\underline{d})'$	[0,1]	P	SP
Linear Measure	$APS(\underline{r},\underline{d}) = \sum_{n=1}^{N} \lvert r_n - d_n \rvert$	[0,1]	N	I
Quadratic Measure	$Q(\underline{r},\underline{d}) = (\underline{r}-\underline{d})(\underline{r}-\underline{d})'$	[0,2]	N	SP
Logarithmic Measure	$L(\underline{r},\underline{d}) = \ln(\underline{r}\underline{d}')$	$(-\infty,0]$	P	SP
Spherical Measure	$S(\underline{r},\underline{d}) = \underline{r}\underline{d}'/(\underline{r}\underline{r}')^{1/2}$	[0,1]	P	SP
Ranked Probability Score	$RPS(\underline{r},\underline{d}) = \sum_{n=1}^{N}(\sum_{m=1}^{n} r_m - \sum_{m=1}^{n} d_m)^2$	[0,N-1]	N	SP
(2) **Relative Measures**[4]				
Distance Measure	$D(\underline{\pi},\underline{r},\underline{d}) = (1/2)+(1/2)[D(\underline{r},\underline{d})-D(\underline{\pi},\underline{d})]$	[0,1]	P	SP
Ratio Skill Score	$RSS(\underline{\pi},\underline{r},\underline{d}) = 1-[Q(\underline{r},\underline{d})/Q(\underline{\pi},\underline{d})]$	$(-\infty,1]$	P	I[5]
Information Ratio	$IR(\underline{\pi},\underline{r},\underline{d}) = 1-[\ln(\underline{r}\underline{d}')/\ln(\underline{\pi}\underline{d}')]$	$(-\infty,\infty)$	P	I[5]
Matrix Skill Score	$MSS(\underline{\pi},\underline{r},\underline{d}) = \sum_{n=1}^{N} d_n \phi(r_n,\pi_n)(1/\pi) + \sum_{n=1}^{N}(1-d_n)\phi(\pi_n,r_n)[1/(1-\pi_n)]$[6]	$[0,\infty)$	P	P

this brief summary of empirical evaluation measures, we describe the form and properties of the measures of bias (or reliability), resolution, and sharpness considered in this study in Table 6-3b.

Key to Table 6-3(b)

[1]P = Positive, N = Negative

[2]The vector $\underline{u}$ is the unity vector, i.e., $\underline{u} = (1,\ldots,1)$.

[3]The vector $\underline{e}$ is the "uniform" vector, i.e., $\underline{e} = (1/N,\ldots,1/N)$.

Table 6-3(b). The name, form, and properties of the measures of bias, resolution, and sharpness considered in this study (see subsections 6.3.2, 6.3.3, and 6.3.4, respectively).

Table 6-3(b)

Name	Form	Properties: Range	Properties: Orientation[1]
(1) Bias Measures			
Vector Measure	$\overline{\Delta^2}(\underline{r}^t,\underline{\bar{d}}^t) = (1/K)\sum_{t=1}^{T} K^t(\underline{r}^t-\underline{\bar{d}}^t)(\underline{r}^t-\underline{\bar{d}}^t)'$	[0,2]	N
Scalar Measure	$\overline{SMB}(r^s,\bar{d}^s) = (1/NK)\sum_{s=1}^{S}(NK)^s(r^s-\bar{d}^s)^2$	[0,1]	N
(2) Resolution Measures			
New Vector Measure	$\overline{\Delta^2}(\underline{\bar{d}}^t,\underline{\bar{d}}) = (1/K)\sum_{t=1}^{T} K^t(\underline{\bar{d}}^t-\underline{\bar{d}})(\underline{\bar{d}}^t-\underline{\bar{d}})'$	[0,(N-1)/N]	P
Original Vector Measure	$\overline{VMR}(\underline{\bar{d}}^t) = (1/K)\sum_{t=1}^{T} K^t\bar{d}^t(\underline{u}-\underline{\bar{d}}^t)'^2$	[0,(N-1)/N]	N
Scalar Measure	$\overline{SMR}(\bar{d}^s) = (1/NK)\sum_{s=1}^{S}(NK)^s\bar{d}^s(1-\bar{d}^s)$	[0,1/4]	N
(3) Sharpness Measures			
Distance Measure	$D(\underline{r},\underline{e}) = (\underline{r}-\underline{e})(\underline{r}-\underline{e})'^3$	[0,(N-1)/N]	P
Information Measure	$IM(\underline{r}) = -\sum_{n=1}^{N} r_n \ln r_n$	[0,ln N]	P

CHAPTER SEVEN. OPERATIONAL EVALUATION MEASURES

In this chapter general measures of the attribute utility are defined for situations in which the evaluator's knowledge is complete (Section 7.1) and incomplete (Section 7.2). Specific measures corresponding to certain special situations are considered as examples. In formulating the specific measures described in Section 7.2, the evaluator's knowledge is assumed to be expressed in terms of the inequality and/or probability measures considered in Chapter 5. Several important properties of operational evaluation measures are discussed in Section 7.3.

7.1 Knowledge of Utilities Complete

In this section we assume that the evaluator's knowledge of the decision maker's utilities is complete. That is, the evaluator is assumed to be able to specify the decision maker's utilities in a precise, or categorical, manner (see Chapter 5). First we define the basic utility measure, and then we describe the utility measures associated with two specific decision situations.

7.1.1 The Utility Measure: Definition

Within the framework of the modified SEU model of a decision situation (see Chapter 4), the decision maker has been assumed to adopt the forecasters probabilities as his subjective probabilities

on the relevant states. Thus, the action selected and, as a result, the consequence subsequently obtained by the decision maker depend upon the meteorologist's forecast. The utility of the consequence that obtains is, then, a reasonable measure of the "value" of the forecast to the decision maker. Such a measure depends upon the meteorologist's forecast, $\underline{r}$, the decision maker's utilities, $\underline{U} = (u_{mn})$, and the observation defined on the states, $\underline{d}$. If we denote this utility measure by $U(\underline{r},\underline{d})$, then

$$U(\underline{r},\underline{d}) = \sum_{m=1}^{M} \delta_m \left(\sum_{n=1}^{N} d_n u_{mn} \right), \tag{7-1}$$

where

$$\delta_m = \begin{cases} 1 & \text{if } \underline{r} \in R_m^N, \\ 0 & \text{otherwise.} \end{cases}$$

Equation (7-1) defines the general utility measure of concern in this study. Note that when $\underline{r} \in R_m^N$ and $d_n = 1$ (i.e., when the forecast $\underline{r}$ is such that the decision maker selects action a_m and state s_n obtains), $U(\underline{r},\underline{d})$ is simply u_{mn}.

Since the utilities of concern are standard utilities (see Chapter 5), the range of the utility measure $U(\underline{r},\underline{d})$ is the closed unit interval [0,1] and $U(\underline{r},\underline{d})$ has a positive orientation.

7.1.2 The Utility Measure: Examples

For illustrative purposes, the utility measure associated with two specific decision situations are described in this subsection. First, consider the three-action, two-state decision situation depicted in Figure 4-6. The standard utility matrix corresponding to this situation has been presented in Figure 5-1a. The utility measure $U(\underline{r},\underline{d})$, a step function with two steps in this situation, is depicted in Figure 7-1. Note, for example, that $U_1(\underline{r}) = 0$ for $0 \leq r_1 < 1/3$, $U_1(\underline{r}) = 2/3$ for $1/3 < r_1 < 1/2$, and $U_1(\underline{r}) = 1$ for $1/2 < r_1 < 1$. We shall assume that the value of $U_n(\underline{r})$ at each step is defined simply as the average of the two "limiting" values of $U_n(\underline{r})$ at that step [i.e., the values of $U_n(\underline{r})$ as the step is approached from the left and the right] (n=1,2), and these values are represented by large dots in Figure 7-1. Then, $U_1(\underline{r}) = 1/6$ for $r_1 = 1/3$ and $U_1(\underline{r}) = 5/6$ for $r_1 = 1/2$. In any case, the values of the measure $U(\underline{r},\underline{d})$ (as well as the values of the expected-utility measures; see Section 7.2) at these steps are not of particular concern in this study.

Second, consider the four-action, three-state decision situation depicted in Figure 4-8, the standard utility matrix for which is presented in Figure 5-1b. The utility measure $U(\underline{r},\underline{d})$, a "plateau" function in this situation, is depicted schematically in Figure 7-2. Specifically, the figures in square brackets are the values of $U_1(\underline{r})$, $U_2(\underline{r})$, and $U_3(\underline{r})$ when $\underline{r} \in R_m^3$ (m=1,2,3,4) and represent the

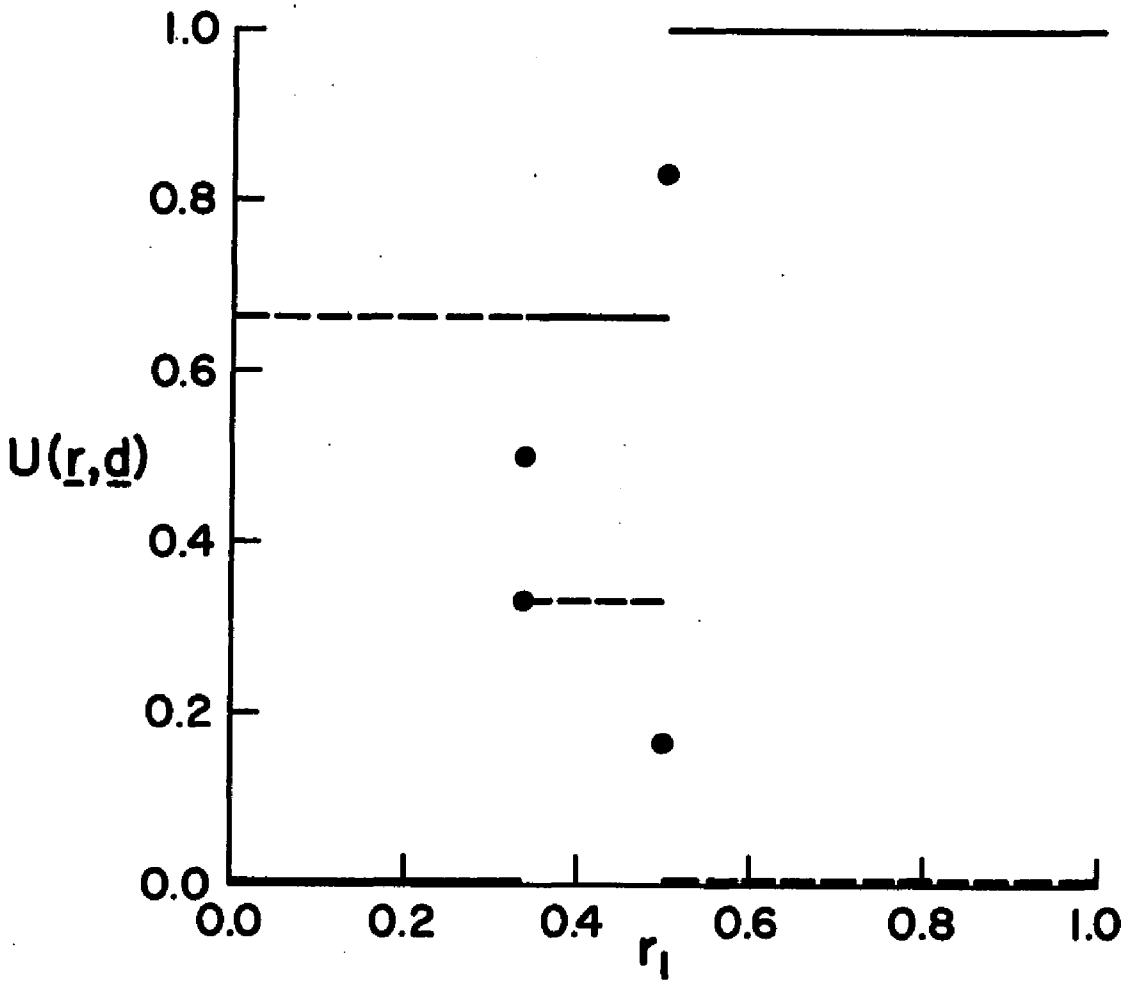


Figure 7-1. The utility measure $U(\underline{r},\underline{d})$ in the decision situation depicted in Figure 4-6 with standardized utilities (see Figure 5-1a). The solid lines denote $U_1(\underline{r})$, while the dashed lines denote $U_2(\underline{r})$. The large dots denote the values of $U(\underline{r},\underline{d})$ at each "step" (see text).

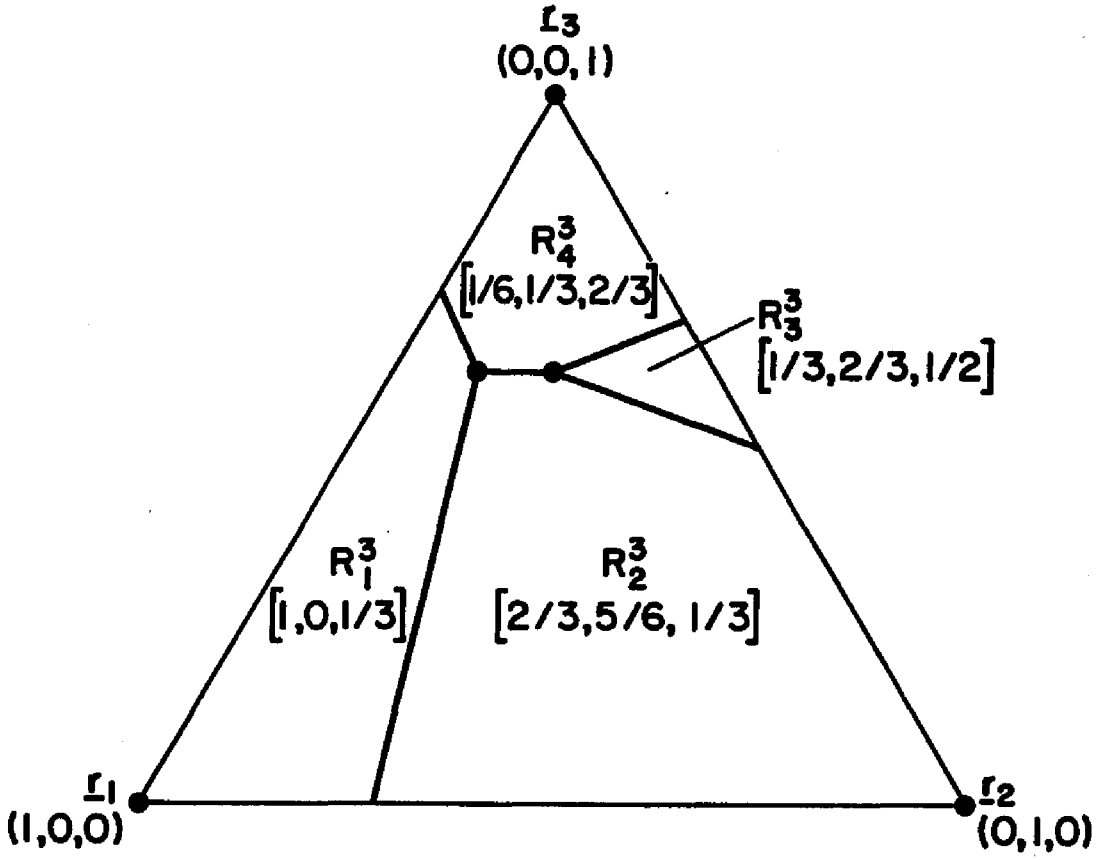


Figure 7-2. The utility measure $U(\underline{r},\underline{d})$ in the decision situation depicted in Figure 4-8 with standardized utilities (see Figure 5-1b). The figures in square brackets represent the values of $U_1(\underline{r})$, $U_2(\underline{r})$, and $U_3(\underline{r})$ when $\underline{r} \in R_m^3$ (m=1,2,3,4).

"heights" of the plateaus in the respective regions of the simplex R^3. Thus, for example, when $\underline{r} \in R_1^3$, $U_1(\underline{r}) = 1$, $U_2(\underline{r}) = 0$, and $U_3(\underline{r}) = 1/3$. As before, the value of $U(\underline{r},\underline{d})$ at the boundaries of two (or more) regions will be defined as the average of the limiting values in the respective regions of concern. Thus, for $\underline{r} \in R_{12}^3$, $U_1(\underline{r}) = 5/6$, $U_2(\underline{r}) = 5/12$, and $U_3(\underline{r}) = 1/3$.

7.2 Knowledge of Utilities Incomplete

In this section we assume that the evaluator's knowledge of the utilities is incomplete. That is, the evaluator is assumed to be able to specify the decision maker's utilities in either an incomplete or a probabilistic manner (see Chapter 5). We first consider the nature of the appropriate evaluation measures when the evaluator's knowledge is expressed in terms of these two classes of measures. Then, we define the basic expected-utility measure, the appropriate measure when the evaluator's knowledge is expressed in probabilistic terms. Finally, we describe the utility and expected-utility measures associated with three specific decision situations.

7.2.1 Utility and Expected-Utility Measures

In Chapter 5 we identified two classes of measures in terms of which the evaluator could express his knowledge of the decision maker's utilities. The class of inequality measures (Class I) corresponded to the situation in which the evaluator's knowledge was incomplete, while the class of probability measures (Class II)

corresponded to the situation in which the evaluator's knowledge was expressed in probabilistic terms. Different evaluation measures are appropriate in these two situations. In this regard, inequality measures simply provide bounds on the values of the utilities $\underline{U} = (u_{mn})$. As we shall discover in subsection 7.2.3 (via the examples), these "restrictions" lead to bounds on the utility of the forecasts. However, the utility measure $U(\underline{r},\underline{d})$ is (still) the appropriate evaluation measure in this situation, and this measure has been described in subsection 7.1.1.

On the other hand, in situations in which the evaluator's knowledge is expressed in terms of a probability distribution the utilities $\underline{U} = (u_{mn})$ are considered to be random variables. In such situations an appropriate measure of the "value" of the forecasts would be an expected-utility measure, where the expectation is taken over the utility measure $U(\underline{r},\underline{d})$ with respect to the probability distribution on the relevant utilities. We define and describe the basic expected-utility measure of concern in this study in the following subsection.

7.2.2 The Expected-Utility Measure: Definition

As indicated above, when the evaluator's knowledge of the utilities is expressed in terms of a probability distribution (a Class II measure), the utilities are random variables and the appropriate measure of utility is an expected-utility measure. If we denote this expected-utility measure by $EU(\underline{r},\underline{d})$, then

$$EU(\underline{r},\underline{d}) = \int_{\Theta} U(\underline{r},\underline{d}) f(\underline{U}) d\underline{U}, \qquad (7\text{-}2)$$

or, from (7-1),

$$EU(\underline{r},\underline{d}) = \int_{\Theta} [\sum_{m=1}^{M} \delta_m (\sum_{n=1}^{N} d_n u_{mn})] f(\underline{U}) d\underline{U}, \qquad (7\text{-}3)$$

where $f(\underline{U})$ denotes a continuous probability distribution defined on the elements of the utility matrix $\underline{U}$ and Θ is the region over which the density function $f(\underline{U})$ is defined. If the distribution is discrete, then the integrals in (7-2) and (7-3) will be replaced by summations and the probability element $f(\underline{U})d\underline{U}$ will be replaced by the probability function $p(\underline{U})$. Equations (7-2) and (7-3) define the general expected-utility measure of concern in this study.

Since the utilities are standard utilities, the range of the expected-utility measure $EU(\underline{r},\underline{d})$ is the closed unit interval [0,1] and $EU(\underline{r},\underline{d})$ has a positive orientation.

7.2.3 The Utility and Expected-Utility Measures: Examples

In order to provide specific examples of utility and expected-utility measures in situations in which the evaluator's knowledge is incomplete, the following decision situations are considered:

(1) the general two-action, two-state decision situation described by one of the four admissible key utility matrices in Figure 5-2;

(2) the special two-action, two-state decision situation described by one of the two admissible key utility matrices in Figure 5-3; and

(3) a special three-action, three-state decision situation.

In each of these situations we describe the utility and/or expected-utility measures corresponding to one or more measures of the evaluator's knowledge.

7.2.3.1 A General Two-Action, Two-State Decision Situation

Consider the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2. In this situation, the indifference point is $\underline{r}^* = (r_1^*, r_2^*)$, where $r_1^* = x_2/(1 - x_1 + x_2)$ and $r_2^* = 1 - r_1^*$, and the utility measure $U(\underline{r},\underline{d})$ is

$$U(\underline{r},\underline{d}) = \delta_1 d_1 + \delta_1(d_1 x_1 + d_2 x_2). \qquad (7\text{-}4)$$

With regard to the evaluator's knowledge of the utilities x_1 and x_2, we shall consider the following specific measures:

(a) a bounded interval measure (a Class I measure);

(b) a bounded interval - uniform measure (a Class IIb measure); and

(c) a beta measure (a Class IIb measure).

(1) Bounded interval measure. This measure defines upper and lower limits on the values of the utilities, namely $\alpha_1 \leq x_1 \leq \beta_1$ and

$\alpha_2 \leq x_2 \leq \beta_2$ [$0 \leq \alpha_i \leq \beta_i \leq 1$ (i=1,2)]. These bounds, together with the expression for the indifference point lead to the identification of bounds on, or specific values for, the utility measure $U(\underline{r},\underline{d})$. These bounds and values are presented in Table 7-1, both for (a) general intervals and (b) specific intervals [namely, $(\alpha_1,\beta_1) = (0.1,0.6)$ and $(\alpha_2,\beta_2) = (0.3,0.8)$]. Note that this (knowledge) measure will provide bounds on, and/or specific values for, the measure $U(\underline{r},\underline{d})$ for at least some values of r_1 when $\alpha_2 \neq 0$ and/or $\beta_1 \neq 1$. The utility measure $U(\underline{r},\underline{d})$ for the specific intervals is depicted in Figure 7-3.

(2) Bounded interval - uniform measure. This measure defines a uniform probability distribution over the region associated with the corresponding bounded interval measure. The expected-utility measure $EU(\underline{r},\underline{d})$ in this situation is

$$EU(\underline{r},\underline{d}) = \begin{cases} d_1 & \text{if } r_1 \leq \alpha_2/(1-\alpha_1+\alpha_2), \\ d_1E(x_1) + d_2E(x_2) & \text{if } r_1 \geq \beta_2/(1-\beta_1+\beta_2), \\ \int\int_{(x_1,x_2)\in\Theta} [\delta_1 d_1 + \delta_2(d_1x_1 + d_2x_2)] \cdot f(x_1,x_2)dx_1dx_2 & \text{otherwise,} \end{cases} \quad (7\text{-}5)$$

where $E(x_i) = (1/2)(\alpha_i + \beta_i)$ (i=1,2), $f(x_1,x_2)$ is the bounded interval - uniform distribution on the utilities x_1 and x_2 [see (5-5)], and Θ is the region over which this distribution is defined. The specific

Utility Measure

$U(\underline{r},\underline{d})$

(a) General intervals: $\alpha_1 \leq x_1 \leq \beta_1$ and $\alpha_2 \leq x_2 \leq \beta_2$ $[0 \leq \alpha_i \leq \beta_i \leq 1$ $(i=1,2)]$

$\underline{r}_1$	$\underline{U}_1(\underline{r})$	$\underline{U}_2(\underline{r})$
$[0,\alpha_2/(1-\alpha_1+\alpha_2))$	$[\alpha_1,\beta_1]$	$[\alpha_2,\beta_2]$
$(\alpha_2/(1-\alpha_1+\alpha_2),\beta_2/(1-\beta_1+\beta_2))$	$[\alpha_1,1]$	$[0,\beta_2]$
$(\beta_2/(1-\beta_1+\beta_2),1]$	1	0

(b) Specific intervals: $(\alpha_1,\beta_1) = (0.1,0.6)$ and $(\alpha_2,\beta_2) = (0.3,0.8)$

$\underline{r}_1$	$\underline{U}_1(\underline{r})$	$\underline{U}_2(\underline{r})$
[0,1/4)	[0.1,0.6]	[0.3,0.8]
(1/4,2/3)	[0.1,1]	[0,0.8]
(2/3,1]	1	0

Table 7-1. The bounds on, or values assumed by, the utility measure $U(\underline{r},\underline{d})$ in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2, when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of a bounded interval measure. (a) General intervals. (b) Specific intervals.

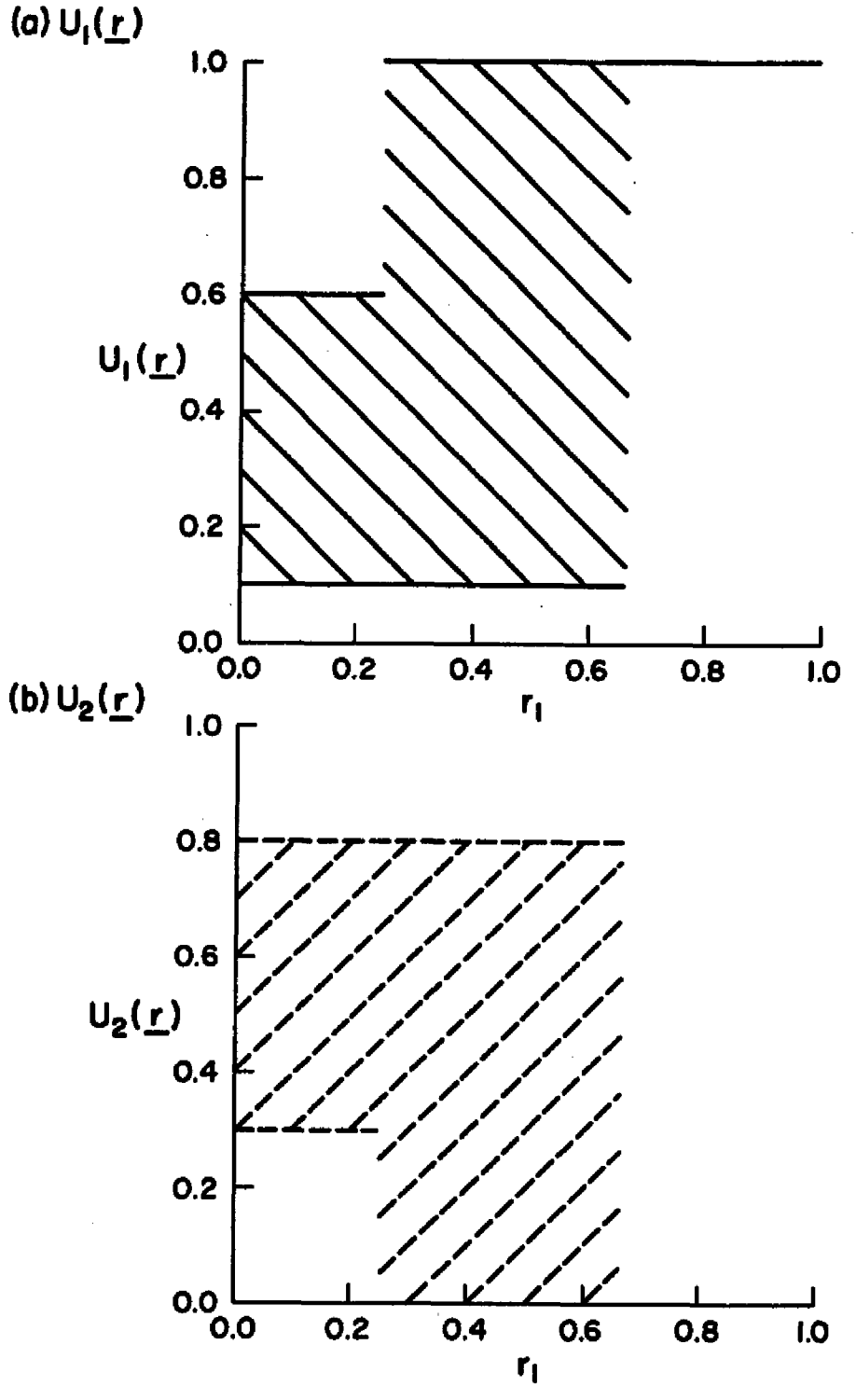


Figure 7-3. The utility measure $U(\underline{r},\underline{d})$ in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2, when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of the specific bounded interval measure $0.1 \leq x_1 \leq 0.6$ and $0.3 \leq x_2 \leq 0.8$ (see Table 7-1b). (a) $U_1(\underline{r})$. (b) $U_2(\underline{r})$.

expressions for $EU(\underline{r},\underline{d})$ when $\alpha_2/(1-\alpha_1+\alpha_2) \leq r_1 \leq \beta_2/(1-\beta_1+\beta_2)$ are rather lengthy and are reproduced in Appendix F. For the specific intervals described in Table 7-1b, the probability measure $f(x_1,x_2)$ is

$$f(x_1,x_2) = \begin{cases} 4 & 0.1 \leq x_1 \leq 0.6 \text{ and } 0.3 \leq x_2 \leq 0.8, \\ 0 & \text{otherwise,} \end{cases} \tag{7-6}$$

and the expected-utility measure $EU(\underline{r},\underline{d})$ in this situation is depicted in Figure 7-4.

(3) Beta measure. This measure defines (independent) beta distributions on the utilities x_1 and x_2. That is, $x_1 \sim \beta(a_1,b_1)$ and $x_2 \sim \beta(a_2,b_2)$. The expected-utility measure $EU(\underline{r},\underline{d})$ in this situation is, in general,

$$EU(\underline{r},\underline{d}) = \int\int_{(x_1,x_2)\in\Theta} [\delta_1 d_1 + \delta_2(d_1x_1 + d_2x_2)f(x_1,x_2)dx_1dx_2, \tag{7-7}$$

where $f(x_1,x_2) = f(x_1)f(x_2)$ and Θ is the unit square [the region over which the distribution $f(x_1,x_2)$ is defined]. The specific expressions for the measure $EU(\underline{r},\underline{d})$ in this situation, in which the $f(x_i)$ $(i=1,2)$ are beta distributions [see (5-7)], are also rather lengthy and are reproduced in Appendix G. We depict $EU(\underline{r},\underline{d})$ in Figures 7-5a, 7-5b, and 7-5c for specific values of the

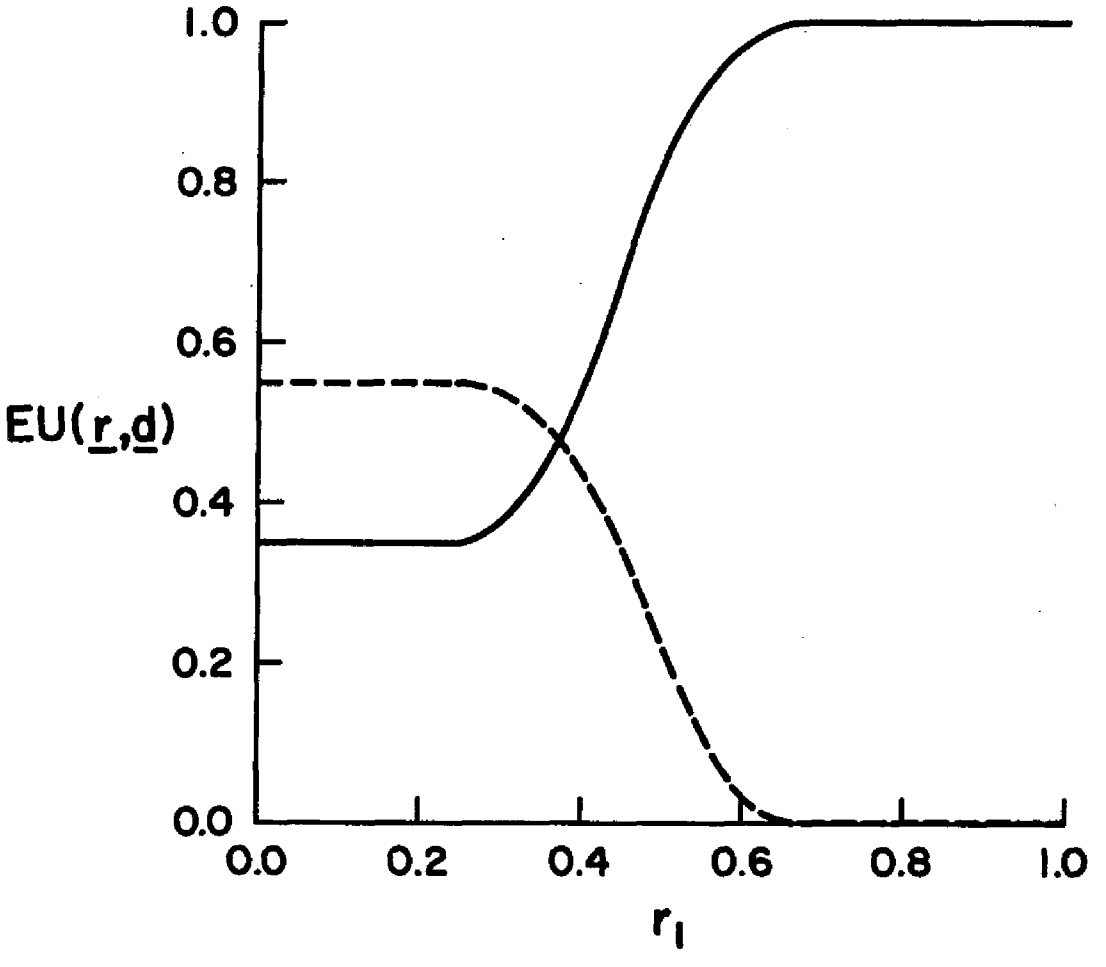


Figure 7-4. The expected-utility measure $EU(\underline{r},\underline{d})$ in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2, when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of the specific bounded interval - uniform measure $f(x_1,x_2) = 4$ for $0.1 \leq x_1 \leq 0.6$ and $0.3 \leq x_2 \leq 0.8$ and $f(x_1,x_2) = 0$ otherwise. The solid curve denotes $EU_1(\underline{r})$ and the dashed curve denotes $EU_2(\underline{r})$.

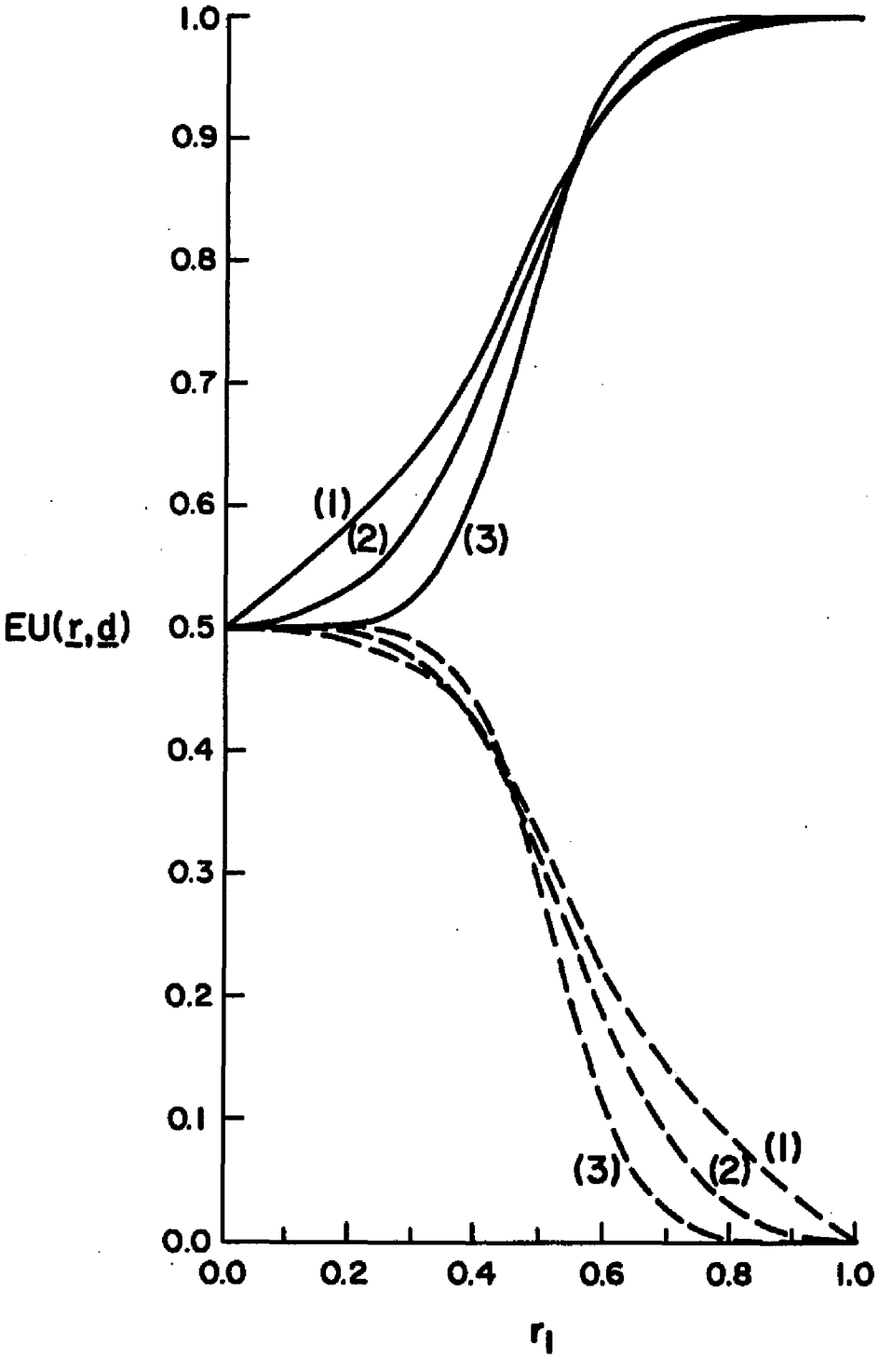


Figure 7-5(a). The expected-utility measure EU($\underline{r}$,$\underline{d}$) in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2, when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of the specific beta measures $x_1 \sim \beta(a_1,b_1)$ and $x_2 \sim \beta(a_2,b_2)$ for which $a_i = b_i$ $(i=1,2)$. (1) $a_i = b_i = 1$. (2) $a_i = b_i = 2$. (3) $a_i = b_i = 5$. The solid curves denote $EU_1(\underline{r})$ and the dashed curves denote $EU_2(\underline{r})$.

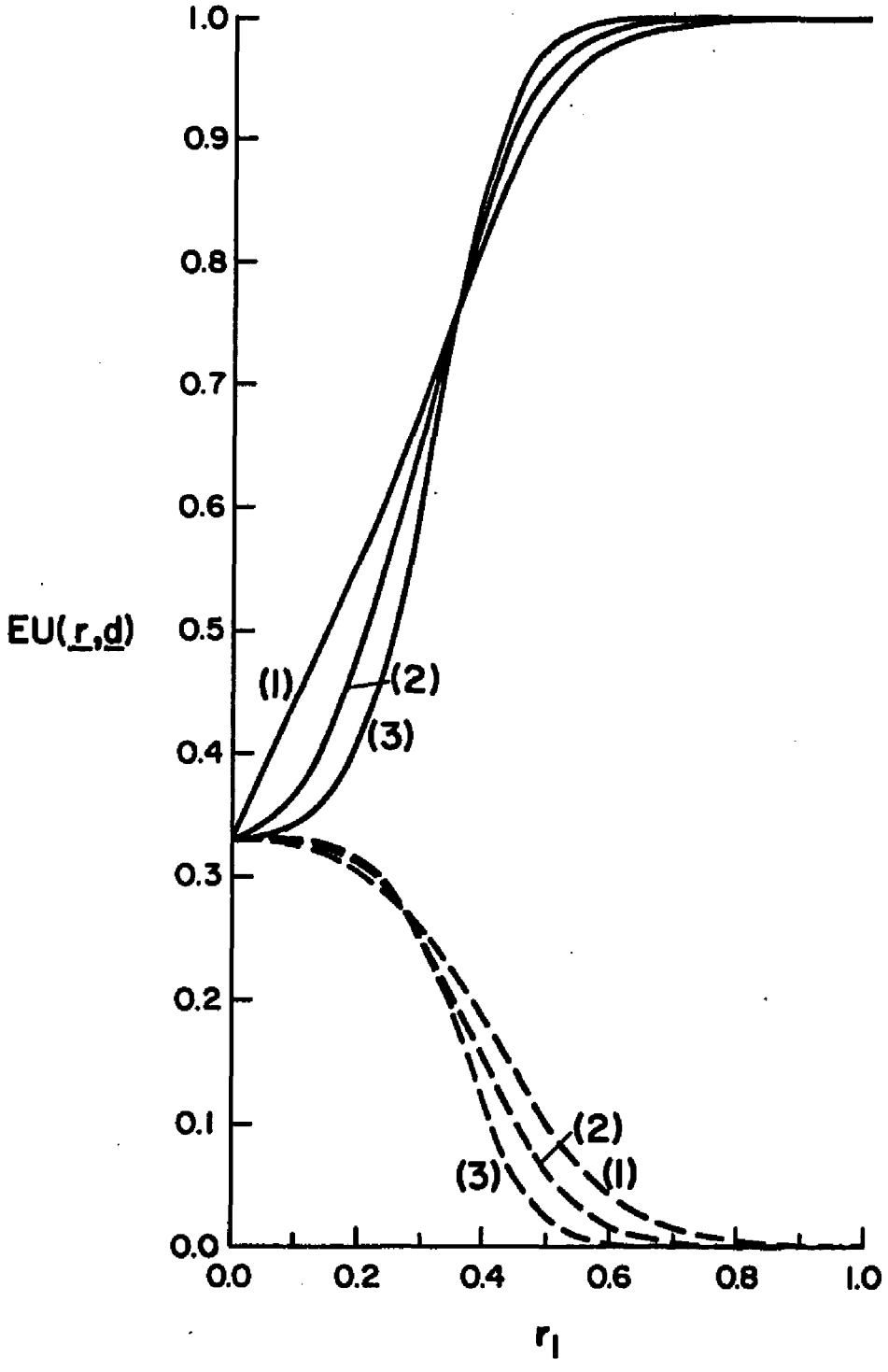


Figure 7-5(b). The expected-utility measure EU($\underline{r}$,$\underline{d}$) in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2, when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of the specific beta measures $x_1 \sim \beta(a_1,b_1)$ and $x_2 \sim \beta(a_2,b_2)$ for which $2a_i = b_i$ $(i=1,2)$. (1) $2a_i = b_i = 2$. (2) $2a_i = b_i = 4$. (3) $2a_i = b_i = 8$. The solid curves denote $EU_1(\underline{r})$ and the dashed curves denote $EU_2(\underline{r})$.

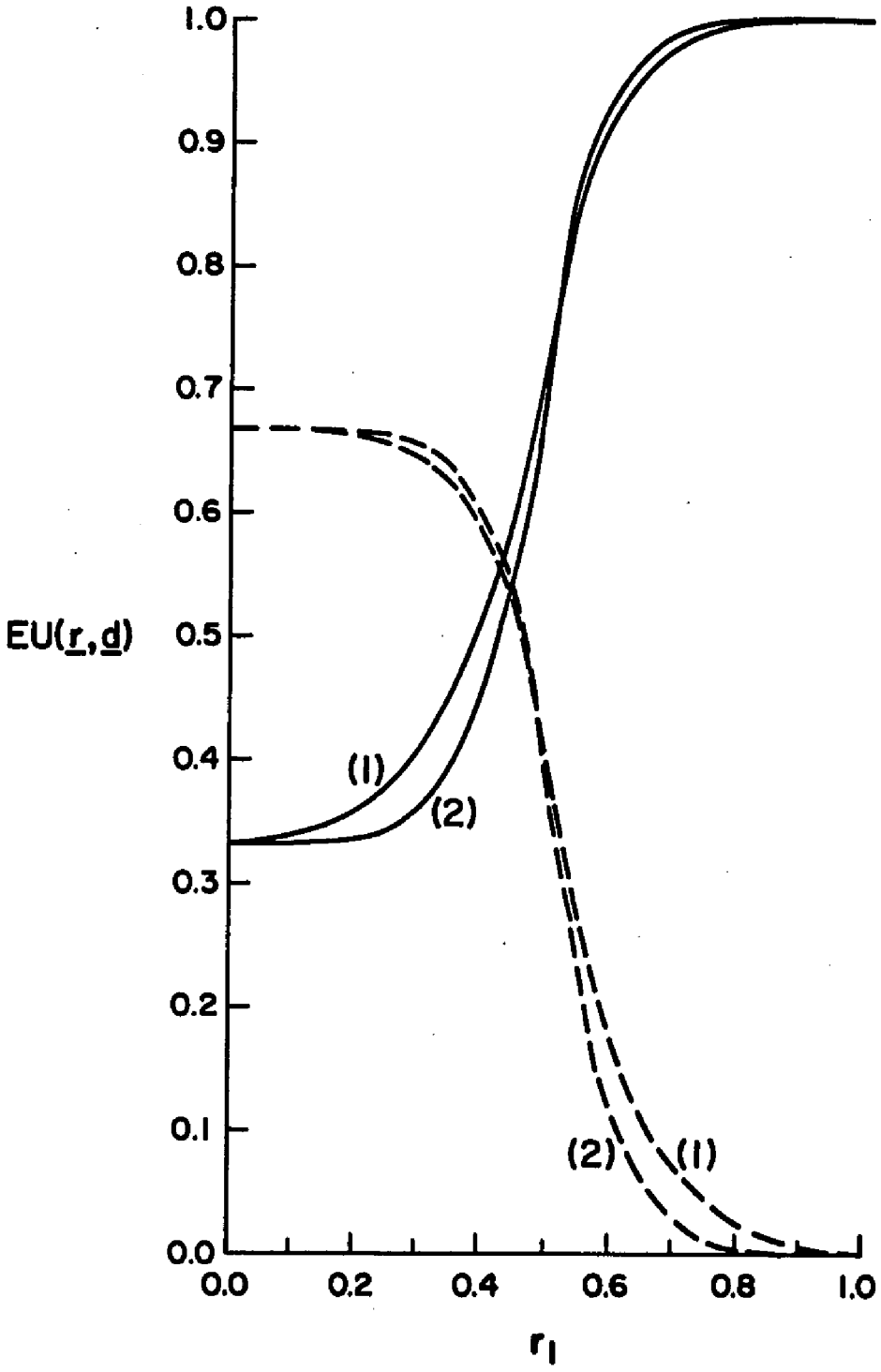


Figure 7-5(c). The expected-utility measure EU($\underline{r}$,$\underline{d}$) in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2, when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of the specific beta measures $x_1 \sim \beta(a_1,b_1)$ and $x_2 \sim \beta(a_2,b_2)$ for which $2a_1 = b_1$ and $a_2 = 2b_2$. (1) $2a_1 = b_1 = a_2 = 2b_2 = 2$. (2) $2a_1 = b_1 = a_2 = 2b_2 = 4$. The solid curves denote $EU_1(\underline{r})$ and the dashed curves denote $EU_2(\underline{r})$.

parameters a_i and b_i when (a) $a_1 = b_1$ [$E(x_1) = E(x_2) = 1/2$], (b) $2a_1 = b_i$ [$E(x_1) = E(x_2) = 1/3$], and (c) $2a_1 = b_1$ and $a_2 = 2b_2$ [$E(x_1) = 1/3$ and $E(x_2) = 2/3$] (i=1,2).

7.2.3.2 A Special Two-Action, Two-State Decision Situation

Consider the special two-action, two-state decision situation described by the key utility matrix (c) in Figure 5-3. Note that this situation is similar to the situation considered in 7.2.3.1 except that $x_1 = x_2 = x$. In this regard, the indifference point is $\underline{r}^* = (x, 1-x)$, and the utility measure $U(\underline{r},\underline{d})$ is

$$U(\underline{r},\underline{d}) = \delta_1 d_1 + \delta_2 x. \qquad (7\text{-}8)$$

With regard to the evaluator's knowledge of the utility x, we shall consider the same measures as those used in the general two-action, two-state situation.

(1) Bounded interval measure. The bounded interval measure in this situation specifies that $\alpha \leq x \leq \beta$. As in the general case, these bounds together with the expression for the indifference point lead to the identification of bounds on, or specific values for, the utility measure $U(\underline{r},\underline{d})$. These bounds and values are presented in Table 7-2, both for (a) the general interval and (b) a specific interval [namely, $(\alpha,\beta) = (0.1,0.6)$]. Note that this measure will bound, or specify values for, the measure $U(\underline{r},\underline{d})$ for at least some values of r_1 when $\alpha \neq 0$ or $\beta \neq 1$. The measure

Utility Measure

$U(\underline{r},\underline{d})$

(a) General interval: $\alpha \leq x \leq \beta \quad (0 \leq \alpha \leq \beta \leq 1)$

$\underline{r}_1$	$U_1(\underline{r})$	$U_2(\underline{r})$
$[0,\alpha)$	$[\alpha,\beta]$	$[\alpha,\beta]$
(α,β)	$[\alpha,1]$	$[0,\beta]$
$(\beta,1]$	1	0

(b) Specific interval: $(\alpha,\beta) = (0.1,0.6)$

$\underline{r}_1$	$U_1(\underline{r})$	$U_2(\underline{r})$
[0,0.1)	[0.1,0.6]	[0.1,0.6]
(0.1,0.6)	[0.1,1]	[0,0.6]
(0.6,1]	1	0

Table 7-2. The bounds on, or values assumed by, the utility measure $U(\underline{r},\underline{d})$ in the special two-action, two-state decision situation described by the key utility matrix (c) in Figure 5-3, when the evaluator's knowledge of the utility x is expressed in terms of a bounded interval measure. (a) General interval. (b) Specific interval.

$U(\underline{r},\underline{d})$ for the specific interval is depicted in Figure 7-6.

(2) Bounded interval - uniform measure. The bounded interval - uniform measure in this situation specifies that $f(x) = 1/(\beta - \alpha)$ for $\alpha \leq x \leq \beta$ and $f(x) = 0$ otherwise. The expected-utility measure $EU(\underline{r},\underline{d})$ in this situation is

$$EU(\underline{r},\underline{d}) = \begin{cases} d_1 & \text{if } r_1 \leq \alpha, \\ E(x) & \text{if } r_1 \geq \beta, \\ \int_{x_1 \epsilon \Theta} (\delta_1 d_1 + \delta_2 x) f(x)dx & \text{otherwise,} \end{cases} \tag{7-9}$$

where $E(x) = (1/2)(\alpha + \beta)$, $f(x)$ is a bounded interval - uniform distribution on the utility x, and Θ is the interval over which this distribution is defined. The specific expression for $EU(\underline{r},\underline{d})$ when $\alpha \leq r_1 \leq \beta$ is, then,

$$EU(\underline{r},\underline{d}) = [1/(\beta - \alpha)](d_1 \int_{\alpha}^{r_1} dx + \int_{r_1}^{\beta} xdx),$$

or

$$EU(\underline{r},\underline{d}) = [1/(\beta - \alpha)][d_1(r_1 - \alpha) + (1/2)(\beta^2 - r_1^2)]. \tag{7-10}$$

For the specific interval described in Table 7-2b, the probability measure $f(x)$ is

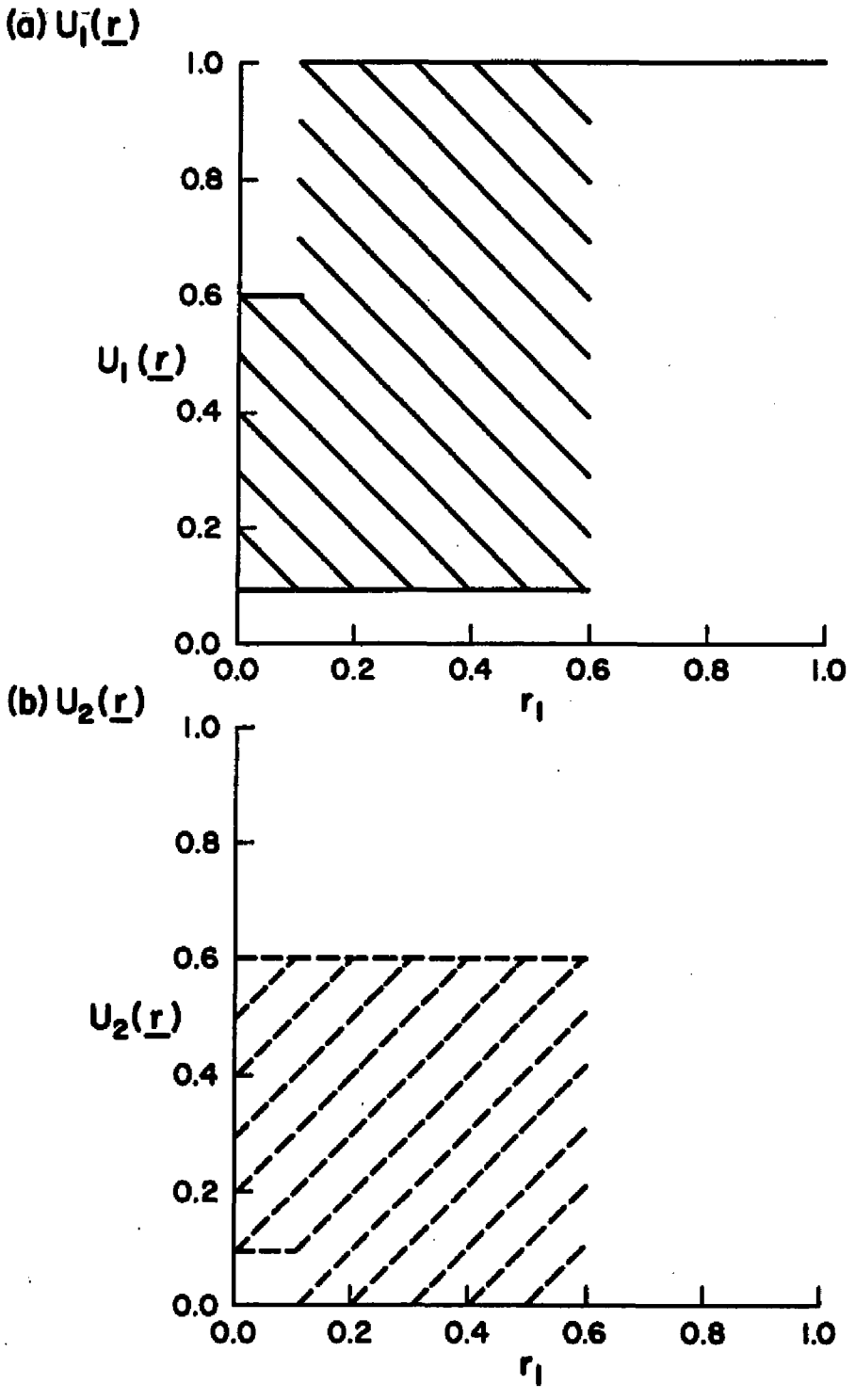


Figure 7-6. The utility measure $U(\underline{r},\underline{d})$ in the special two-action, two-state decision situation described by the key utility matrix (c) in Figure 5-3, when the evaluator's knowledge of the utility x is expressed in terms of the specific bounded interval measure $0.1 \leq x \leq 0.6$ (see Table 7-2b). (a) $U_1(\underline{r})$. (b) $U_2(\underline{r})$.

$$f(x) = \begin{cases} 2 & \text{if } 0.1 \leq x \leq 0.6, \\ 0 & \text{otherwise,} \end{cases} \tag{7-11}$$

and the measure $EU(\underline{r},\underline{d})$ in this situation is depicted in Figure 7-7.

(3) Beta measure. This measure specifies that the utility x has a beta distribution with parameters a and b. That is, $x \sim \beta(a,b)$. The expected utility measure $EU(\underline{r},\underline{d})$ in this situation is, in general,

$$EU(\underline{r},\underline{d}) = \int_{x\in\Theta} (\delta_1 d_1 + \delta_2 x) f(x)dx, \tag{7-12}$$

in which Θ, the interval over which the distribution f(x) is defined, is the unit interval [0,1]. The specific expression for the measure $EU(\underline{r},\underline{d})$ in this situation is, from (5-7),

$$EU(\underline{r},\underline{d}) = [(a+b-1)!/(a-1)!(b-1)!][d_1 \int_0^{r_1} x^{a-1}(1-x)^{b-1}dx + \int_{r_1}^{1} x^{a}(1-x)^{b-1}dx],$$

or, when a and b are positive integers (see Chapter 5),

$$EU(\underline{r},\underline{d}) = d_1 \sum_{i=a}^{a+b-1} \binom{a+b-1}{i} r_1^i (1-r_1)^{a+b-1-i} + [a/(a+b)][1 - \sum_{i=a+1}^{a+b} \binom{a+b}{i} r_1^i (1-r_1)^{a+b-i}. \tag{7-13}$$

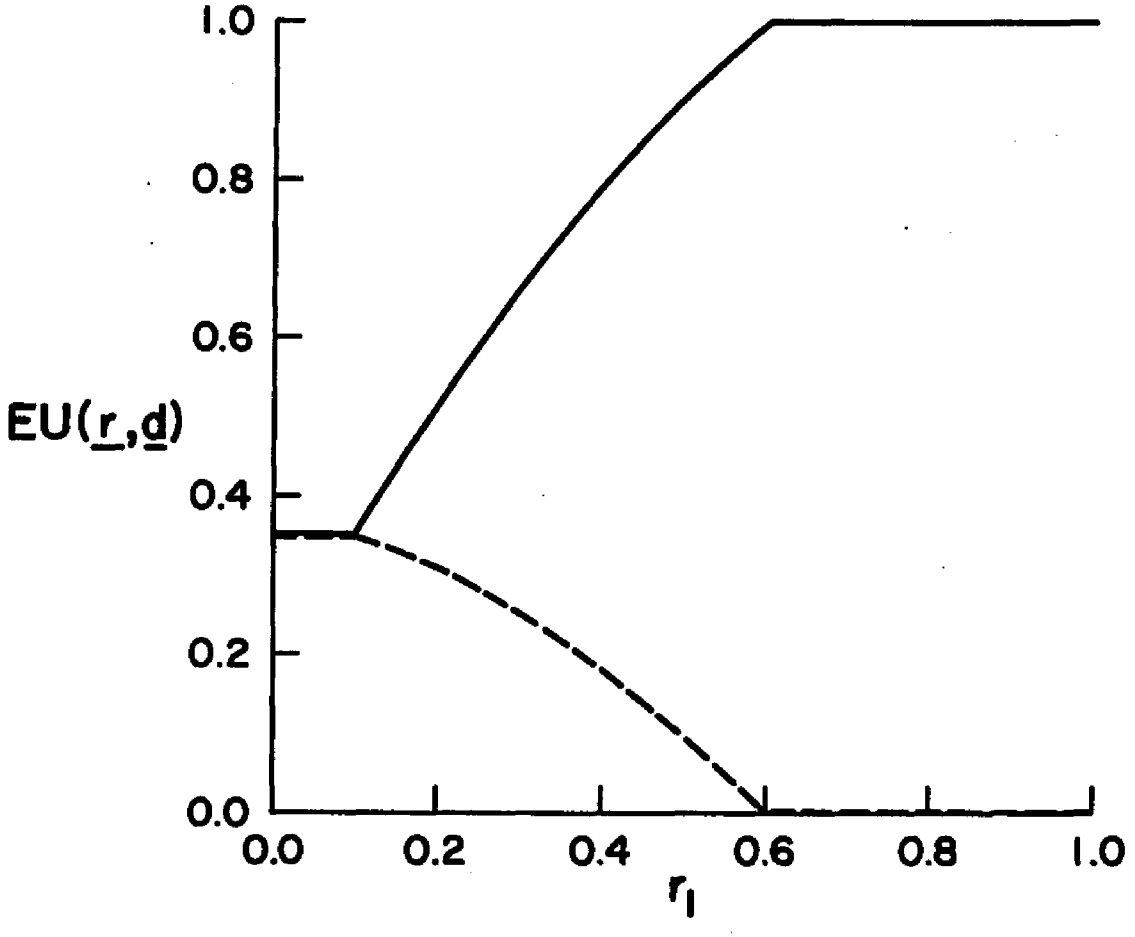


Figure 7-7. The expected-utility measure EU($\underline{r}$,$\underline{d}$) in the special two-action, two-state decision situation described by the key utility matrix (c) in Figure 5-3, when the evaluator's knowledge of the utility x is expressed in terms of the specific bounded interval - uniform measure $f(x) = 2$ for $0.1 \leq x \leq 0.6$ and $f(x) = 0$ otherwise. The solid curve denotes $EU_1(\underline{r})$ and the dashed curve denotes $EU_2(\underline{r})$.

The measure $EU(\underline{r},\underline{d})$ is depicted in Figure 7-8 for selected values of a and b when $E(x)$ $[=a/(a+b)]$ equals (a) 1/2 and (b) 1/3.

7.2.3.3 A Special Three-Action, Three-State Decision Situation

As a final example, we consider the special three-action, three-state decision situation presented in Figure 7-9. The framework for operational evaluation in this situation is depicted within the simplex R^3 in Figure 7-10. The utility measure $U(\underline{r},\underline{d})$ is

$$U(\underline{r},\underline{d}) = \delta_1(d_1 + d_2x) + \delta_2(d_1x + d_2 + d_3x) + \delta_3(d_2x + d_3). \quad (7\text{-}14)$$

The expected-utility measure $EU(\underline{r},\underline{d})$ in this situation is, in general,

$$EU(\underline{r},\underline{d}) = \int_{x\in\Theta} [\delta_1(d_1 + d_2x) + \delta_2(d_1x + d_2 + d_3x) + \delta_3(d_2x + d_3)]f(x)dx, \quad (7\text{-}15)$$

where Θ is the interval over which the probability distribution $f(x)$ is defined. In this example, we shall assume that the evaluator's knowledge of the utility x is expressed in terms of a uniform distribution. That is, $x \sim \beta(1,1)$. The specific expressions for the measure $EU(\underline{r},\underline{d})$ in this case are presented in Table 7-3. The values of the measures $EU_1(\underline{r})$ and $EU_2(\underline{r})$ are depicted in Figures 7-11a and 7-11b, respectively, within the framework of the simplex

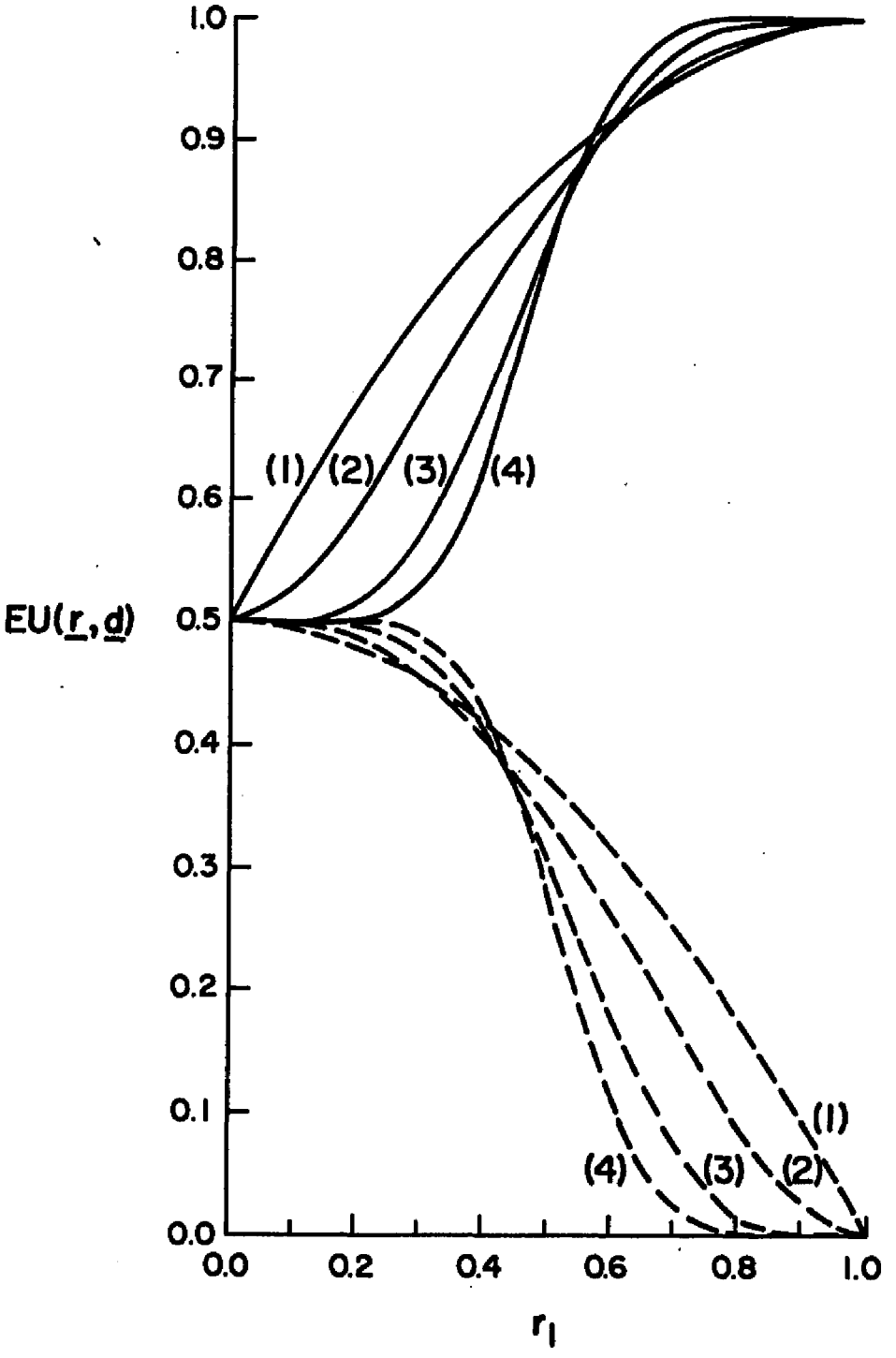


Figure 7-8(a). The expected-utility measure EU(r,d) in the special two-action, two-state decision situation described by the key utility matrix (c) in Figure 5-3, when the evaluator's knowledge of the utility x is expressed in terms of the specific beta measures $x \sim \beta(a,b)$ for which $a = b$. (1) $a = b = 1$. (2) $a = b = 2$. (3) $a = b = 5$. (4) $a = b = 10$. The solid curves denote $EU_1(r)$ and the dashed curves denote $EU_2(r)$.

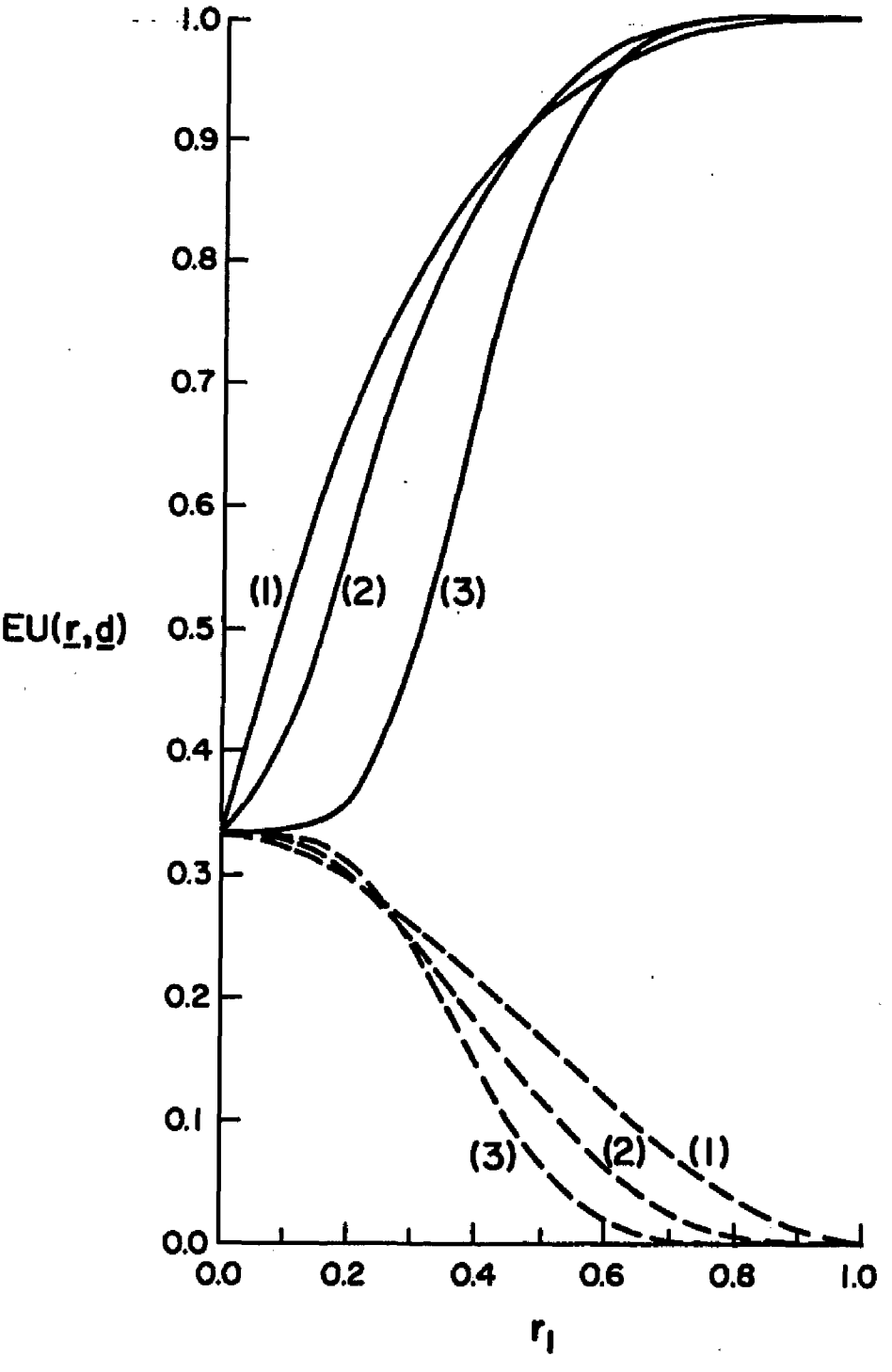
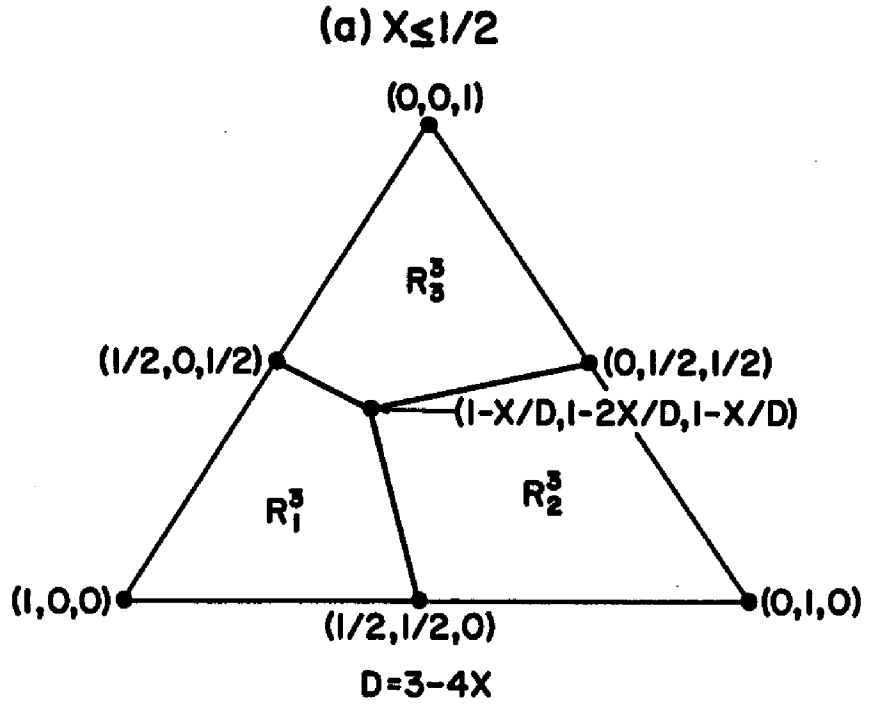


Figure 7-8(b). The expected-utility measure EU(r,d) in the special two-action, two-state decision situation described by the key utility matrix (c) in Figure 5-3, when the evaluator's knowledge of the utility x is expressed in terms of the specific beta measures $x \sim \beta(a,b)$ for which $2a = b$. (1) $2a = b = 2$. (2) $2a = b = 4$. (3) $2a = b = 8$. The solid curves denote $EU_1(r)$ and the dashed curves denote $EU_2(r)$.

Utility Matrix $\underline{U}$

		States			
		s_1	s_2	s_3	Expected Utilities
	a_1	1	x	0	$U_1 = r_1 + r_2 x$
Actions	a_2	x	1	x	$U_2 = r_1 x + r_2 + r_3 x$
	a_3	0	x	1	$U_3 = r_2 x + r_3$
Probability Vector	$\underline{r}$	r_1	r_2	r_3	

Figure 7-9. A special three-action, three-state decision situation.



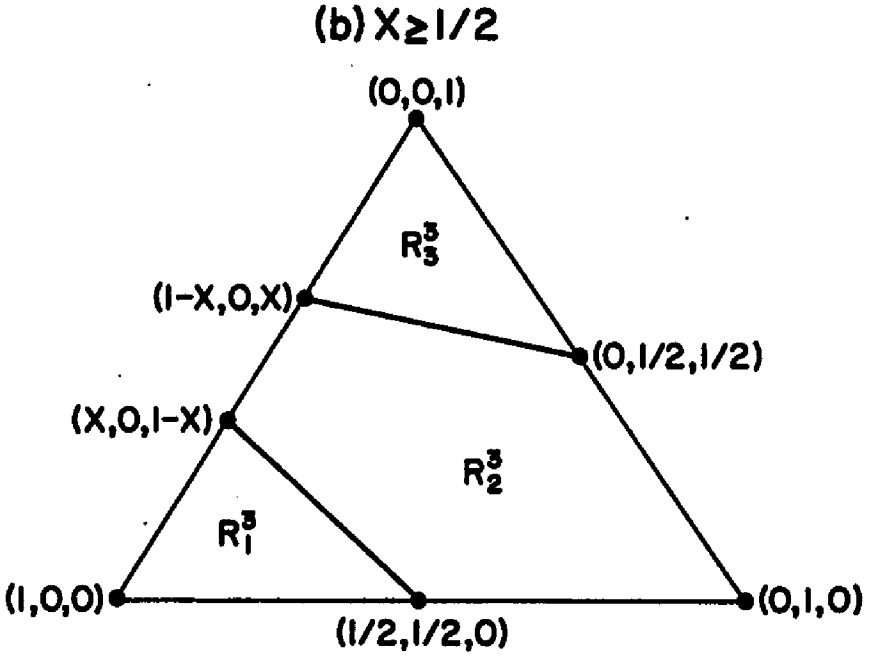


Figure 7-10. The framework for operational evaluation for the special three-action, three-state decision situation described in Figure 7-9 when (a) $0 \leq x \leq 1/2$ and (b) $1/2 \leq x \leq 1$.

Conditions on Probabilities	$EU_1(\underline{r})$	$EU_2(\underline{r})$	$EU_3(\underline{r})$
$r_1 \geq r_2 \geq r_3$ or $r_1 \geq r_3 \geq r_2$	$1-(1/2)(r_3/D)^2$	$(1/2)+(1/2)(r_3/D)^2$	$(r_3/D)-(1/2)(r_3/D)^2$
$r_2 \geq r_1 \geq r_3$ or $r_2 \geq r_3 \geq r_1$	$1/2$	1	$1/2$
$r_3 \geq r_1 \geq r_2$ or $r_3 \geq r_2 \geq r_1$	$(r_1/D)-(1/2)(r_1/D)^2$	$(1/2)+(1/2)(r_1/D)^2$	$1-(1/2)(r_1/D)^2$

$$D = 1-2r_2$$

Table 7-3. The expressions for the expected-utility measure $EU(\underline{r},\underline{d})$ in the special three-action, three-state decision situation described in Figure 7-9, when the evaluator's knowledge of the utility x is expressed in terms of a uniform distribution $x \sim \beta(1,1)$.

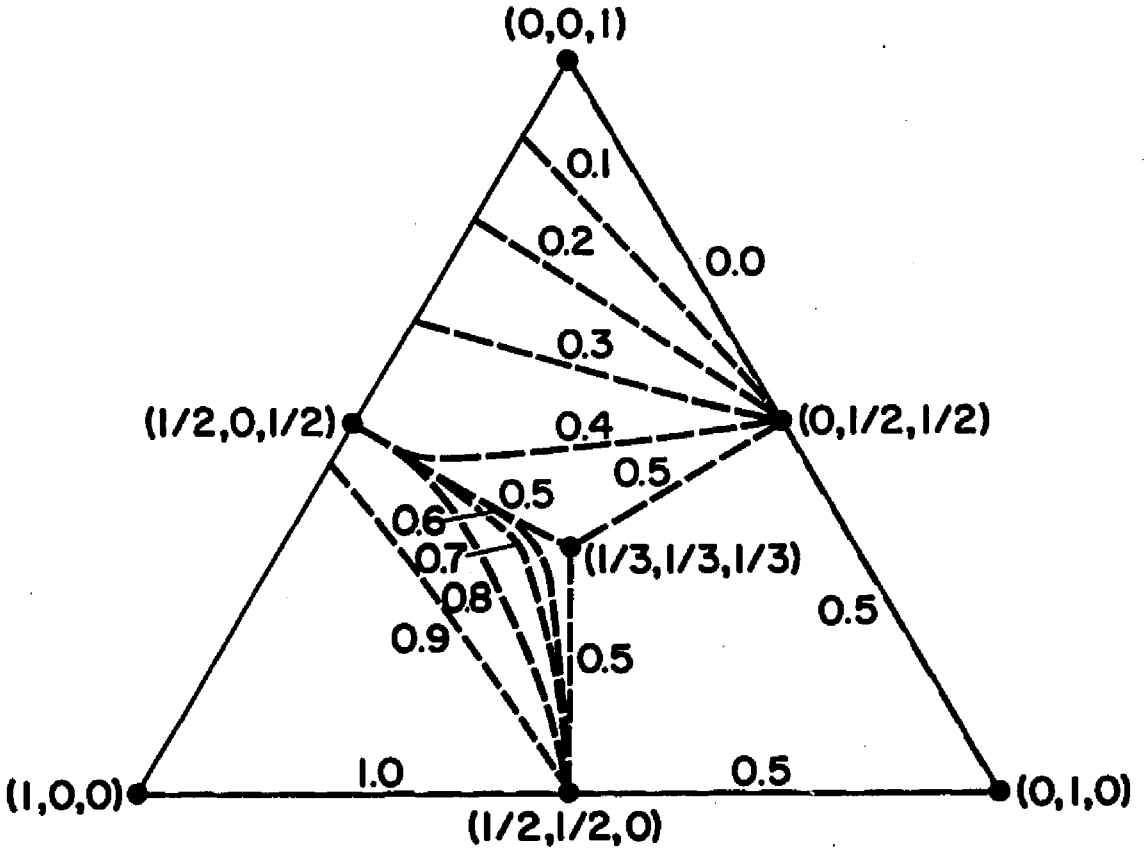


Figure 7-11(a). The expected-utility measure $EU_1(\underline{r})$ in the special three-action, three-state decision situation described in Figure 7-9, when the evaluator's knowledge of the utility x is expressed in terms of a uniform probability distribution $x \sim \beta(1,1)$. $EU_1(\underline{r})$ becomes $EU_3(\underline{r})$ when the labels on the vertices $(1,0,0)$ and $(0,0,1)$ are interchanged.

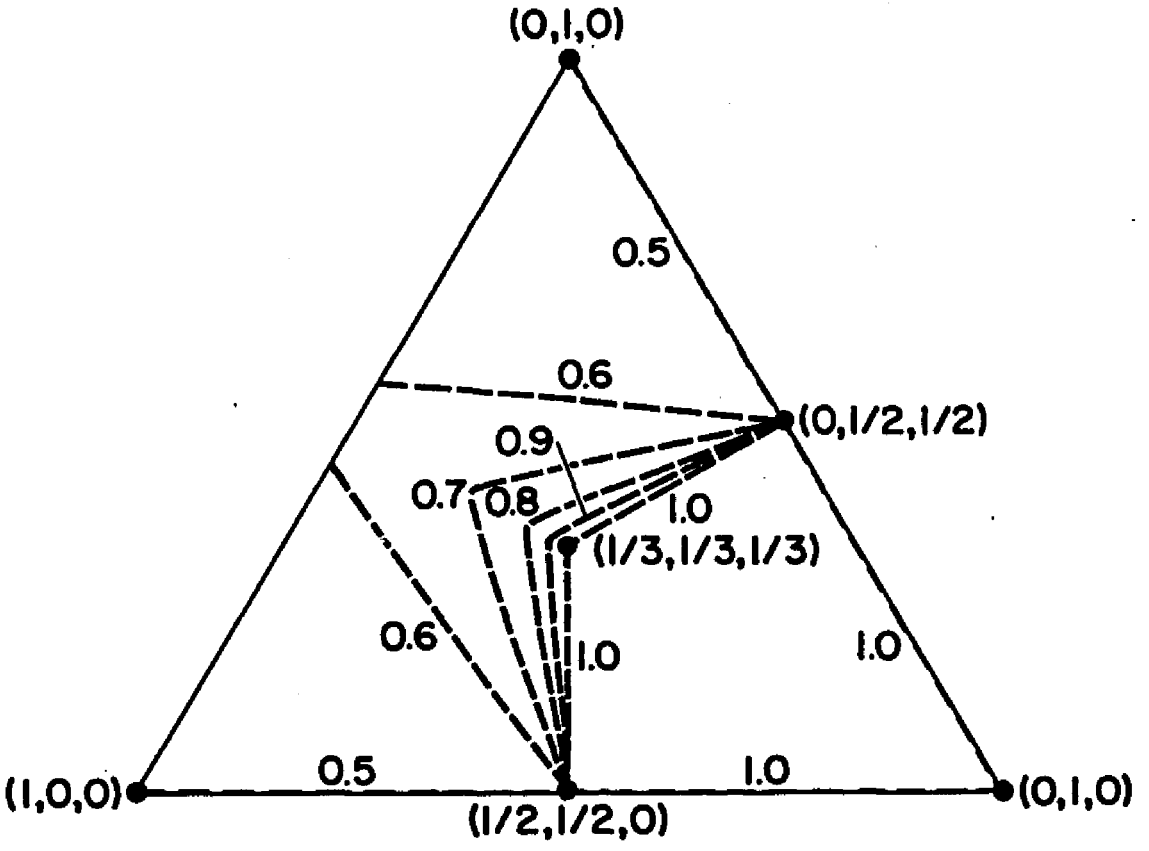


Figure 7-11(b). The expected-utility measure $EU_2(\underline{r})$ in the special three-action, three-state decision situation described in Figure 7-9, when the evaluator's knowledge of the utility x is expressed in terms of a uniform probability distribution $x \sim \beta(1,1)$.

R^3. Since the utility matrix $\underline{U}$ is symmetric in this situation (see Figure 7-9), the measure $EU_3(\underline{r})$ can be obtained simply by interchanging the labels on the vertices (1,0,0) and (0,0,1) in Figure 7-11a. The form of the isograms of expected utility (i.e., the "lines" of equal expected utility) in Figures 7-11a and 7-11b is indicated in Figure 7-12. Note that these isograms are curves in one region and lines in two other regions and that the measure $EU(\underline{r},\underline{d})$ assumes a constant value in the remaining region.

7.3 Some Properties of Utility and Expected-Utility Measures

In this section we describe some of the properties of utility and expected-utility measures in general and specific decision situations. These "results" are summarized in terms of various statements, conjectures, and speculations. A formal proof of each statement is included whenever possible. Conjectures and speculations are supported by brief heuristic arguments.

7.3.1 Utility Measures

The utility measure $U(\underline{r},\underline{d})$ in (7-1) is a proper scoring rule, and a general proof of this statement is contained in Appendix D. Thus, this measure does not offer the meteorologist any incentive to make his forecast $\underline{r}$ differ from his judgment $\underline{p}$.

The measure $U(\underline{r},\underline{d})$ is depicted in Figures 7-1 and 7-2 for two decision situations in which the evaluator's knowledge of the utilities is assumed to be complete. In the three-action, two-state

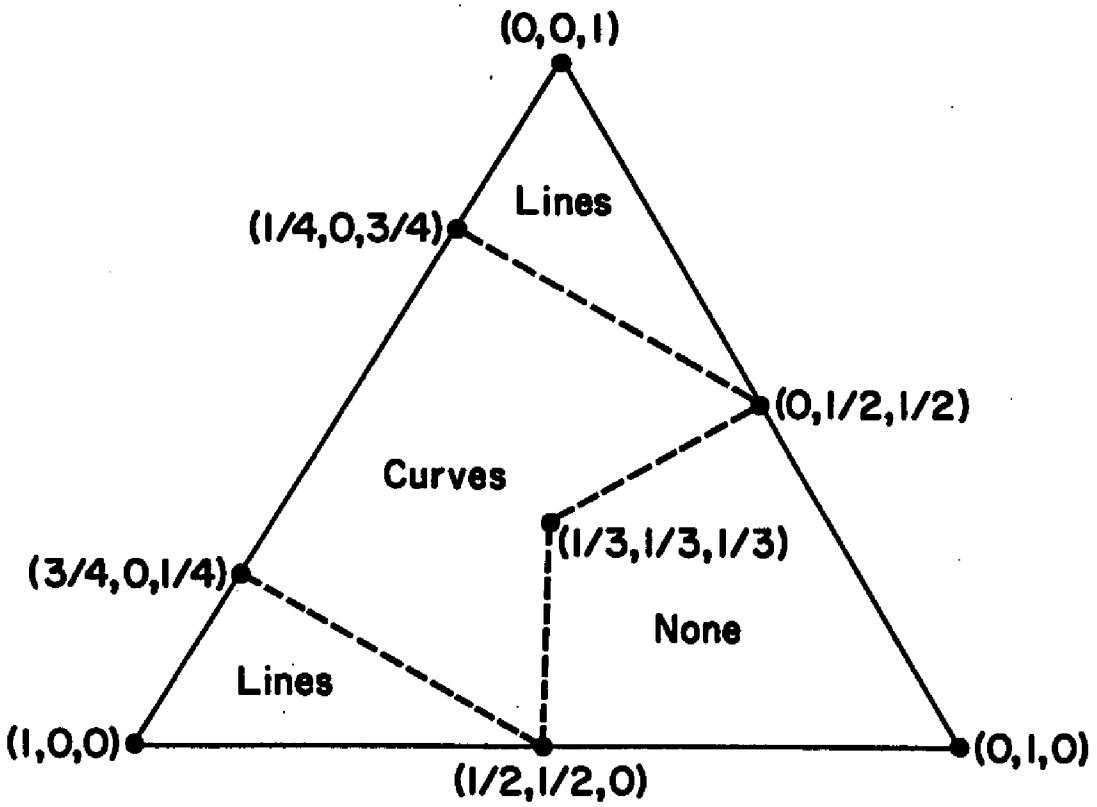


Figure 7-12. The form of the isograms of expected utility (i.e., the "lines" of equal expected utility) within the framework of the simplex R^3 for the three-action, three-state decision situation described in Figure 7-9. The term "none" means that all of the points in the region are of equal expected utility.

situation considered in Figure 7-1, the simplex R^2 is divided into three intervals R_m^2 (m=1,2,3) and the measure $U(\underline{r},\underline{d})$ is a step function with steps at the indifference points $r_1 = 1/3$ and $r_1 = 1/2$. Note that $U_1(\underline{r})$ is an increasing function of r_1 and that $U_2(\underline{r})$ is an increasing function of r_2 $(=1-r_1)$. That is, $U_n(\underline{r})$ either remains constant or increases as r_n increases (n=1,2). In the four-action, three-state situation considered in Figure 7-2, R^3 is divided into four convex regions R_m^3 (m=1,2,3,4), and $U(\underline{r},\underline{d})$ is a plateau function with "steps" along the indifference lines separating the respective regions. A careful examination of Figure 7-2 will reveal that $U_n(\underline{r})$ is an increasing function of r_n along directed line segments (see paragraph 6.3.1.1) through the vertex $\underline{d}_n$ (n=1,2,3). That is, the value of $U_n(\underline{r})$ either remains constant or increases as r_n increases along directed line segments through the vertex $\underline{d}_n$ (n=1,2,3). We extend this result to the general M-action, N-state situation by means of CONJECTURE 7-1.

CONJECTURE 7-1

The general utility measure $U_n(\underline{r})$ in M-action, N-state decision situations is an increasing function of r_n along directed line segments through the vertex $\underline{d}_n$ (n=1,...,N).

While we shall not attempt to prove this conjecture, we suggest that this result follows from the fact that the regions R_m^N (m=1,...,M)

are convex.

The measure $U(\underline{r},\underline{d})$ is also the appropriate measure of "value" in situations in which the evaluator's knowledge is expressed in terms of inequality measures (see Section 7.2). In this regard, the utility measures $U_n(\underline{r})$ (n=1,2) are depicted in Figures 7-3 and 7-6 for general and special two-action, two-state decision situations, respectively, in which the evaluator's knowledge is expressed in terms of specific bounded interval measures. CONJECTURE 7-1 indicates that these (utility) measures should be increasing functions of r_n (n=1,2), and an examination of Figures 7-3 and 7-6 provides additional support for this statement. Of course, since this measure of knowledge only bounds the values of the relevant utilities, the measure $U(\underline{r},\underline{d})$ can assume a range of values for certain values of r_1.

7.3.2 Expected-Utility Measures

The expected-utility measure $EU(\underline{r},\underline{d})$ in (7-3) is (1) a <u>proper</u> scoring rule if the probability distribution $f(\underline{U})$ is a non-negative function [i.e., if $f(\underline{U})\geq 0$] and (2) a <u>strictly proper</u> scoring rule if this distribution is a positive function [i.e., $f(\underline{U})>0$]. In Appendix E we show that the expected-utility measure associated with a specific decision situation is a proper [under (1)] or strictly proper [under (2)] scoring rule, and we offer a heuristic argument (together with some other evidence) to support the statement that the expected-utility measure $EU(\underline{r},\underline{d})$ in M-action, N-state decision situations also

possesses these properties. Thus, the measure $EU(\underline{r},\underline{d})$ encourages the meteorologist to make his forecast $\underline{r}$ correspond (exactly) to his judgment $\underline{p}$ when $f(\underline{U})$ is positive and offers the meteorologist no incentive to make his forecast $\underline{r}$ differ from his judgment $\underline{p}$ when $f(\underline{U})$ is non-negative.

The expected-utility measure $EU(\underline{r},\underline{d})$ is depicted in Figures 7-4 and 7-7 for the general and special two-action, two-state situations, respectively, in which the evaluator's knowledge is expressed in terms of specific bounded interval - uniform measures. In both situations the measure $EU(\underline{r},\underline{d})$ is a strictly increasing function of r_n for those values of r_n which correspond to the region (general situation) or interval (special situation) over which the relevant utilities possess a uniform distribution (n=1,2). That is, $U_n(\underline{r})$ increases as r_n increases (n=1,2). For the remaining values of r_n, $EU_n(\underline{r})$ is a constant. Thus, over the entire range of values of r_n $(0 \leq r_n \leq 1)$, $EU_n(\underline{r})$ is an increasing function of r_n (n=1,2).

The measure $EU(\underline{r},\underline{d})$ is depicted in Figures 7-5 and 7-8 for these same general and special situations, respectively, when the evaluator's knowledge is expressed in terms of specific (independent) beta measures. In both situations, and for all of the specific measures of knowledge, the measure $EU_n(\underline{r})$ is a strictly increasing function of r_n (n=1,2).

In Figure 7-11 we depict the measure $EU(\underline{r},\underline{d})$ for a special three-action, three-state situation (see Figure 7-9) in which the evaluator's

knowledge is expressed in terms of a uniform distribution. In this situation $EU_n(\underline{r})$ is a strictly increasing function of r_n along directed line segments through the vertex $\underline{d}_n$ (n=1,2,3; n=1 and n=3 in Figure 7-11a and n=2 in Figure 7-11b). We extend the results described in this paragraph and in the previous two paragraphs to the general M-action, N-state situation by means of CONJECTURE 7-2.

CONJECTURE 7-2

The general expected-utility measure $EU_n(\underline{r})$ in M-action, N-state decision situations is (1) an increasing function of r_n along directed line segments through the vertex $\underline{d}_n$ when the evaluator's knowledge is expressed in terms of a non-negative probability distribution and (2) a strictly increasing function of r_n along directed line segments through the vertex $\underline{d}_n$ when the evaluator's knowledge is expressed in terms of a positive probability distribution (n=1,...,N).

While we shall not present a proof of this conjecture, we suggest that this result also follows from the fact that the regions R_m^N (m=1,...,M) are convex.

Consider the expected-utility measures depicted in Figure 7-8 for the special two-action, two-state situation in which the evaluator's knowledge of the utility x is expressed in terms of a beta distribution [i.e., $x \sim \beta(a,b)$]. Specific measures are presented which correspond to selected pairs of values of the parameters a and

b. In particular, c = a = b [=E(x)=1/2] in Figure 7-8a and c = 2a = b [=E(x)=1/3] in Figure 7-8b. As c increases the expected-utility measures in Figures 7-8a and 7-8b approach step functions with (single) steps at $r_1 = 1/2$ and $r_1 = 1/3$, respectively. The expected-utility measures in the general two-action, two-state situation exhibit a similar behavior (see Figure 7-5). These results are, of course, consistent with the fact that, as the evaluator's knowledge increases (i.e., as c increases), the expected-utility measures should approach utility measures.

7.3.3 Utility and Expected-Utility Measures: A Summary

The general utility and expected-utility measures formulated in Sections 7.1 and 7.2 have been shown to be proper and strictly proper (or proper) scoring rules, respectively. Thus, these measures encourage meteorologists to make their forecasts correspond to their judgments (or at least offer the forecaster no incentive to hedge). We have also shown that the utility measure $U_n(\underline{r})$ and the expected-utility measure $EU_n(\underline{r})$ are increasing and strictly increasing (or increasing) functions, respectively, of r_n along directed line segments through the vertex $\underline{d}_n$ of the simplex R^N (n=1,...,N). This result will be of particular importance when we consider the relationships between empirical and operational evaluation measures in Chapter 8. Finally, we have shown that expected-utility measures approach utility measures as the evaluator's knowledge of the relevant utilities increases. This result indicates that the proposed framework for

operational evaluation is internally consistent.

To complete this brief summary of operational evaluation, we describe the form and properties of the utility and expected-utility measures in Table 7-4.

Name	Form	Range	Properties: Orientation[1]	Properties: SP/P/I[2]
Utility Measure	$U(\underline{r},\underline{d}) = \sum_{m=1}^{M} \delta_m \left(\sum_{n=1}^{N} d_n u_{mn} \right)$	[0,1]	P	P
Expected Utility Measure	$EU(\underline{r},\underline{d}) = \int_{\Theta} \sum_{m=1}^{M} \delta_m \left(\sum_{n=1}^{N} d_n u_{mn} \right) f(\underline{U}) d\underline{U}$[3]	[0,1]	P	SP/P[4]

Key

[1] P = Positive, N = Negative

[2] SP = Strictly Proper, P = Proper, I = Improper

[3] If the probability distribution representing the evaluator's knowledge of the utilities $\underline{U} = (u_{mn})$ is discrete, the integral is replaced by a summation and $f(\underline{U})d\underline{U}$ is replaced by $p(\underline{U})$ (see subsection 7.2.2).

[4] $EU(\underline{r},\underline{d})$ is strictly proper if $f(\underline{U})>0$ and proper if $f(\underline{U})\geq 0$.

Table 7-4. The name, form, and properties of the operational evaluation measures considered in this study (see subsections 7.1.1 and 7.2.2).

CHAPTER EIGHT. EMPIRICAL AND OPERATIONAL EVALUATION MEASURES: SOME RELATIONSHIPS

In this chapter we are concerned with the relationships between the basic measures for empirical and operational evaluation described in Chapters 6 and 7. However, relationships between measures of attributes such as validity and utility on these scales will, of course, differ from measure to measure as well as from decision situation to decision situation. On the other hand, in comparative evaluation in which the evaluator is concerned with the relative validity and relative utility of forecasts prepared by two (or more) forecasters or forecasting procedures, even the existence of relationships on an ordinal scale between measures of these attributes would be of considerable importance. Furthermore, if the evaluator is a meteorologist (as opposed to a decision maker), relationships between the relative validity and relative utility of such forecasts would be of particular interest in view of the difficulties inherent in any attempt on the part of the forecaster to determine the absolute or relative utility of his forecasts directly.

We shall consider both general relationships between measures of validity and measures of utility on an ordinal scale and specific relationships between such measures in special decision situations. With regard to general relationships, the problem of concern can be described as follows: If two forecasters, A and B say, prepare forecasts on the same occasions and if A's forecasts are more valid than

B's forecasts, then what if any relationships exist on an ordinal scale between the utility of A's and B's forecasts and what if any assumptions must be satisfied for these relationships to hold? We consider this question in Section 8.1, first for individual forecasts and then for collections of forecasts.

The relationships between specific validity and utility measures in special decision situations are examined in Section 8.2. We first consider a special two-action, two-state decision situation and then we examine a M-action, M-state generalization of this situation.

8.1 General Decision Situations

In this section we investigate the ordinal relationships between measures of (primary) validity and measures of utility and expected utility for individual forecasts (subsection 8.1.1) and collections of forecasts (subsection 8.1.2). We then illustrate these relationships using specific examples in subsection 8.1.3. In subsection 8.1.4 we offer some speculations regarding the existence of ordinal relationships between measures of validity and measures of utility in general situations.

We shall denote the (primary) validity of A's forecasts, as measured by $D(\underline{r},\underline{d})$, by VA and the utility and expected utility of A's forecasts, as measured by $U(\underline{r},\underline{d})$ and $EU(\underline{r},\underline{d})$, by UA and EUA, respectively. Similarly, we shall denote the validity and utility and

expected utility of B's forecasts by VB and UB and EUB, respectively. Finally, we shall refer to two forecasts as <u>comparable</u> if they correspond to points on the same directed line segment through the vertex $\underline{d}_n$ of the simplex R^N (n=1,...,N).

8.1.1 Individual Forecasts

The validity measure $D_n(\underline{r})$ is a <u>strictly increasing</u> function of r_n along directed line segments through the vertex $\underline{d}_n$ (n=1,...,N) (see Chapter 6). That is, if A's and B's forecasts are comparable and if $r_n^A > r_n^B$, then VA > VB.[1] The utility measure $U_n(\underline{r})$, on the other hand, is an <u>increasing</u> function of r_n along directed line segments through the vertex $\underline{d}_n$ (see Chapter 7). That is, if A's and B's forecasts are comparable and if $r_n^A > r_n^B$, then UA $\geq$ UB. The expected-utility measure $EU(\underline{r},\underline{d})$ for Class IIa (i.e., non-negative probability) measures is also an increasing function of r_n for comparable forecasts (see Chapter 7). However, for Class IIb (i.e., positive probability) measures, the measure $EU(\underline{r},\underline{d})$ is a <u>strictly increasing</u> function of r_n for comparable forecasts (see Chapter 7). That is, if A's and B's forecasts are comparable and if $r_n^A > r_n^B$

---

[1] We shall consider only the situation in which $r_n^A > r_n^B$. If $r_n^A < r_n^B$, then we can relabel the forecasters. We shall also not consider forecasts for which $r_n^A = r_n^B$.

then EUA > EUB. The relationships, on an ordinal scale, between the measures of concern are summarized in RESULT 8-1.

RESULT 8-1 (N States, One Forecast)

Assumption. A's and B's forecasts are comparable.

Results. (1) Class I: If VA > VB, then UA $\geq$ UB.

(2a) Class IIa: If VA > VB, then EUA $\geq$ EUB.

(2b) Class IIb: If VA > VB, then EUA > EUB.

Thus, for individual, comparable forecasts, the (primary) validity measure D($\underline{r}$,$\underline{d}$) represents an ordinal measure on utility and expected utility, i.e., the measure D($\underline{r}$,$\underline{d}$) is able to order, or rank, forecasters according to the utility or expected utility of their forecasts.

Unfortunately, A's and B's forecasts will not often be comparable in N-state (N>2) situations. However, on many occasions the forecasts may be "nearly" comparable, i.e., the forecasts may correspond to points on two line segments which are "close" to one another. On such occasions the measure D($\underline{r}$,$\underline{d}$) represents an ordinal measure on utility and expected utility for most, but not all, decision situations and states of the evaluator's knowledge. In two-state (N=2) situations, however, all line segments are directed line segments (see Chapter 6). Thus, all forecasts are comparable in two-state situations. Therefore, the ordinal relationships

described in RESULT 8-1 hold, in general, in two-state situations.

Since, as indicated in Chapter 6, all strictly proper scoring rules are strictly increasing functions of r_n along directed line segments through the vertex $\underline{d}_n$, RESULT 8-1 (and the other results to be presented in this chapter) holds for all such scoring rules. Of course, if the scoring rule of concern is a strictly decreasing function of r_n, then certain inequalities in RESULT 8-1 must be reversed.

8.1.2 Collections of Forecasts

For collections of K (K>1) forecasts, we shall be concerned with the relationships, on an ordinal scale, between the average "accuracy" and average "value" of the forecasts.

Although the validity $D(\underline{r},\underline{d})$ represents an ordinal measure on the utility $U(\underline{r},\underline{d})$ and expected utility $EU(\underline{r},\underline{d})$ for individual, comparable forecasts, the average validity $\overline{D}(\underline{r},\underline{d})$ is not, in general, an ordinal measure on the average utility $\overline{U}(\underline{r},\underline{d})$ or average expected utility $\overline{EU}(\underline{r},\underline{d})$ for collections of individually comparable forecasts. However, we can identify a special situation, in which, under certain assumptions, average validity is an ordinal measure on average utility and average expected utility. The relationships, on an ordinal scale, between the measures of concern in this situation are summarized in RESULT 8-2.

RESULT 8-2 (N States, K Forecasts)

Assumptions. (1) $DA_k > DB_k$ for all k (k=1,...,K).

(2) A's and B's forecasts are individually comparable.

Results. The ordinal relationships described in RESULT 8-1 hold for average validity, average utility, and average expected utility.

Thus, for collections of individually comparable forecasts, the average validity $\overline{D}(\underline{r},\underline{d})$ represents an ordinal measure on the average utility $\overline{U}(\underline{r},\underline{d})$ and average expected utility $\overline{EU}(\underline{r},\underline{d})$ in this special situation.

The assumptions that $DA_k > DB_k$ for all k (k=1,...,K) and that A's and B's forecasts are individually comparable, which yield RESULT 8-2 are, of course, very restrictive assumptions. That is, collections of A's and B's forecasts will not often satisfy these assumptions in N-state ($N \geq 2$) situations. On the other hand, since all forecasts are comparable in two-state situations, the ordinal relationships described in RESULT 8-2 hold under somewhat less restrictive assumptions in such situations.

8.1.3 Some Examples

In this subsection we present sample forecasts and consider general and special situations which illustrate the nature of the relationships and scope of the results described in subsections 8.1.1 and 8.1.2. For the purposes of this presentation, we consider two-state (N=2) and N-state (N>2) situations separately in paragraphs

8.1.3.1 and 8.1.3.2, respectively.

8.1.3.1 Two-State (N=2) Situations

In Table 8-1 we present a sample collection of forecasts prepared by forecasters A and B for a two-state situation. Consider A's and B's forecasts on the first occasion (k=1). Note that $VA_1 > VB_1$. Thus, $UA_1 \geq UB_1$ for all decision situations. In particular, for the decision situation described in Figure 5-1a, $UA_1 = UB_1$ (=1). Now, consider A's and B's forecasts on the second occasion (k=2). Since $VA_2 > VB_2$, $UA_2 \geq UB_2$ for all decision situations. For the decision situation described in Figure 5-1a, UA_2 (=2/3) > UB_2 (=0). Thus, as RESULT 8-1 indicates, the validity measure $D(\underline{r},\underline{d})$ represents an ordinal measure on utility for individual (comparable) forecasts. A similar relationship holds for expected utility, in which case the decision maker's utilities are assumed to be random variables with a joint probability distribution (see Chapters 5 and 7). If the probability distribution of concern is (strictly) positive, then the ordinal relationships between the expected utilities of A's and B's forecasts will involve strict inequalities.

Now, consider the forecasts presented in Table 1 as two collections of forecasts. Note that $VA_k \not> VB_k$ for all k (k=1,2,3). Thus, even though $\overline{VA} > \overline{VB}$, either $\overline{UA} \geq \overline{UB}$ or $\overline{UA} \leq \overline{UB}$ in general. For the decision situation described in Figure 5-1a, $\overline{UA}$ (=0.778) > $\overline{UB}$ (=0.667). Now, consider subcollections of this collection of forecasts. First, note that $VA_k > VB_k$ (k=1,2). Thus, $\overline{UA} \geq \overline{UB}$ in general, and for the particular decision situation of concern,

Forecast Number k	Forecasts $\underline{r} = (r_1, r_2)$ $\underline{r}_k^A$	$\underline{r}_k^B$	Observation $\underline{d} = (d_1, d_2)$ $\underline{d}_k$	Validity Measure $D(\underline{r},\underline{d})$ VA_k	VB_k	Utility Measure $U(\underline{r},\underline{d})$ UA_k	UB_k
1	(0.95,0.05)	(0.55,0.45)	(1,0)	0.9975	0.7975	1	1
2	(0.30,0.70)	(0.60,0.40)	(0,1)	0.9100	0.6400	2/3	0
3	(0.45,0.55)	(0.55,0.45)	(1,0)	0.6975	0.7975	2/3	1
Average				0.8683	0.7450	0.778	0.667

Table 8-1. A sample collection of three forecasts prepared by forecasters A and B, the relevant observations, and the validities and utilities of the forecasts according to the measures $D(\underline{r},\underline{d})$ and $U(\underline{r},\underline{d})$, respectively, for two-state situations. The utilities relate to the decision situation presented in Figures 4-6 and 5-1a.

$\overline{UA}$ (=0.833) > $\overline{UB}$ (=0.500). On the other hand, although $\overline{VA}$ (=0.8475) > $\overline{VB}$ (=0.7975) for the subcollection which includes the first and third forecasts, VA_k, VB_k (k=1,3) and, for the decision situation of concern, $\overline{UA}$ (=0.833) < $\overline{UB}$ (=1.000). Therefore, as RESULT 8-2 indicates, the average validity is, in general, an ordinal measure on the average utility only if the validity of each of A's forecasts exceeds the validity of the corresponding member of B's set of forecasts. Similar results hold for expected-utility measures in situations in which the evaluator's knowledge of the relevant utilities is expressed in terms of a probability distribution. As indicated in subsection 8.1.2, collections of forecasts, even in two-state situations, will not often satisfy the assumptions which yield RESULT 8-2.

8.1.3.2 N-State (N>2) Situations

In Table 8-2 we present sample forecasts prepared by forecasters A and B for three-state situations. These forecasts are depicted in Figure 8-1 within the framework of the simplex R^3 and the decision situation presented in Figures 4-8 and 5-1b. Consider A's and B's forecasts on the first occasion (k=1). Note, in Figure 8-1, that the points $\underline{r}_1^A$ and $\underline{r}_1^B$ do <u>not</u> lie on the same line segment through $\underline{d}_2$. Thus, A's and B's forecasts on this occasion are not comparable. Therefore, in general, either $UA_1 \geq UB_1$ or $UA_1 \leq UB_1$ (even though $VA_1 > VB_1$). In particular, for the decision situation described in Figures 4-8 and 5-1b, UA_1 (=2/3) < UB_1 (=5/6) even though VA_1

Forecast Number	Forecasts $\underline{r} = (r_1, r_2, r_3)$		Observation $\underline{d} = (d_1, d_2, d_3)$	Validity Measure $D(\underline{r},\underline{d})$		Utility Measure $U(\underline{r},\underline{d})$	
k	$\underline{r}_k^A$	$\underline{r}_k^B$	$\underline{d}_k$	VA_k	VB_k	UA_k	UB_k
1	(0.05,0.40,0.55)	(0.30,0.15,0.55)	(0,1,0)	0.6675	0.4425	2/3	5/6
2	(0.15,0.80,0.05)	(0.30,0.60,0.10)	(0,1,0)	0.9675	0.8700	5/6	5/6
3	(0.15,0.70,0.15)	(0.20,0.50,0.30)	(0,1,0)	0.9325	0.8100	5/6	5/6
Average				0.8558	0.7075	0.778	0.833

Table 8-2. A sample collection of forecasts prepared by forecasters A and B, the relevant observations, and the validities and utilities of the forecasts according to the measures $D(\underline{r},\underline{d})$ and $U(\underline{r},\underline{d})$, respectively, for three-state situations. The utilities relate to the decision situation presented in Figures 4-8 and 5-1b.

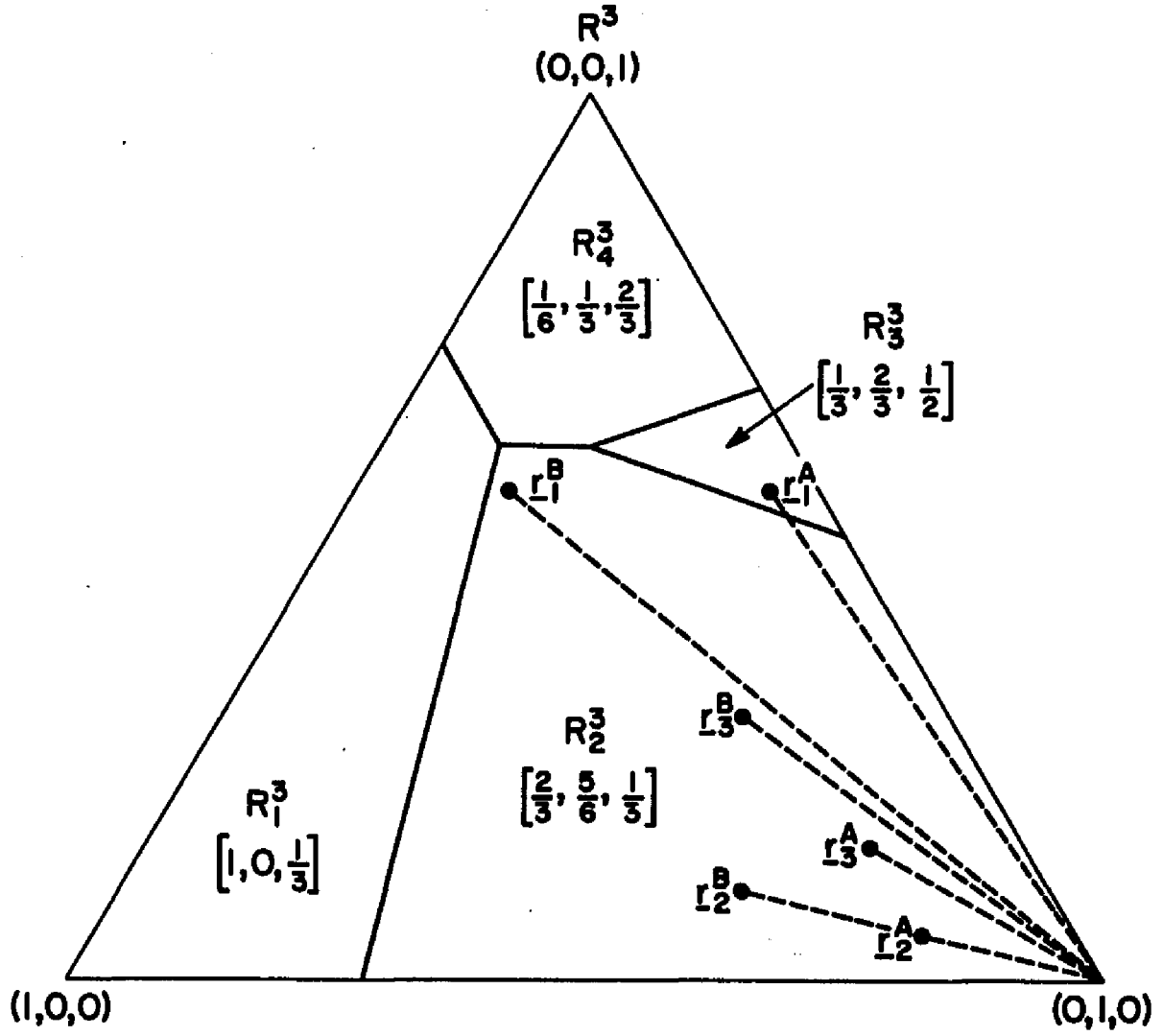


Figure 8-1. The representation of the forecasts presented in Table 8-2 within the framework of the simplex R^3 and the decision situation presented in Figures 4-8 and 5-1b.

(=0.6675) > VB_1 (=0.4425). Now, consider A's and B's forecasts on the second occasion (k=2). Note, in Figure 8-1, that the points $\underline{r}_2^A$ and $\underline{r}_2^B$ lie on the same line segment through $\underline{d}_2$. Thus, A's and B's forecasts on this occasion are comparable, and, since $VA_2 > VB_2$, $UA_2 \geq UB_2$ for all decision situations. For the decision situation depicted in Figures 4-8 and 5-1b, $UA_2 = UB_2$ (=5/6). Thus, as indicated in subsection 8.1.1, the validity measure $D(\underline{r},\underline{d})$ represents an ordinal measure on utility for individual, comparable forecasts. A similar relationship holds for expected utility. Unfortunately, as indicated in subsection 8.1.1, forecasts are not often comparable in N-state situations. However, if the forecasts of concern are nearly comparable, then these ordinal relationships can be expected to hold in most situations. For example, A's and B's forecasts on the third occasion (k=3) correspond to points on two line segments through $\underline{d}_2$ which are close to one another. Thus, since $VA_3 > VB_3$, $UA_3 \geq UB_3$ for most, but not all, decision situations.

With regard to collections of forecasts, the sample forecasts presented in Table 8-2 indicate that, unless the forecasts of concern are individually comparable, average validity is not an ordinal measure on average utility (or on average expected utility). Note that the forecasts on the first and third occasions are not comparable. Thus, even though $VA_k > VB_k$ for all k (k=1,2,3) (and, as a result, $\overline{VA} > \overline{VB}$), either $\overline{UA} \geq \overline{UB}$ or $\overline{UA} \leq \overline{UB}$ in general. In particular, for the decision situation depicted in Figure 8-1, $\overline{UA}$

($=0.778$) < $\overline{UB}$ ($=0.833$).

As indicated in paragraph 8.1.3.1, similar results hold for expected utility in situations in which the evaluator's knowledge of the utilities is expressed in terms of a probability distribution.

8.1.4 Some Speculations

The results and examples presented in the previous subsections of this section (8.1) indicate that ordinal relationships between measures of validity and measures of utility need exist only in very special situations. Therefore, in the absence of specific information concerning the utilities of the relevant decision makers, the fact that A's forecasts are more valid than B's forecasts does not, in general, guarantee that the utility of A's forecasts will be greater than (or even equal to) the utility of B's forecasts. This result raises an important and interesting question: In the absence of any proof of the existence of <u>general</u> ordinal relationships (at least at the present time), what "evidence" do we have that any increases in the validity of probability forecasts in the future will lead to corresponding increases in their utility? While we cannot offer any formal mathematical proof that such (ordinal) relationships do indeed exist, we will speculate briefly regarding the possible existence of such relationships. In this regard, consider Figure 8-2 in which we have depicted the decision situation described in Figures 4-8 and 5-1b within the framework of the simplex R^3. This figure also contains the arcs of two circles centered at

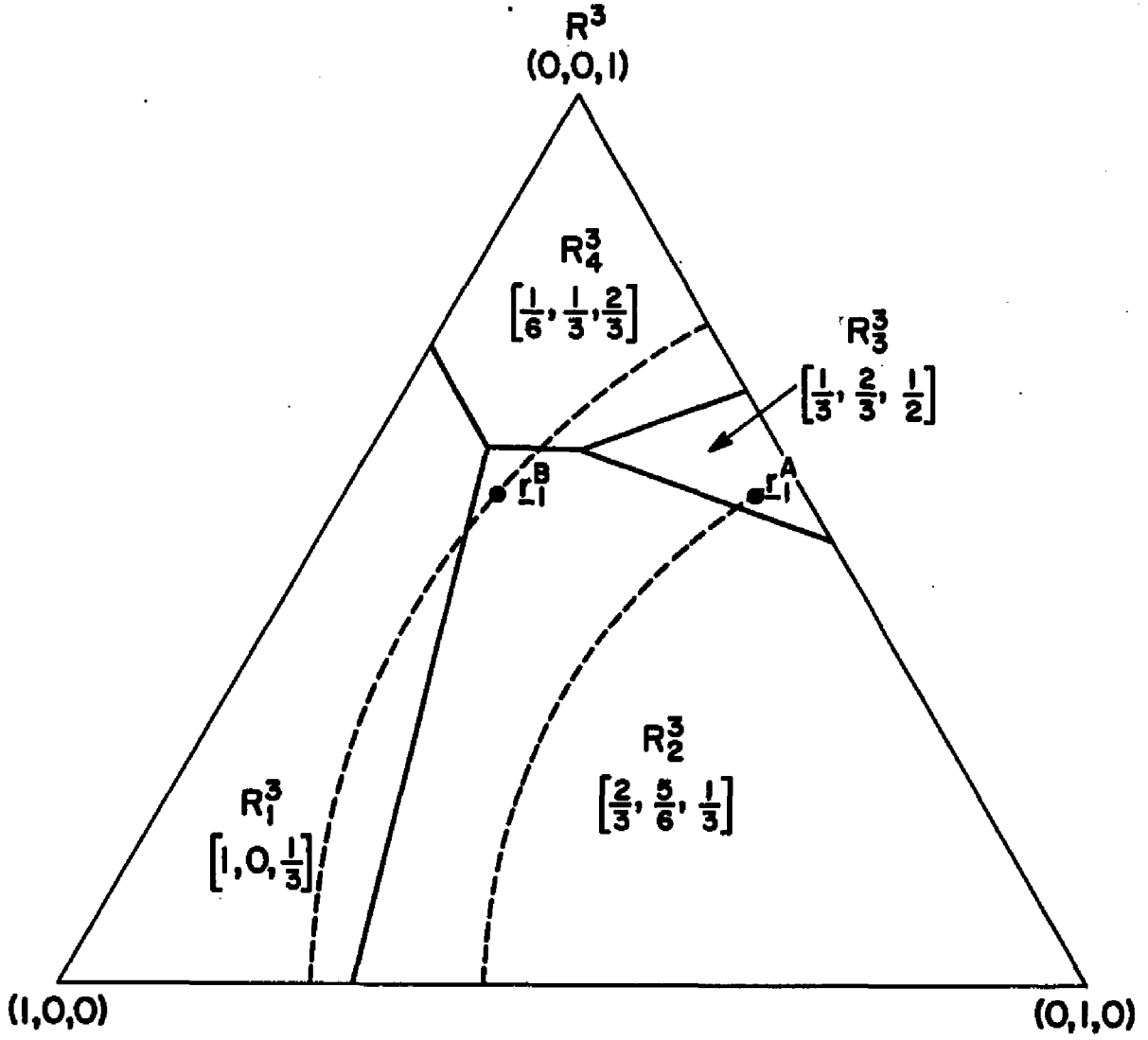


Figure 8-2. The decision situation presented in Figures 4-8 and 5-1b depicted within the framework of the simplex R^3. Each arc represents a set of forecasts of equal validity according to the validity measure $D_2(\underline{r})$. These arcs were chosen to include the forecasts $\underline{r}_1^A$ and $\underline{r}_1^B$ (see Table 8-2 and Figure 8-1).

the vertex $\underline{d}_2$ [=(0,1,0)], each of which represents a set of forecasts of equal validity according to the distance measure $D(\underline{r},\underline{d})$. Each of these arcs was chosen to include a forecast from the set of forecasts presented in Table 8-2 and depicted in Figure 8-1. Specifically, the "inner" arc (i.e., the arc closest to the vertex $\underline{d}_2$) includes the forecast $\underline{r}_1^A$, while the "outer" arc includes the forecast $\underline{r}_1^B$. Now, we have already shown in the previous subsection (8.1.3) that, although $VA_1 > VB_1$ according to the measure $D_2(\underline{r})$, $UA_1 < UA_2$. However, consider the two arcs in their entirety rather than just the two points $\underline{r}_1^A$ and $\underline{r}_1^B$, and suppose that forecasters A and B have made sets of forecasts that have fallen on the inner and outer arcs, respectively. Almost all of A's forecasts will fall within the region R_2^3 for which $U_2(\underline{r})$ is a maximum (5/6), while many of B's forecasts will fall in regions (namely, R_1^3 and R_4^3) for which $U_2(\underline{r})$ <u>necessarily</u> assumes a smaller value (0 and 1/3, respectively). Thus, for most (but not all) distributions of forecasts on these two arcs, the average utility of A's forecasts, $\overline{UA}$, will be greater than the (average) utility of B's forecasts, $\overline{UB}$. Note that this statement holds for any two distinct arcs in the simplex R^3. Further, the closer an arc is to the vertex $\underline{d}_n$ the more likely the forecasts are to achieve the maximum (average) utility regardless of the nature of the decision situation. We summarize this discussion in terms of CONJECTURE 8-1.

CONJECTURE 8-1

If the average validity of A's forecasts is greater than the average validity of B's forecasts, then the average utility of A's forecasts will, in almost all situations, be greater than the average utility of B's forecasts.

While CONJECTURE 8-1 obviously lacks rigorous proof, we believe that this result also depends primarily upon the fact that the regions R_m^N (m=1,...,M) are convex.

We should note that CONJECTURE 8-1 also applies to the expected utility of A's and B's forecasts in situations in which the evaluator's knowledge of the relevant utilities is expressed in probabilistic terms.

8.2 Special Decision Situations

In this section we examine the relationships between expected-utility measures and measures of (primary) validity in two special decision situations. First, we consider a special two-action, two-state situation which is equivalent to the cost-loss ratio decision situation. Then, we investigate a M-action, M-state generalization of the cost-loss situation. In both situations, we assume that the evaluator's knowledge of the utilities of concern is expressed in terms of a uniform probability distribution. Finally, we briefly discuss the nature of the general relationship between measures of

validity and measures of utility and expected utility.

8.2.1 A Two-Action, Two-State Situation

Consider the special two-action, two-state decision situation described by the key utility matrix (c) in Figure 5-3. If we assume that the evaluator's knowledge of the utility x is expressed in terms of a uniform probability distribution [i.e., if $x \sim \beta(1,1)$], then, from (7-13),

$$EU(\underline{r},\underline{d}) = d_1 r_1 + (1/2)(1 - r_1^2),$$

or

$$EU(\underline{r},\underline{d}) = (1/2)(1 + d_1) - (1/2)(r_1 - d_1)^2. \qquad (8\text{-}1)$$

In the two-state situation, the primitive (absolute) distance measure $\Delta^2(\underline{r},\underline{d})$ can be expressed as follows:

$$\Delta^2(\underline{r},\underline{d}) = 2(r_1 - d_1)^2$$

[see (6-3)]. Thus, (8-1) can be rewritten as follows:

$$EU(\underline{r},\underline{d}) = (1/2)(1 + d_1) - (1/4)\Delta^2(\underline{r},\underline{d}) \qquad (8\text{-}2)$$

(see Murphy, 1966). Therefore, the expected-utility measure $EU(\underline{r},\underline{d})$

is equivalent to the primitive distance measure $\Delta^2(\underline{r},\underline{d})$. That is, the expected utility of a forecast $\underline{r}$ in this special situation depends only upon the score assigned by the primitive distance measure $\Delta^2(\underline{r},\underline{d})$ and the observation $\underline{d}$. In Appendix H we show that $EU(\underline{r},\underline{d})$ is a linear function of $\Delta^2(\underline{r},\underline{d})$ in this decision situation if and only if the evaluator's knowledge of the utility x is expressed in terms of a uniform distribution.

In order to obtain an expected-utility measure for this decision situation whose values are less dependent upon the state that obtains, we shall consider the decision situation described by this same utility matrix when the states (i.e., the columns) are interchanged. If we let $EU^*(\underline{r},\underline{d})$ denote the expected-utility measure in this situation, then

$$EU^*(\underline{r},\underline{d}) = (1/2)(1 + d_2) - (1/4)\Delta^2(\underline{r},\underline{d}). \qquad (8\text{-}3)$$

Further, if we let $TEU(\underline{r},\underline{d})$ denote the sum of the measures $EU(\underline{r},\underline{d})$ and $EU^*(\underline{r},\underline{d})$, then

$$TEU(\underline{r},\underline{d}) = (3/2) - (1/2)\Delta^2(\underline{r},\underline{d}). \qquad (8\text{-}4)$$

The range of the measure $TEU(\underline{r},\underline{d})$ is the closed interval $[1/2,3/2]$ and $TEU(\underline{r},\underline{d})$ has a positive orientation. If we let $REU(\underline{r},\underline{d})$ denote a "representative" expected-utility measure for this decision situation (the special two-action, two-state situation with a uniform

probability distribution on the utility x) which is equivalent to TEU($\underline{r}$,$\underline{d}$) in (8-4) and which possesses a unit interval range, then

$$\text{REU}(\underline{r},\underline{d}) = 1 - (1/2)\Delta^2(\underline{r},\underline{d}). \tag{8-5}$$

Now, from (6-4),

$$\text{REU}(\underline{r},\underline{d}) = \text{D}(\underline{r},\underline{d}). \tag{8-6}$$

That is, the representative expected-utility measure for this situation is identical to the absolute distance measure, the basic absolute (primary) validity measure in this study.

The special two-action, two-state decision situation considered in this subsection is equivalent, in a decision-making sense, to the familiar cost-loss ratio decision situation (see Chapter 5). Specifically, $x = 1 - (C/L)$, where C and L are the cost of taking protective action and the loss incurred if protective action is not taken and adverse weather occurs, respectively. Thus, since the quadratic, or probability, score $Q(\underline{r},\underline{d})$ is identical to the primitive distance measure $\Delta^2(\underline{r},\underline{d})$ [see (6-3) and (6-19)], the former can also be considered to be a measure of expected-utility in this special situation (Murphy, 1966).

8.2.2 A M-Action, M-State Situation

A M-action, M-state generalization of the cost-loss ratio decision situation has been described by Epstein (1969). In this

situation the set a of actions $\{a_1,\ldots,a_M\}$ consists of M decreasing levels of protection and the set s of states $\{s_1,\ldots,s_M\}$ consists of M decreasing degrees of severity of the weather. The cost-loss matrix in this situation when M = 5 is presented in Figure 8-3. When state s_j obtains, the representative expected-utility measure (see subsection 8.2.1) in this situation, when the utility x is assumed to possess a uniform probability distribution, is $REU_j(\underline{r})$, where, from Epstein (op. cit.),

$$REU_j(\underline{r}) = (3/2) - [1/2(N-1)]\sum_{n=1}^{N-1}[\sum_{m=1}^{n} r_m)^2 + (\sum_{m=n+1}^{N} r_m)^2] - [1/(N-1)]\sum_{n=1}^{N}|n-j|r_n. \quad (j=1,\ldots,N) \tag{8-7}$$

Murphy (1970) has shown that $REU_j(\underline{r})$ in (8-7) can be expressed as

$$REU_j(\underline{r}) = 1 - [1/(N-1)][\sum_{n=1}^{j-1}(\sum_{m=1}^{n} r_m)^2 + \sum_{n=j}^{N-1}(\sum_{m=n+1}^{N} r_m)^2]. \quad (j=1,\ldots,N) \tag{8-8}$$

Further, Murphy (1971) has shown that $REU(\underline{r},\underline{d})$ in this situation can be expressed as

$$REU(\underline{r},\underline{d}) = 1 - [1/(N-1)][\sum_{n=1}^{N}(\sum_{m=1}^{n} r_m - \sum_{m=1}^{n} d_m)^2]. \tag{8-9}$$

A comparison of (8-9) and (6-27) reveals that

		States				
		s_1	s_2	s_3	s_4	s_5
	a_1	x	x	x	x	x
	a_2	(3x+1)/4	3x/4	3x/4	3x/4	3x/4
Actions	a_3	(2x+2)/4	(2x+1)/4	x/2	x/2	x/2
	a_4	(x+3)/4	(x+2)/4	(x+1)/4	x/4	x/4
	a_5	1	3/4	1/2	1/4	0

Figure 8-3. The cost-loss matrix in the M-action, M-state generalization of the cost-loss ratio decision situation when M = 5. The variable x represents the cost-loss ratio (i.e., x = C/L).

$$REU(\underline{r},\underline{d}) = RPS(\underline{r},\underline{d}). \qquad (8\text{-}10)$$

Therefore, the representative expected-utility measure for this situation is identical[1] to the ranked probability score.

8.2.3 General Relationships between Validity and Utility Measures: Discussion

In subsections 8.2.1 and 8.2.2 we have shown that two familiar measures of validity are equivalent to the (representative) expected-utility measures associated with two special decision situations when the evaluator's knowledge is expressed in terms of a uniform distribution. Murphy (1969a) and Stael von Holstein (1970b) have discussed and/or formulated expected-utility measures associated with these same situations when the evaluator's knowledge was expressed in more general terms (e.g., in terms of a beta distribution). The process of formulating these expected-utility measures generally leads to very complex expressions (cf. Chapter 7) which bear no obvious relationship to known measures of validity (or other empirical attributes). Evidently, then, one desirable property of empirical

---

[1]The ranked probability score is generally expressed simply as $RPS(\underline{r},\underline{d}) = \sum_{n=1}^{N}(\sum_{m=1}^{n} r_m - \sum_{m=1}^{n} d_m)^2$ (see Murphy, 1971), which is still equivalent to the measure $REU(\underline{r},\underline{d})$.

evaluation measures is simplicity. In any case, the question arises as to the general relationship between measures of validity and measures of utility. A definitive answer to this question has been provided by Savage (1971), who states that "... every proper [strictly proper in the terminology of this thesis] scoring rule can be viewed as a share in a business and ... every such share leads to an at least weakly proper [proper] scoring rule ..." (p. 799). Clearly, this result has important implications for forecast evaluation in general and for the relationship between empirical and operational evaluation measures in particular, and these implications are in need of further study (see Chapter 9).

CHAPTER NINE. CONCLUSION

This chapter consists of three sections. In Section 9.1, we briefly summarize and discuss the principal results and conclusions forthcoming from this study. Several possible extensions to the study are indicated in Section 9.2, and we briefly describe an alternative (geometrical) framework for the evaluation of probability forecasts. The final section (9.3) contains a brief postscript.

9.1 Summary and Discussion

In this study we initially emphasized the importance of the steps in the preliminary (or formulation) phase of the evaluation process. These steps include the identification of the purposes for evaluation and the description, in quantitative terms, of the desirable attributes of the forecasts with respect to these purposes. In Chapter 2 we identified two general purposes, to determine the (absolute and relative) degree of correspondence between the forecasts and the observations and to determine the "value" of the forecasts to users or decision makers. Each of these purposes leads to the identification of a distinct form of evaluation, and we have referred to these forms as empirical evaluation and operational evaluation, respectively. In general, the former is of concern to the meteorologist, while the latter is of concern to the decision maker. However, since they may often be interested in both forms of evaluation, we have used the term "evaluator" to describe these and other

individuals concerned with the problem of determinating the "goodness" of forecasts.

Normative and substantive attributes of probability forecasts were described in Chapter 3. The normative attributes relate to the assessment process (i.e., the process of formulating the forecasts), while the substantive attributes relate to the evaluation process. With regard to the former, the subjective view of probability, the view to which we subscribe in this study, prescribes that (subjective) probabilities should be both consistent and honest. The latter requirement implies that a meteorologist's assessments or forecasts should correspond to his judgments. One method of encouraging the forecaster to make honest assessments (i.e., not to hedge) is to reward him in accordance with a strictly proper scoring rule. Improper scoring rules, on the other hand, encourage hedging. Thus, the question of whether an evaluation measure is strictly proper or improper became an important consideration in this study (even though we are primarily concerned with the evaluation problem rather than the assessment problem). With regard to substantive attributes, desirable properties of probability forecasts with respect to both forms of evaluation were identified. The principal attributes with respect to empirical evaluation are validity and bias (or reliability). These attributes refer to the correspondence between forecasts (i.e., probabilities) and observations on an individual and collective basis, respectively. Other secondary attributes were also described. The principal (and only) attribute

considered with respect to operational evaluation was utility (in a decision-theoretic sense).

Frameworks for empirical and operational evaluation were described in Chapter 4 in terms of their primitive elements. First, a geometrical framework, the regular simplex, was identified, within which probability forecasts could be represented as "points." When the observations were also associated with points in this geometrical framework, the simplex became a framework for empirical evaluation. The framework for operational evaluation was based upon a modified version of the subjective, expected-utility (SEU) model of a decision situation. The SEU model is, of course, consistent with the subjective view of probability which we have adopted in this study. The principal assumption which we have made in applying the SEU model to the evaluation problem was to assume that the decision maker adopts the meteorologist's probabilities as his own subjective probabilities. This assumption appears to be reasonable in view of our primary concern in this study with the evaluation of the meteorologist's forecasts. We then demonstrated that this (modified) model could be represented within the framework of the regular simplex and that, as a result, this "augmented" simplex represents a framework for operational evaluation. Finally, we indicated that the fact that the set of primitive elements for the framework for operational evaluation subsumes the set of primitive elements for the framework for empirical evaluation implies that the former (and its geometrical representation) is a comprehensive framework for

evaluation. The existence of such a framework is of particular importance when the question of the relationships between empirical and operational evaluation is considered.

In Chapter 5, the specification of the primitive elements which constitute the evaluation frameworks was considered. The evaluator was assumed to be able to specify all of the primitives except the (decision maker's) utilities in a complete, i.e., precise, manner. The evaluator's knowledge of the utilities could be complete or incomplete, and, in the latter case, his knowledge was assumed to be expressed in terms of inequality or probability measures. Several examples of such measures were considered. Further, we indicated that the treatment of the utilities as an uncertain quantity seems to be particularly appropriate when the evaluator is a meteorologist, since the latter's knowledge of the utilities is seldom if ever complete. Two possible interpretations of the probabilistic measures of knowledge in this context were discussed.

Within the (geometrical) framework for empirical evaluation described in Chapter 4, the distance between the forecasts and the observations is a natural measure of their degree of correspondence, and distance measures of the principal and supplemental attributes were formulated in Chapter 6. These measures were then compared with other measures proposed heretofore, with particular emphasis upon absolute and relative measures of the attribute validity. With

regard to absolute measures, a number of strictly proper scoring rules were described, and we indicated that (other) definitive criteria upon which to base the choice of a particular measure are lacking. In this regard, however, all strictly proper scoring rules are (believed to be) increasing functions of the probability assigned to the state that obtains along certain "directed" line segments in the simplex (i.e., when the forecasts are comparable). Since this property is of particular importance in our study of the relationships between measures of validity and measures of utility, any strictly proper scoring rule can be used as a standard of comparison. With regard to relative measures, the distance measure was shown to be the only strictly proper measure, although two other relative measures were approximately strictly proper for large collections of forecasts ($K \geq 100$). Measures of the attributes bias (or reliability), resolution, and sharpness were also described and compared. With regard to the first two attributes, distance measures based upon a vector partition of the primitive distance measure were preferred, since we believe that forecasts should be considered to be vectors rather than scalars. Two tables were presented that summarize the form and properties of the measures considered. In general, the members of the family of distance measures compare favorably with their "competitors."

Measures of the "value" of forecasts were formulated in Chapter 7 for situations in which the evaluator's knowledge of the utilities

was both complete and incomplete. Examples of the expected-utility measures associated with particular decision situations and states of the evaluator's knowledge were presented. The behavior and properties of these measures were investigated. In particular, we demonstrated that utility and expected-utility measures are proper and strictly proper (or proper) scoring rules, respectively. Further, on the basis of the examples and other considerations, we speculated that these measures are increasing and strictly increasing (or increasing) functions of the probability (r_j say) assigned to the state that obtains (s_j) for comparable forecasts. Finally, we showed (via examples) that, as expected, expected-utility measures approach utility measures as the evaluator's knowledge increases (i.e., as his probability distribution becomes more concentrated).

The results obtained in Chapters 6 and 7 with regard to empirical and operational evaluation measures, respectively, provided the basis for an investigation of the relationships between measures of validity and utility in Chapter 8. We were concerned in this chapter with both the ordinal relationships between these measures in general as well as the specific relationships between such measures in special situations. With regard to the former, the fact that both (strictly proper) validity measures and utility and expected-utility measures are strictly increasing (or increasing) functions of r_j for comparable forecasts means that validity measures provide an ordinal measure on utility and expected-utility for such

forecasts. While two forecasts will seldom fall precisely on the same line segment, they may fall on "adjacent" line segments, in which case the ordinal relationships will hold for most decision situations. Moreover, since all line segments are directed line segments in the two-state (N=2) situation, this ordinal relationship holds in general in this situation for individual forecasts. Unfortunately, for collections of forecasts this relationship appears to hold, in general, only if the individual relationship holds for each pair of forecasts in the respective collections. In summary, we were able to show that the fact that one forecast is more valid than another forecast does not guarantee that the former will be of greater utility than the latter unless they are comparable (except in the two-state situation in which all forecasts are comparable). On the other hand, we speculated on the basis of the properties of the framework for operational evaluation (i.e., convexity) that, in most situations, the fact that one set of forecasts is more valid than another set implies that the former set will be of greater utility than the latter set.

With regard to specific relationships between measures of validity and measures of expected-utility, we showed that the measures of expected-utility associated with two special decision situations were equivalent (i.e., linearly related) to two familiar strictly proper scoring rules, namely the distance or quadratic measure $D(\underline{r},\underline{d})$ and the ranked probability score $RPS(\underline{r},\underline{d})$. These results provide

specific evidence of the statement by Savage (1971, p. 799) that every strictly proper scoring rule corresponds to a particular decision situation and every decision situation corresponds to a proper or strictly proper scoring rule. As indicated in Chapter 8, this result obviously has important implications for the evaluation problem in meteorology and other contexts.

9.2 Some Extensions

Many possible extensions to this study suggest themselves, and we indicate a number of such possibilities in subsection 9.2.1. In addition, we briefly describe an alternative geometrical framework for the evaluation of probability forecasts in subsection 9.2.2.

9.2.1 A Potpourri of Extensions

The specific (as opposed to the general) purposes for evaluation need further study and such an investigation could lead to the identification of new forms, or sub-forms, of evaluation. Furthermore, such studies might identify important new attributes and/or lead to refined definitions of attributes proposed heretofore.

While the evaluation frameworks described in this study possess certain attractive features (at least to the author), the advantages and disadvantages of this "approach" to the evaluation problem are in need of further study. In addition, this framework should be compared with the frameworks proposed by other investigators (Raiffa, 1969; Roberts, 1968).

With regard to the specification of the primitive elements in any such framework, the implications of a lack of complete knowledge regarding primitives other than the decision maker's utilities should be investigated. From a meteorological point of view, empirical studies of the meteorologist's state of knowledge with respect to decision maker's utilities would be of some interest.

The need to identify additional criteria which can be used as a basis for comparing empirical evaluation measures is apparent. Such criteria might include the sensitivity of strictly proper scoring rules to hedging (see, for example, Murphy and Winkler, 1970), and the formulation of "indicators" of the degree to which scoring rules actually "measure" the relevant attributes. Empirical studies involving the use of strictly proper and improper scoring rules in connection with probability assessment tasks would also be valuable to determine the magnitude of the impact of such measures on hedging.

With regard to operational evaluation, the need exists to identify representative utility and/or expected-utility measures which provide reasonable measures of the "value" of probability forecasts to specific groups of decision makers. In addition, the speculations regarding the properties of these measures should be proved either true or false.

We consider the studies of the relationships between empirical and operational evaluation measures to be a matter of the greatest importance. In particular, a much more comprehensive study of the

conditions under which validity measures, and measures of other attributes, are or are not ordinal measures on utility and expected-utility is needed. In addition, the expected-utility measures associated with a variety of decision situations and states of knowledge should be investigated to identify simple and easy to use measures such as the ranked probability score (Epstein, 1969; see also Murphy, 1971) which are also measures of "value" for representative decision situations.

9.2.2 An Alternative Framework

In this subsection we briefly describe an alternative geometrical framework for the evaluation of probability forecasts.

Consider the ranked probability score RPS($\underline{r}$,$\underline{d}$) in (6-27), and let

$$R_n = \sum_{m=1}^{n} r_m \tag{9-1}$$

and

$$D_n = \sum_{m=1}^{n} d_m. \tag{9-2}$$

Then,

$$RPS(\underline{r},\underline{d}) = (\underline{R} - \underline{D})(\underline{R} - \underline{D})' = \sum_{n=1}^{N} (R_n - D_n)^2, \tag{9-3}$$

where $\underline{R} = (R_1,\ldots,R_N)$ and $\underline{D} = (D_1,\ldots,D_N)$,[1] We shall refer to $\underline{R}$ and $\underline{D}$ as a cumulative forecast and a cumulative observation, respectively (see Murphy, 1971, 1972c). If we denote the set of all possible cumulative forecasts by Π^N, then, from (9-1), Π^N is an irregular (N-1)-dimensional simplex defined by

$$\Pi^N = \{(R_1,\ldots,R_N) \mid R_i \geq 0, R_k \geq R_j \text{ for } i<j<k<N,\ R_N = 1\} \qquad (9\text{-}4)$$

(Murphy and Stael von Holstein, 1975). A cumulative forecast $\underline{R} = (R_1,\ldots,R_N)$ corresponds to a point in this simplex (and vice versa), and the N vertices of the simplex represent the N cumulative observations $\underline{D}_n$ and the N categorical forecasts $\underline{R}_n$ (n=1,...,N).

The simplex for three-state (N=3) situations, Π^3, is depicted in a cartesian coordinate system in Figure 9-1 as a right, isosceles triangle (heavy solid lines) on the top face of the unit cube. This triangle is then a framework for the representation of cumulative forecasts and observations in three-state situations. The regular two-dimensional simplex, the equilateral triangle R^3, which represents the set of all forecasts $\underline{r}$, is shown in dashed lines. Since a one-to-one correspondence exists between a forecast $\underline{r}$ and a cumulative

---

[1] We have chosen to express the measure RPS($\underline{r}$,$\underline{d}$), or RPS($\underline{R}$,$\underline{D}$), in its simplest (equivalent) form in this subsection.

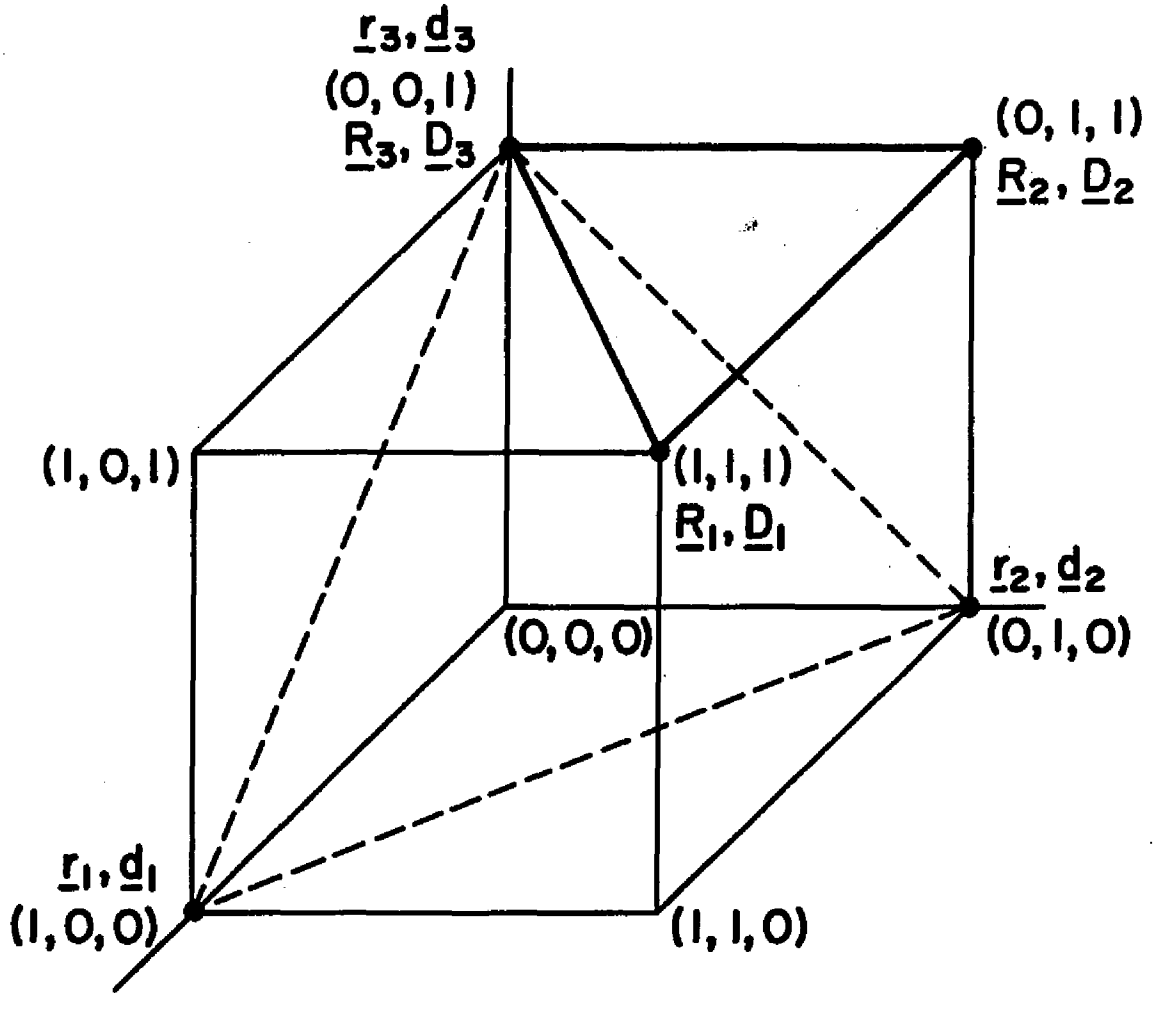


Figure 9-1. The framework for the representation of forecasts and observations in three-state (N=3) situations, the equilateral triangle R^3 (dashed lines), depicted in a cartesian coordinate system. The corresponding framework for the representation of cumulative forecasts and observations, a right, isosceles triangle Π^3 (heavy solid lines) on the top face of the unit cube (light solid lines), is also depicted in this figure.

forecast $\underline{R}$, a one-to-one correspondence exists between points in the two triangles.

The (Euclidean) distance between a cumulative forecast $\underline{R}$ and the relevant observation $\underline{D}$ in this framework is $\Delta(\underline{R},\underline{D})$, where

$$\Delta(\underline{R},\underline{D}) = [\sum_{n=1}^{N}(R_n - D_n)^2]^{1/2} \qquad (9\text{-}5)$$

Thus, from (9-3) and (9-5),

$$RPS(\underline{R},\underline{D}) = \Delta^2(\underline{R},\underline{D}), \qquad (9\text{-}6)$$

i.e., the RPS is the square of the distance between $\underline{R}$ and $\underline{D}$. The cumulative forecast $\underline{R} = (0.2,0.7,1.0)$, corresponding to the forecast $\underline{r} = (0.2,0.5,0.3)$, is depicted within the framework of the triangle Π^3 in a two-dimensional cartesian coordinate system in Figure 9-2. As previously indicated, the vertices of this triangle represent the cumulative observations $\underline{D}_n$ and the categorical cumulative forecasts $\underline{R}_n$ (n=1,2,3). The dashed lines in Figure 9-2 represent the distances between the cumulative forecast $\underline{R}$ and the cumulative observations $\underline{D}_1 = (1,1,1)$, $\underline{D}_2 = (0,1,1)$, and $\underline{D}_3 = (0,0,1)$. The squares of these distances are $RPS_1(\underline{R})$, $RPS_2(\underline{R})$, and $RPS_3(\underline{R})$, respectively.

Now, Murphy (1970) has demonstrated that the $RPS(\underline{R},\underline{D})$ is a weighted distance measure, while the measure $D(\underline{r},\underline{d})$ [or $Q(\underline{r},\underline{d})$] is an unweighted distance measure. More recently, Murphy and Stael von

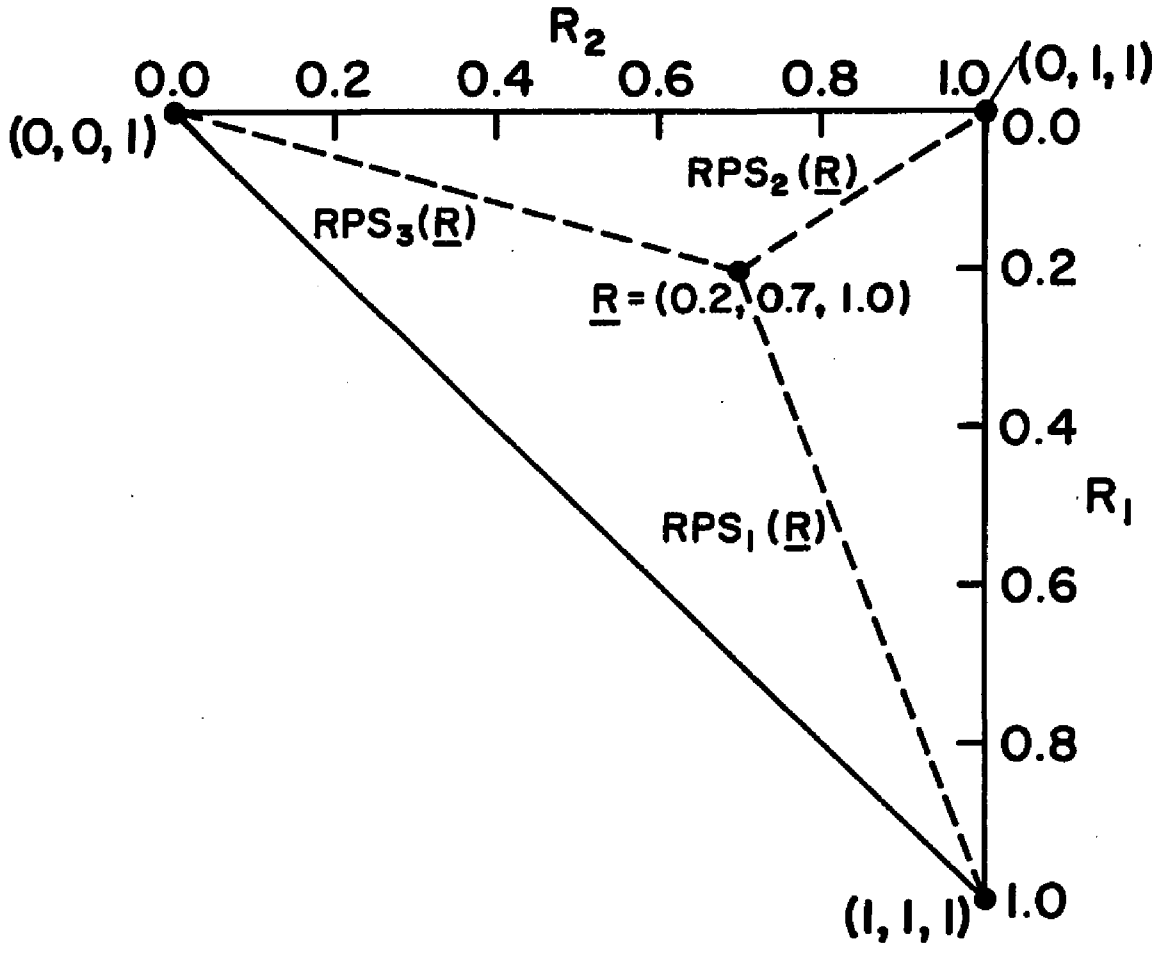


Figure 9-2. The cumulative forecast $\underline{R}$ = (0.2,0.7,1.0) depicted within the framework for the representation of cumulative forecasts and observations in three-state (N=3) situations, a right, isosceles triangle Π^3 in a two-dimensional cartesian coordinate system. The dashed lines represent the distances between the cumulative forecast $\underline{R}$ and the three cumulative observations $\underline{D}_1$ = (1,1,1), $\underline{D}_2$ = (0,1,1), and $\underline{D}_3$ = (0,0,1). The squares of these distances are $RPS_1(\underline{R})$, $RPS_2(\underline{R})$, and $RPS_3(\underline{R})$, respectively.

Holstein (1974) have shown that these measures are simply two different members of the family of quadratic scoring rules $GQ(\underline{r},\underline{d})$, where

$$GQ(\underline{r},\underline{d}) = (\underline{r} - \underline{d})\underline{A}(\underline{r} - \underline{d})', \qquad (9\text{-}7)$$

where $\underline{A}$ is an N-by-N matrix. The forms of the matrix $\underline{A}$ for the measures (a) $D(\underline{r},\underline{d})$ and (b) $RPS(\underline{R},\underline{D})$ are presented in Figure 9-3. At this time, we speculate that the members of this family of scoring rules will be strictly proper if $\underline{A}$ is symmetric and non-singular. In any case, the measure $GQ(\underline{r},\underline{d})$ appears to define a family of measures each of which is related to a different irregular simplex and each of which assigns different weights to the probabilities. The members of this family of scoring rules and their geometrical representations are obviously in need of further detailed study.

9.3 A Postscript

We conclude this thesis by presenting the following "innocent" comment by Goodman and Kruskal (1959, p. 153) which was the point of departure for this study:

> "If we attempt to construct a measure of association between probabilistic forecasts and actual events later observed, we are faced with association between an essentially continuous distribution on a k-1 dimensional simplex (k categories, probabilities for each that sum to one) and a

(a) $D(\underline{r},\underline{d})$

$$\underline{A} = \begin{matrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{matrix}$$

(b) $RPS(\underline{r},\underline{d})$

$$\underline{A} = \begin{matrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 \end{matrix}$$

Figure 9-3. The forms of the matrix $\underline{A}$ in Equation (9-7) when $N = 5$ for (a) the distance measure $D(\underline{r},\underline{d})$ and (b) the ranked probability score $RPS(\underline{r},\underline{d})$.

discrete distribution on k points (for the actual events)." The author hopes that some comment, speculation, or conjecture in this thesis will stimulate further interest and studies in this important area.

APPENDIX A

THE CLASS OF DISTANCE MEASURES $\Delta^k(\underline{r},\underline{d})$: "PROPER" AND IMPROPER SCORING RULES

In this appendix we show that the only member of the class of power functions of the form $\Delta^k(\underline{r},\underline{d})$, where

$$\Delta^k(\underline{r},\underline{d}) = [\sum_{n=1}^{N}(r_n - d_n)^2]^{k/2}, \qquad \text{(A-1)}$$

which is a strictly proper scoring rule is the squared distance measure $\Delta^2(\underline{r},\underline{d})$.

When state s_j obtains (i.e., when $d_j = 1$), the measure $\Delta^k(\underline{r},\underline{d})$ in (A-1) becomes $\Delta_j^k(\underline{r})$, where

$$\Delta_j^k(\underline{r}) = [(1 - r_j)^2 + \sum_{\substack{n=1 \\ n\neq j}}^{N} r_n^2]^{k/2},$$

or

$$\Delta_j^k(\underline{r}) = (1 - 2r_j + \sum_{n=1}^{N} r_n^2)^{k/2}. \qquad \text{(A-2)}$$

If we let $\Delta^k(\underline{r},\underline{p})$ denote the forecaster's subjective expected score (see Chapter 3), then

$$\Delta^k(\underline{r},\underline{p}) = \sum_{j=1}^{N} p_j \Delta_j^k(\underline{r}),$$

or, from (A-2),

$$\Delta^k(\underline{r},\underline{p}) = \sum_{j=1}^{N} p_j (1 - 2r_j + \sum_{n=1}^{N} r_n^2)^{k/2}. \qquad \text{(A-3)}$$

Now, taking the derivative of $\Delta^k(\underline{r},\underline{p})$ in (A-3) with respect to r_i, we obtain

$$\partial\Delta^k(\underline{r},\underline{p})/\partial r_i = (kr_i) \sum_{j=1}^{N} p_j (1 - 2r_j + \sum_{n=1}^{N} r_n^2)^{(k/2)-1} - kp_i (1 - 2r_j + \sum_{n=1}^{N} r_n^2)^{(k/2)-1},$$

and setting this result equal to zero yields

$$(r_i) \sum_{j=1}^{N} p_j (1 - 2r_j + \sum_{n=1}^{N} r_n^2)^{(k/2)-1} - p_i (1 - 2r_j + \sum_{n=1}^{N} r_n^2)^{(k/2)-1} = 0. \qquad \text{(A-4)}$$

Let

$$a_j = (1 - 2r_j + \sum_{n=1}^{N} r_n^2)^{(k/2)-1}. \qquad \text{(A-5)}$$

Then, (A-4) can be rewritten as follows:

$$r_i = (a_i / \sum_{j=1}^{N} p_j a_j) p_i. \qquad \text{(A-6)}$$

Now, $r_i = p_i$ in (A-6) if and only if

$$a_i = \sum_{j=1}^{N} p_j a_j. \qquad \text{(A-7)}$$

Note that (A-7) must hold for all i (i=1,...,N), which implies that the a_i must all be equal (i.e., $a_1 = \ldots = a_N$). From (A-5), this condition will hold, in general, only if k = 2 [i.e., only if $a_j = 1$ for all j (j=1,...,N)]. It can also be shown that the function $\Delta^k(\underline{r},\underline{p})$ in (A-3) is strictly convex. Thus, the only member of the class of power functions of the form $\Delta^k(\underline{r},\underline{d})$ which is a strictly proper scoring rule is the squared, or primitive, distance measure $\Delta^2(\underline{r},\underline{d})$.

APPENDIX B

THE INFORMATION RATIO $IR(\underline{\pi},\underline{r},\underline{d})$: AN IMPROPER SCORING RULE

In this appendix we show that the information ratio $IR(\underline{\pi},\underline{r},\underline{d})$ is an improper scoring rule and determine the "optimal" forecast $\underline{r}^* = (r_1^*,\ldots,r_N^*)$.

When $d_j = 1$, the measure $IR(\underline{\pi},\underline{r},\underline{d})$ becomes $IR_j(\underline{\pi},\underline{r})$, where, from (6-31),

$$IR_j(\underline{\pi},\underline{r}) = 1 - (\ln r_j/\ln \pi_j). \qquad \text{(B-1)}$$

If we let $IR(\underline{\pi},\underline{r},\underline{p})$ denote the forecaster's (subjective) expected score, then

$$IR(\underline{\pi},\underline{r},\underline{p}) = \sum_{j=1}^{N} p_j[1 - (\ln r_j/\ln \pi_j),$$

or

$$IR(\underline{\pi},\underline{r},\underline{p}) = 1 - \sum_{j=1}^{N} p_j(\ln r_j/\ln \pi_j). \qquad \text{(B-2)}$$

Now, let $F(\underline{r})$ denote $IR(\underline{\pi},\underline{r},\underline{p})$ in (B-2), let

$$G(\underline{r}) = \sum_{n=1}^{N} r_n - 1,$$

and let

$$H(\underline{r}) = F(\underline{r}) + \lambda G(\underline{r}), \qquad \text{(B-3)}$$

where λ is a Lagrange multiplier. Then, maximizing $H(\underline{r})$ in (B-3) is equivalent to maximizing the expected score $IR(\underline{\pi},\underline{r},\underline{p})$ in (B-2). Thus, taking the derivative of $H(\underline{r})$ with respect to r_m (say) and setting this derivative equal to zero yields

$$\partial H(\underline{r})/\partial r_m = -(p_m/\ln \pi_m)(1/r_m) + \lambda,$$

or

$$r_m^* = (1/\lambda)(p_m/\ln \pi_m). \quad (m=1,\ldots,N) \qquad \text{(B-4)}$$

Then, since $\sum_{n=1}^{N} r_n = 1$,

$$\lambda = \sum_{n=1}^{N} (p_n/\ln \pi_n).$$

Thus, r_m^* in (B-4) becomes

$$r_m^* = (p_m/\ln \pi_m)/\sum_{n=1}^{N} (p_n/\ln \pi_n). \quad (m=1,\ldots,N) \qquad \text{(B-5)}$$

Note that $r_m^* \neq p_m$ in (B-5) unless $\pi_n = 1/N$ for all n (n=1,...,N)

or unless $p_m = 0$ or 1. Thus, the measure $IR(\underline{\pi},\underline{r},\underline{d})$ is an <u>improper</u> scoring rule for individual forecasts.[1]

For a collection of forecasts, we shall adopt the same approach as that used in connection with the ratio skill score (see paragraph 6.3.1.2). The expected score on the (K+1)st occasion is $IR_{K+1}(\underline{\pi},\underline{r},\underline{p})$, where

$$\overline{IR_{K+1}(\underline{\pi},\underline{r},\underline{p})} = 1 - \sum_{k=1}^{K} \ln r_k^\circ \sum_{j=1}^{N}\{p_{j,K+1}/[\sum_{k=1}^{K} \ln \pi_k^\circ + \pi_{j,K+1}]\} - \sum_{j=1}^{N}\{p_{j,K+1} \ln r_{j,K+1}/[\sum_{k=1}^{K} \ln \pi_k^\circ + \ln \pi_{j,K+1}]\}, \quad \text{(B-6)}$$

in which r_k° and π_k° are the forecast and climatological probabilities, respectively, associated with the state that obtains on the kth occasion (k=1,...,K). The method of Lagrange multipliers yields the following result:

$$r^*_{m,K+1} = \{p_{m,K+1}/[\sum_{k=1}^{K} \ln \pi_k^\circ + \ln \pi_{m,K+1}]\}/\sum_{n=1}^{N}\{p_{n,K+1}/[\sum_{k=1}^{K} \ln \pi_k^\circ + \ln \pi_{n,K+1}]\}. \quad \text{(B-7)}$$

Note that $r^*_{m,K+1} \neq p_{m,K+1}$ unless $\pi_{n,K+1} = 1/N$ for all n (n=1,...,N) or unless $p_{m,K+1} = 0$ or 1. Thus, the measure $IR(\underline{\pi},\underline{r},\underline{d})$ is also <u>improper</u> for a collection of forecasts. However, for large collections of forecasts (K>100), $\sum_{k=1}^{K} \ln \pi_k^\circ >> \ln \pi_{n,K+1}$ for all n (assuming that $\pi_{n,K+1} \neq 0$), in which case $r^*_{m,K+1} \simeq p_{m,K+1}$ and $IR(\underline{\pi},\underline{r},\underline{d})$ is approximately strictly proper.

---

[1]The function $IR(\underline{\pi},\underline{r},\underline{p})$ is obviously strictly convex.

APPENDIX C

THE MATRIX SKILL SCORE MSS($\underline{\pi},\underline{r},\underline{d}$): A PROPER SCORING RULE

In this appendix we show that the matrix skill score MSS($\underline{\pi},\underline{r},\underline{d}$) is a proper scoring rule.

When $d_j = 1$, the measure MSS($\underline{\pi},\underline{r},\underline{d}$) becomes $MSS_j(\underline{\pi},\underline{r})$, where, from (6-36),

$$MSS_j(\underline{\pi},\underline{r}) = \phi(r_j,\pi_j)(1/\pi_j) + \sum_{\substack{n=1 \\ n\neq j}}^{N} \phi(\pi_n,r_n)[1/(1-\pi_n)]. \qquad \text{(C-1)}$$

If we let MSS($\underline{\pi},\underline{r},\underline{p}$) denote the forecaster's (subjective) expected score, then

$$MSS(\underline{\pi},\underline{r},\underline{p}) = \sum_{j=1}^{N} p_j \{\phi(r_j,\pi_j)(1/\pi_j) + \sum_{\substack{n=1 \\ n\neq j}}^{N} \phi(\pi_n,r_n)[1/(1-\pi_n)]\},$$

or

$$MSS(\underline{\pi},\underline{r},\underline{p}) = \sum_{n=1}^{N} [\phi_n(r_n,\pi_n)(p_n/\pi_n) + \phi_n(\pi_n,r_n)(\sum_{\substack{m=1 \\ m\neq n}}^{N} p_m / \sum_{\substack{m=1 \\ m\neq n}}^{N} \pi_m)]. \qquad \text{(C-2)}$$

Now, since MSS($\underline{\pi},\underline{r},\underline{p}$) in (C-2) is a discrete rather than a continuous function of $\underline{r}$, we shall proceed in the following manner. First, note that each probability r_n (n=1,...,N) appears in only two

terms in (C-2). Thus, in determining the "optimal" forecast r_n^* $(n=1,\ldots,N)$, we need only consider these two terms.[1] Suppose that $p_1 > \pi_1$. If we denote the sum of the relevant terms in (C-2) by $MSS^1(\underline{\pi},\underline{r},\underline{p})$, then

$$MSS^1(\underline{\pi},\underline{r},\underline{p}) = \phi_1(r_1,\pi_1)(p_1/\pi_1) + \phi(\pi_1,r_1)\left(\sum_{m=2}^{N} p_m\right)\Big/\left(\sum_{m=2}^{N}\pi_m\right). \tag{C-3}$$

Now, $p_1 > \pi_1$ implies that $\sum_{m=2}^{N} p_m < \sum_{m=2}^{N}\pi_m$. Thus, $MMS^1(\underline{\pi},\underline{r},\underline{p})$ is maximized by setting $r_1 > \pi_1$. Suppose that $p_1 < \pi_1$. Then, $\sum_{m=2}^{N} p_m > \sum_{m=2}^{N}\pi_m$ and $MSS^1(\underline{\pi},\underline{r},\underline{p})$ is maximized by setting $r_1 < \pi_1$. This argument can be used to show that if $p_n > (<)\ \pi_n$, then the forecaster should choose $r_n > (<)\ \pi_n$ in order to maximize his expected score $MSS^n(\underline{\pi},\underline{r},\underline{p})$ $(n=1,\ldots,N)$. Since the terms $MSS^n(\underline{\pi},\underline{r},\underline{p})$ are "independent," the forecaster maximizes his (total) expected score $MSS(\underline{\pi},\underline{r},\underline{p})$ by maximizing each of these individual terms. Thus, the scoring rule $MSS(\underline{\pi},\underline{r},\underline{d})$ is a proper, but not a strictly proper, scoring rule. The forecaster will realize the maximum expected score by setting $r_n = p_n$ for all n, but other values of the r_n may also yield this same (maximum) expected score $(n=1,\ldots,N)$.

---

[1] The r_n^* must, of course, satisfy the usual conditions on probabilities. That is, $r_n^* \geq 0$ and $\sum_n r_n^* = 1$ $(n=1,\ldots,N)$.

APPENDIX D

THE UTILITY MEASURE $U(\underline{r},\underline{d})$: A PROPER SCORING RULE

In this appendix we show that the utility measure $U(\underline{r},\underline{d})$ is a proper scoring rule.

When $d_j = 1$, the utility measure $U(\underline{r},\underline{d})$ becomes $U_j(\underline{r})$, where, from (7-1),

$$U_j(\underline{r}) = \sum_{m=1}^{M} \delta_m u_{mj}. \qquad \text{(D-1)}$$

The measure $U(\underline{r},\underline{d})$ is based upon the assumption that the decision maker selects the action a_m, (say) that maximizes his expected utility.[1] In order to emphasize this fact, we shall rewrite (D-1) as follows:

$$U_j(\underline{r}) = u_{m',j}(\underline{r} \in R_{m'}). \qquad \text{(D-2)}$$

If we let $U(\underline{r},\underline{p})$ denote the forecaster's expected score, then

---

[1] We must also assume that the forecaster's utility function is linearly related to the decision maker's utility function (i.e., to the utility matrix $\underline{U}$). This assumption may not be unreasonable in many practical situations.

$$U(\underline{r},\underline{p}) = \sum_{j=1}^{N} p_j U_j(\underline{r}),$$

or, from (D-2),

$$U(\underline{r},\underline{p}) = \sum_{j=1}^{N} p_j u_{mj}(\underline{r} \epsilon R_m), \qquad \text{(D-3)}$$

where $\underline{p} = (p_1,\ldots,p_N)$ denotes the forecaster's judgment. Suppose that $\underline{p} \epsilon R_{m^*}$. Then,

$$U(\underline{p},\underline{p}) = \sum_{j=1}^{N} p_j u_{m^*j}(\underline{p} \epsilon R_{m^*}). \qquad \text{(D-4)}$$

Now, from (D-3) and (D-4),

$$U(\underline{r},\underline{p}) = \sum_{j=1}^{N} p_j u_{mj}(\underline{r} \epsilon R_m) \leq U(\underline{p},\underline{p}) = \sum_{j=1}^{N} p_j u_{m^*j}(\underline{p} \epsilon R_{m^*}), \qquad \text{(D-5)}$$

with equality holding in (D-5) only when $\underline{r} \epsilon R_{m^*}$. Thus, the utility measure $U(\underline{r},\underline{d})$ is a proper, but not a strictly proper, scoring rule, since any $\underline{r} \epsilon R_{m^*}$ (not only $\underline{r} = \underline{p}$) will yield the same maximum expected score. This result has also been obtained independently by Raiffa (1969) and Stael von Holstein (1970a).

APPENDIX E

THE EXPECTED-UTILITY MEASURE EU($\underline{r}$,$\underline{d}$): A "PROPER" SCORING RULE

In this appendix we show that the expected-utility measure EU($\underline{r}$,$\underline{d}$) in a special two-action, two-state decision situation is (1) a strictly proper scoring rule when the evaluator's knowledge is expressed in terms of a positive probability distribution and (2) a proper scoring rule when the evaluator's knowledge is expressed in terms of a non-negative probability distribution. We also provide a heuristic argument and some additional evidence to support CONJECTURE 7-2 which asserts that the general expected-utility measure in the M-action, N-state decision situation possesses these properties.

Consider the special two-action, two-state situation described by the key utility matrix (c) in Figure 5-3. The utility measure U($\underline{r}$,$\underline{d}$) in this situation is, from (7-8),

$$U(\underline{r},\underline{d}) = \delta_1 d_1 + \delta_2 x. \tag{E-1}$$

The expected-utility measure EU($\underline{r}$,$\underline{d}$) is, from (7-12),

$$EU(\underline{r},\underline{d}) = \int_{x\in\Theta} (\delta_1 d_1 + \delta_2 x) f(x) dx,$$

or, when Θ is the closed unit interval [0,1],

$$EU(\underline{r},\underline{d}) = d_1\int_0^{r_1} f(x)dx + \int_{r_1}^{1} xf(x)dx. \qquad \text{(E-2)}$$

If we let $EU(\underline{r},\underline{p})$ denote the forecaster's expected score, then

$$EU(\underline{r},\underline{p}) = \sum_{j=1}^{N} p_j EU_j(\underline{r}),$$

or, from (E-2),

$$EU(\underline{r},\underline{p}) = p_1[\int_0^{r_1} f(x)dx + \int_{r_1}^{1} xf(x)dx] + p_2\int_{r_1}^{1} xf(x)dx. \qquad \text{(E-3)}$$

Now, differentiating (E-3) with respect to r_1 and setting the derivative equal to zero yields

$$\partial EU(\underline{r},\underline{p})/\partial r_1 = p_1[f(r_1) - r_1 f(r_1)] - p_2 r_1 f(r_1),$$

or

$$f(r_1^*)(p_1 - r_1^*) = 0. \qquad \text{(E-4)}$$

If $f(x)$ is a positive function, then $f(x) > 0$ and, from (E-4),

$$r_1^* = p_1.$$

Thus, the measure EU($\underline{r}$,$\underline{d}$) in (E-2) is a strictly proper scoring rule when $f(x) > 0$. On the other hand, if $f(x)$ is (only) a non-negative function, then $f(x) \geq 0$ and r_1^* need not necessarily be set equal to p_1 to maximize the expected score EU($\underline{r}$,$\underline{p}$). However, note that the forecaster can still achieve the maximum expected score by setting $r_1^* = p_1$ (i.e., other values of r_1 may yield the same expected score but no other value of r_1 will yield a greater expected score). Thus, EU($\underline{r}$,$\underline{d}$) is a proper scoring rule when $f(x) \geq 0$. The expected-utility measures associated with other special decision situations and other measures of knowledge have also been shown to be either strictly proper or proper (see Murphy, 1969a,b; Stael von Holstein, 1970b).

With regard to whether or not the expected-utility measure associated with the general M-action, N-state situation is a strictly proper (or proper) scoring rule, we offer a partly heuristic argument based largely on the work of Savage (1971). Savage (op. cit., p. 799) makes the following comments:[1]

> "(E)very proper [strictly proper] scoring rule can be viewed as a share in a business and ... every such share leads to an at least weakly proper [proper] scoring rule ... . (A) person knowing that an act in a specified economic situation

---

[1] The words and phrases in square brackets are due to the author of this thesis (AHM).

> is to be chosen for him in accordance with his announced system of personal probabilities will have no incentive to announce a false one [i.e., the "economic situation" is a proper scoring rule]. However, if the convex function arising from the family of acts has flat places - technically, is not strictly convex - then the person has no positive incentive to distinguish among certain systems of probabilities, and the scoring rule is only weakly proper [proper]."

The last sentence in the above quotation and other comments in this paper (Savage, op. cit.) indicate that a strictly convex function (arising from a family of acts) would lead to a strictly proper scoring rule. In this regard, the general utility measure $U(\underline{r},\underline{d})$ in (7-1) is a convex function. Further, since the expected-utility measure $EU(\underline{r},\underline{d})$ in (7-3) is simply the integral of $U(\underline{r},\underline{d})$ with respect to the elements of the utility matrix $\underline{U}$, we speculate that $EU(\underline{r},\underline{d})$ is a strictly convex function if $f(\underline{U})$ is positive and a convex function if $f(\underline{U})$ is non-negative. This speculation, then, leads to the conclusion that the measure $EU(\underline{r},\underline{d})$ possesses the properties described in CONJECTURE 7-2.

APPENDIX F

THE EXPECTED-UTILITY MEASURE EU($\underline{r}$,$\underline{d}$) IN A GENERAL TWO-ACTION, TWO-STATE DECISION SITUATION WITH A BOUNDED INTERVAL - UNIFORM MEASURE ON THE UTILITIES X_1 AND X_2

In this appendix we describe the expected-utility measure EU($\underline{r}$,$\underline{d}$) in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2 when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of a bounded interval - uniform measure [see (5-5)].

The utility measure U($\underline{r}$,$\underline{d}$) in this situation can be expressed as follows:

$$U(\underline{r},\underline{d}) = \delta_1 d_1 + \delta_2(d_1 x_1 + d_2 x_2). \qquad \text{(F-1)}$$

Then, the expected-utility measure is EU($\underline{r}$,$\underline{d}$), where

$$EU(\underline{r},\underline{d}) = \int\int_{(x_1,x_2)\in\Theta} [\delta_1 d_1 + \delta_2(d_1 x_1 + d_2 x_2)] f(x_1,x_2)\,dx_1\,dx_2,$$

or

$$EU(\underline{r},\underline{d}) = \{1/[(\beta_1-\alpha_1)(\beta_2-\alpha_2)]\} \int\int_{(x_1,x_2)\in\Theta} [\delta_1 d_1 + \delta_2(d_1 x_1 + d_2 x_2)]\,dx_1\,dx_2, \qquad \text{(F-2)}$$

where Θ represents the two-dimensional rectangular region for which $\alpha_1 \leq x_1 \leq \beta_1$ and $\alpha_2 \leq x_2 \leq \beta_2$ $(0 \leq \alpha_1 \leq \beta_1 \leq 1, 0 \leq \alpha_2 \leq \beta_2 \leq 1)$. When the limits of integration in (F-2) are considered, the following six cases can be identified:

(1) Case 1: $1 - (r_2/r_1)\alpha_2 \leq \alpha_1$

(2) Case 2: $\alpha_1 \leq 1 - (r_2/r_1)\alpha_2 \leq \beta_1$ and $1 - (r_2/r_1)\beta_2 \leq \alpha_1$

(3) Case 3: $\alpha_1 \leq 1 - (r_2/r_1)\alpha_2 \leq \beta_1$ and $\alpha_1 \leq 1 - (r_2/r_1)\beta_2 \leq \beta_1$

(4) Case 4: $\beta_1 \leq 1 - (r_2/r_1)\alpha_2$ and $1 - (r_2/r_1)\beta_2 \leq \alpha_1$

(5) Case 5: $\beta_1 \leq 1 - (r_2/r_1)\alpha_2$ and $\alpha_1 \leq 1 - (r_2/r_1)\beta_2 \leq \beta_1$

(6) Case 6: $\beta_1 \leq 1 - (r_2/r_1)\beta_2$

For illustrative purposes, the terms and limits of integration in (F-2) are indicated below for Case 2:

$$EU(\underline{r},\underline{d}) = \{1/[(\beta_1-\alpha_1)(\beta_2-\alpha_2)]\}\Big[\int_{\alpha_2}^{u} d_1\Big(\int_{\alpha_1}^{v} dx_1\Big)dx_2 + \int_{\alpha_2}^{u}\int_{v}^{\beta_1}(d_1x_1+d_2x_2)dx_1dx_2 + \int_{u}^{\beta_2}\int_{\alpha_1}^{\beta_1}(d_1x_1+d_2x_2)dx_1dx_2\Big], \tag{F-3}$$

where $u = (r_1/r_2)(1 - \alpha_1)$ and $v = 1 - (r_2/r_1)x_2$. The specific expressions for the expected-utility measure $EU(\underline{r},\underline{d})$ in all six of these cases are presented in Table F-1.

Table F-1. The expected-utility measure EU($\underline{r}$,$\underline{d}$) in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2 when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of a bounded interval[1]- uniform measure.

Case	EU($\underline{r}$,$\underline{d}$)
1	$d_1\xi_1 + d_2\xi_2$
2	$(d_1/\Delta_1\Delta_2)\{(\rho/3)(1-\alpha_1)^3+(1/6\rho^2)\alpha_2^3+\xi_1\Delta_1\beta_2-\alpha_2[(1/2)-\alpha_1+(1/2)\beta_1^2]\} + (d_2/\Delta_1\Delta_2)[(-\rho^2/6)(1-\alpha_1)^3-(1/3\rho)\alpha_2^3+(1/2)\alpha_2^2(1-\beta_1)+(1/2)\Delta_1\beta_2^2]$
3	$(d_1/\Delta_1)\{(-1/6\rho^2)(\alpha_2^2+\alpha_2\beta_2+\beta_2^2)+[(1/2)-\alpha_1+(1/2)\beta_1^2]\} + (d_2/\Delta_1)[(1/3\rho)(\alpha_2^2+\alpha_2\beta_2+\beta_2^2)-\xi_2(1-\beta_1)]$
4	$(d_1/\Delta_2)\{(\xi_1\beta_2-\alpha_2)+\rho[1-2\xi_1+(1/3)(\alpha_1^2+\alpha_1\beta_1+\beta_1^2]\} + (d_2/\Delta_2)\{(-\rho^2/2)[1-2\xi_1+(1/3)(\alpha_1^2+\alpha_1\beta_1+\beta_1^2)]+(1/2)\beta_2^2\}$
5	$(d_1/\Delta_1\Delta_2)\{(-\rho/3)(1-\beta_1)^3-(1/6\rho^2)\beta_2^3-\Delta_1\alpha_2+\beta_2[(1/2)-\alpha_1+(1/2)\beta_1^2]\} + (d_2/\Delta_1\Delta_2)[(\rho^2/6)(1-\beta_1)^3+(1/3\rho)\beta_2^3-(1/2)\beta_2^2(1-\beta_1)]$
6	d_1

Key

$\xi_i = (\alpha_i+\beta_i)/2 \quad (i=1,2), \quad \Delta_i = \beta_i - \alpha_i \quad (i=1,2), \quad \rho = r_1/r_2$

APPENDIX G

THE EXPECTED-UTILITY MEASURE EU($\underline{r}$,$\underline{d}$) IN A GENERAL TWO-ACTION, TWO-STATE DECISION SITUATION WITH INDEPENDENT BETA MEASURES ON THE UTILITIES X_1 AND X_2

In this appendix we describe the expected-utility measure EU($\underline{r}$,$\underline{d}$) in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2 when the evaluator's knowledge of the utilities x_1 and x_2 is expressed in terms of a beta measure [see (5-4)].

The utility measure U($\underline{r}$,$\underline{d}$) in this situation can be expressed as follows:

$$U(\underline{r},\underline{d}) = \delta_1 d_1 + \delta_2(d_1 x_1 + d_2 x_2). \tag{G-1}$$

Then, the expected-utility measure is EU($\underline{r}$,$\underline{d}$), where

$$EU(\underline{r},\underline{d}) = \int\int_{(x_1,x_2)} [\delta_1 d_1 + \delta_2(d_1 x_1 + d_2 x_2)] f_1(x_1) f_2(x_2) dx_1 dx_2, \tag{G-2}$$

in which

$$f_i(x_i) = [(a_i+b_i-1)!/(a_i-1)!(b_i-1)!] x_i^{a_i-1} (1-x_i)^{b_i-1} \tag{G-3}$$

$$0<x_i<1 \text{ and } a_i, b_i>0$$

(i=1,2). When the limits of integration in (G-2) are considered, two cases can be identified: (1) Case 1: $0 \leq r_1 \leq 1/2$ and (2) Case 2: $1/2 \leq r_1 \leq 1$. The integral expressions for EU($\underline{r}$,$\underline{d}$) in these cases are presented in Table G-1.

Table G-1. The integral expressions for the expected-utility measure $EU(\underline{r},\underline{d})$ in the general two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2 with independent beta measures on the utilities x_1 and x_2.

<u>Case 1</u>: $0 \leq r_1 \leq 1/2$

$$EU(\underline{r},\underline{d}) = d_1 \int_0^1 f_1(x_1)\left[\int_0^u f_2(x_2)dx_2\right]dx_1$$
$$+ d_1\{E(x_1) - \int_0^1 x_1 f_1(x_1)\left[\int_0^u f_2(x_2)dx_2\right]dx_1\}$$
$$+ d_2\{E(x_2) - \int_0^1 f_1(x_1)\left[\int_0^u x_2 f_2(x_2)dx_2\right]dx_1\}$$

<u>Case 2</u>: $1/2 \leq r_1 \leq 1$

$$EU(\underline{r},\underline{d}) = d_1 \int_0^1 f_2(x_2)\left[\int_0^v f_1(x_1)dx_1\right]dx_2$$
$$+ d_1\{E(x_1) - \int_0^1 f_2(x_2)\left[\int_0^v x_1 f_1(x_1)dx_1\right]dx_2\}$$
$$+ d_2\{E(x_2) - \int_0^1 x_2 f_2(x_2)\left[\int_0^v f_1(x_1)dx_1\right]dx_2\}$$

<u>Key</u>

$$E(x_i) = a_i/(a_i + b_i) \quad (i=1,2)$$
$$u = (r_1/r_2)(1 - x_1)$$
$$v = 1 - (r_2/r_1)x_2$$

The specific expressions for the expected-utility measure $EU(\underline{r},\underline{d})$ in these two cases, when the parameters a_i and b_i $(i=1,2)$ are assumed to be positive integers (see paragraph 5.3.2.3), are presented in Table G-2.

Table G-2(a). The specific expression for the expected-utility measure $EU(\underline{r},\underline{d})$ in the two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2 with independent beta measures on the utilities x_1 and x_2 assuming positive integer parameter values. Case 1: $0 \leq r_1 \leq 1/2$.

$$EU(r,d) = d_1E(x_1) + d_2E(x_2)$$

$$+[(c_2-1)!(c_1-1)!/(a_1-1)!(b_1-1)!](1/r)^{b_1}$$

$$\cdot\{d_1 \sum_{i=a_2}^{c_2-1}(1/i!)\sum_{j=0}^{a_1-1}\binom{a_1-1}{j}(-1/r)^{a_1-1-j}$$

$$\cdot(c_1-2+i-j)!\sum_{k=c_1-1+i-j}^{c-2-j}[1/k!(c-2-j-k)!]r^k(1-r)^{c-2-j-k}$$

$$- d_1 \sum_{i=a_2}^{c_2-1}(1/i!)\sum_{j=0}^{a_1}\binom{a_1}{j}(-1/r)^{a_1-j}(c_1-1+i-j)!$$

$$\cdot\sum_{k=c_1+i-j}^{c-1-j}[1/k!(c-1-j-k)!]r^k(1-r)^{c-1-j-k}$$

$$- d_2a_2 \sum_{i=a_2+1}^{c_2}(1/i!)\sum_{j=0}^{a_1-1}\binom{a_1-1}{j}(-1/r)^{a_1-1-j}$$

$$\cdot(c_1-2+i-j)!\sum_{k=c_1-1+i-j}^{c-1-j}[1/k!(c-1-j-k)!]r^k(1-r)^{c-1-j-k}\}$$

Key

$$E(x_i) = a_i/(a_i+b_i) \quad (i=1,2)$$

$$c_i = a_i + b_i \quad (i=1,2)$$

$$c = c_1 + c_2$$

$$r = (r_1/r_2)$$

Table G-2(b). The specific expression for the expected-utility measure $EU(\underline{r},\underline{d})$ in the two-action, two-state decision situation described by the key utility matrix (e) in Figure 5-2 with independent beta measures on the utilities x_1 and x_2 assuming positive integer parameter values.[1] Case 2: $1/2 \leq r_1 \leq 1$.

$$EU(\underline{r},\underline{d}) = d_1E(x_1) + d_2E(x_2)$$

$$+[(c_1-1)!(c_2-1)!/(a_2-1)!(b_2-1)!]r^{a_2}$$

$$\cdot\{d_1 \sum_{i=a_1}^{c_1-1}[1/(c_1-1-i)!]\sum_{j=0}^{b_2-1}\binom{b_2-1}{j}(-r)^{b_2-1-j}$$

$$\cdot(c-3-i-j)!\sum_{k=c-2-i-j}^{c-2-j}[1/k!(c-2-j-k)!](1/r)^k[1-(1/r)]^{c-2-j-k}$$

$$- d_1a_1 \sum_{i=a_1+1}^{c_1}[1/(c_1-i)!]\sum_{j=0}^{b_2-1}\binom{b_2-1}{j}(-r)^{b_2-1-j}(c-2-i-j)!$$

$$\cdot\sum_{k=c-1-i-j}^{c-1-j}[1/k!(c-1-j-k)!](1/r)^k[1-(1/r)]^{c-1-j-k}$$

$$- d_2r\sum_{i=a_1}^{c_1-1}[1/(c_1-1-i)!]\sum_{j=0}^{b_2-1}\binom{b_2-1}{j}(-r)^{b_2-1-j}$$

$$\cdot(c-2-i-j)!\sum_{k=c-1-i-j}^{c-1-j}[1/k!(c-1-j-k)!](1/r)^k[1-(1/r)]^{c-1-j-k}\}$$

Key

$E(x_i) = a_i/(a_i+b_i) \quad (i=1,2)$

$c_i = a_i + b_i \quad (i=1,2)$

$c = c_1 + c_2$

$r = (r_1/r_2)$

APPENDIX H

THE CONDITIONS UNDER WHICH THE EXPECTED-UTILITY MEASURE EU($\underline{r}$,$\underline{d}$) IS LINEARLY RELATED TO THE DISTANCE MEASURE $\Delta^2(\underline{r},\underline{d})$ IN A SPECIAL TWO-ACTION, TWO-STATE DECISION SITUATION

In this appendix we show that the expected-utility measure EU($\underline{r}$,$\underline{d}$) is linearly related to the distance measure $\Delta^2(\underline{r},\underline{d})$ in the special two-action, two-state decision situation described by the key utility matrix (c) in Figure 5-3 only when the utility x possesses a uniform probability distribution.

The utility measure U($\underline{r}$,$\underline{d}$) in this situation is expressed as follows:

$$U(\underline{r},\underline{d}) = \delta_1 d_1 + \delta_2 x. \qquad \text{(H-1)}$$

[see (7-8)]. The expected-utility measure EU($\underline{r}$,$\underline{d}$) in this situation, when the evaluator's knowledge of the utility x is expressed in terms of a probability distribution f(x) ($0 \leq x \leq 1$), is, from (H-1),

$$EU(\underline{r},\underline{d}) = \int_{x\varepsilon\Theta} (\delta_1 d_1 + \delta_2 x) f(x)dx,$$

or

$$EU(\underline{r},\underline{d}) = d_1 \int_0^{r_1} f(x)dx + \int_{r_1}^{1} xf(x)dx. \qquad \text{(H-2)}$$

Now, the measure $EU(\underline{r},\underline{d})$ is a linear function of the distance measure $\Delta^2(\underline{r},\underline{d})$ if and only if

$$EU(\underline{r},\underline{d}) = a\Delta^2(\underline{r},\underline{d}) + b,$$

where a and b are constants. Then, from (6-3) and (H-2),

$$d_1\int_0^{r_1} f(x)dx + \int_{r_1}^{1} xf(x)dx = 2a(r_1 - d_1)^2 + b. \qquad \text{(H-3)}$$

Now, differentiating (H-3) with respect to r_1 yields

$$d_1 f(r_1) - r_1 f(r_1) = 4a(r_1 - d_1),$$

or

$$f(r_1) = -4a. \qquad \text{(H-4)}$$

Since a is a constant, (H-4) indicates that the expected-utility measure $EU(\underline{r},\underline{d})$ is linearly related to the distance measure $\Delta^2(\underline{r},\underline{d})$ only when the utility x possesses a uniform probability distribution. Further, since $0 \leq x \leq 1$, $f(x) = 1$ and $a = -1/4$.

BIBLIOGRAPHY

Aczel, J., and J. Pfanzagl, 1966: Remarks on the measurement of subjective probability and information. Metrika, 11, 91-105.

Aitcheson, J., 1970: Choice Against Chance: An Introduction to Statistical Decision Theory. Reading, Mass., Addison-Wesley.

Allen, R. A., and E. M. Vernon, 1951: Objective weather forecasting. Compendium of Meteorology. Boston, Mass., American Meteorological Society, pp. 796-801.

Alpert, M., and H. Raiffa, 1969: A progress report on the training of probability assessors. Cambridge, Mass., Harvard University, unpublished manuscript, 31 pp.

American Meteorological Society, 1968: Proceedings of the First National Conference on Statistical Meteorology (Hartford, 1968). Boston, Mass., American Meteorological Society, 179 pp.

——, 1971: Preprint Volume, International Symposium on Probability and Statistics in Atmospheric Sciences (Honolulu, 1971). Boston, Mass., American Meteorological Society, 180 pp.

——, 1973: Preprint Volume, Third Conference on Probability and Statistics in Atmospheric Science (Boulder, 1973). Boston, Mass., American Meteorological Society, 276 pp.

Bartos, J. A., 1969: The assessment of probability distributions for future security prices. Bloomington, Indiana University, Graduate School of Business, Ph.D. Dissertation.

Besson, L., 1905: Essai de prevision methodique de temps (Attempts at methodical weather forecasting). Observatoire Municipal de Montsouris (Paris), Annales, 6, 473-495.

Blair, T. A., 1921: Local forecast studies - summer rainfall. Monthly Weather Review, 49, 183-190.

——, 1924: Local forecast studies - winter precipitation. Monthly Weather Review, 52, 79-85.

Bocchieri, J. R., and H. R. Glahn, 1973: Computer prediction of the probability of snow. Weatherwise, 26, 264-269.

Brelsford, W. H., and R. H. Jones, 1967: Estimating probabilities. Monthly Weather Review, 95, 750-756.

Brier, G. W., 1944: Verification of a forecaster's confidence and the use of probability statements in forecasting. Washington, D.C., U.S. Weather Bureau, Research Paper No. 16, 10 pp.

——, 1946: A study of quantitative precipitation forecasting in the TVA basin. Washington, D.C., U.S. Weather Bureau, Research Paper No. 26, 40 pp.

——, 1950: Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78, 1-3.

——, 1957: Effect of errors in estimating probabilities on the usefulness of probability forecasts. Bulletin of the American Meteorological Society, 38, 76-78.

——, and R. A. Allen, 1951: Verification of weather forecasts. Compendium of Meteorology. Boston, Mass., American Meteorological Society, pp. 841-848.

——, 1958: Forecast verification. In Some Applications of Statistics to Meteorology (H. A. Panofsky and G. W. Brier). University Park, Pennsylvania State University, pp. 191-208.

Bross, I. D. J., 1953: Design for Decision. New York, N.Y., The MacMillan Co., 276 pp.

Brown, T. A., 1970: Probabilistic forecasts and reproducing scoring systems. Santa Monica, Calif., The Rand Corporation, RM-6299-ARPA, 65 pp.

——, 1973: An experiment in probabilistic forecasting. Santa Monica, Calif., Rand Corporation, R-444-ARPA, 52 pp.

Bryan, J. G., 1968: Development of a computational technique for multi-group probability forecasting using a linear exponential-quotient (LEQ) model. Proceedings of the First National Conference on Statistical Meteorology (Hartford, 1968). Boston, Mass., American Meteorological Society, pp. 61-75.

Cooke, W. E., 1906a: Forecasts and verifications in Western Australia. Monthly Weather Review, 34, 23-24.

——, 1906b: Weighting forecasts. Monthly Weather Review, 34, 274-275.

Derouin, R. G., and G. F. Cobb, 1970: Weather Bureau forecast verification scores 1968-69 and some performance trends from 1966. Silver Spring, Md., ESSA, Weather Bureau, Technical Memorandum WBTM FCST-15, 50 pp.

——, and ——, 1972: National Weather Service May 1970 - April 1971 public forecast verification summary. Silver Spring, Md., NOAA, National Weather Service, Technical Memorandum NWS FCST-17, 92 pp.

Dickey, W. W., 1949: Estimating the probability of a large fall in temperature at Washington, D.C. Monthly Weather Review, 77, 67-78.

Dobryshman, E. M., 1972: Review of forecast verification techniques. Geneva, Switzerland, World Meteorological Organization, Technical Note No. 120, 51 pp.

Edwards, W., 1954: The theory of decision making. Psychological Bulletin, 51, 380-417.

Epstein, E. S., 1962: A Bayesian approach to decision making in applied meteorology. Journal of Applied Meteorology, 1, 169-177.

——, 1966: Quality control for probability forecasts. Monthly Weather Review, 94, 487-494.

——, 1969: A scoring system for probabilities of ranked categories. Journal of Applied Meteorology, 8, 985-987.

——, and A. H. Murphy, 1965: A note on the attributes of probabilistic predictions and the probability score. Journal of Applied Meteorology, 4, 297-299.

Ferguson, T. S., 1967: Mathematical Statistics: A Decision Theoretic Approach. New York, N.Y., Academic Press, 396 pp.

de Finetti, B., 1937: La prevision: ses lois logiques, ses sources subjectives. Annales de Institut Henri Poincare, 7, 1-68.

——, 1962: Does it make sense to speak of 'good probability appraisers'? In The Scientist Speculates: An Anthology of Partly-Baked Ideas (I.J. Good, Editor). New York, N.Y., Basic Books, pp. 357-364.

——, 1965: Methods for discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 18, 87-123.

Fishburn, P. C., 1964: Decision and Value Theory. New York, N.Y., John Wiley and Sons, 451 pp.

Gentry, R. C., 1950: Forecasting local showers in Florida during the summer. Monthly Weather Review, 78, 41-49.

Glahn, H. R., 1964: The use of decision theory in meteorology. Monthly Weather Review, 92, 383-388.

——, 1965: Objective weather forecasting by statistical methods. The Statistician, 15, 111-142.

——, 1974: Problems in the use of probability forecasts. Preprint Volume, Fifth Conference on Weather Forecasting and Analysis (St. Louis, 1974). Boston, Mass., American Meteorological Society, pp. 32-35.

——, and D. L. Jorgenson, 1970: Climatological aspects of the Brier P-score. Monthly Weather Review, 98, 136-141.

Good, I. J., 1952: Rational decisions. Journal of the Royal Statistical Society, Series B, 14, 107-114.

——, 1959: Kinds of probability. Science, 129, 443-447.

Goodman, L. A., and W. H. Kruskal, 1954: Measures of association for cross classifications. Journal of the American Statistical Association, 49, 732-764.

——, and ——, 1959: Measures of association for cross classifications. II: Further discussion and references. Journal of the American Statistical Association, 54, 123-163.

Grayson, C. J., 1960: Decisions under Uncertainty: Drilling Decisions by Oil and Gas Operators. Boston, Mass., Harvard University Press.

Gringorten, I. I., 1950: Forecasting by statistical inferences. Journal of Meteorology, 7, 388-394.

——, 1951: The verification and scoring of weather forecasts. Journal of the American Statistical Association, 46, 279-296.

——, 1952: An objective aid to local forecasting. Bulletin of the American Meteorological Society, 33, 202-207.

——, 1955: Methods of objective weather forecasting. Advances in Geophysics, Volume 2. New York, N.Y., Academic Press, pp. 57-92.

——, 1958: On the comparison of one or more sets of probability forecasts. Journal of Meteorology, 15, 283-287.

——, 1959: Probability estimates of the weather in relation to operational decisions. Journal of Meteorology, 16, 663-671.

——, 1965: A measure of skill in forecasting a continuous variable. Journal of Applied Meteorology, 4, 47-53.

——, 1967: Verification to determine and measure forecasting skill. Journal of Applied Meteorology, 6, 742-747.

Hallenbeck, C., 1920: Forecasting precipitation in percentages of probability. Monthly Weather Review, 48, 645-647.

Hampton, J. M., P. G. Moore, and H. Thomas, 1973: Subjective probability and its measurement. Journal of the Royal Statistical Society, Series A, 136, 21-42.

Hendrickson, A. D., and R. J. Buehler, 1971: Proper scores for probability forecasters. Annals of Mathematical Statistics, 42, 1916-1921.

Hoffman, J., and C. R. Peterson, 1972: A scoring rule to train probability assessors. Ann Arbor, University of Michigan, Engineering Psychology Laboratory, Technical Report 037230-4-T, 14 pp.

Holloway, J. L., and M. A. Woodbury, 1955: Application of information theory and discriminant function analysis to weather forecasting and forecast verification. Philadelphia, University of Pennsylvania, Institute for Cooperative Research, Technical Report No. 1 [Contract Nonr 551(07)], 85 pp.

Hughes, L. A., 1965: On the probability forecasting of the occurrence of precipitation. Kansas City, Mo., Environmental Science Services Administration, Weather Bureau, Technical Note 20-CR-3, 36 pp.

——, 1967: Probability verification results (6-month and 18-month). Kansas City, Mo., ESSA, Weather Bureau, Technical Memorandum WBTM CR-17, 24 pp.

——, 1968: Probability verification results (24 months). Kansas City, Mo., ESSA, Weather Bureau, Technical Memorandum WBTM CR-19, 29 pp.

Johnson, D. H., 1957: Forecast verification: a critical survey of the literature. London, England, Air Ministry, Meteorological Research Committee, M.R.P 1956, S.C. 11/237, 40 pp.

Jorgenson, D. L., 1953: Estimating precipitation at San Francisco from concurrent meteorological variables. Monthly Weather Review, 81, 101-110.

Kelly, C. W., and C. R. Peterson, 1971: Probability estimates and probabilistic procedures in current-intelligence analysis. Gaithersburg, Md., International Business Machines Corporation, Federal Systems Division, Report on Phase I (FSC71-5047), 76 pp.

Khinchin, A. I., 1957: Mathematical Foundations of Information Theory. New York, N.Y., Dover Publications, Inc., 120 pp.

Klein, W. H., 1970: The forecast research program of the Techniques Development Laboratory. Bulletin of the American Meteorological Society, 51, 133-142.

——, and H. R. Glahn, 1974: Forecasting local weather by means of model output statistics. Bulletin of the American Meteorological Society, 55, 1217-1227.

Lindley, D. V., 1971: Making Decisions. New York, N.Y., Wiley-Interscience, 195 pp.

Luce, R. D., and H. Raiffa, 1957: Games and Decisions. New York, N.Y., John Wiley and Sons, 509 pp.

Lund, I. A., 1955: Estimating the probability of a future event from dichotomously classified predictors. Bulletin of the American Meteorological Society, 36, 325-328.

Malone, T. F., 1957: Applied meteorology. Meteorological Monographs, 3, No. 16, pp. 152-159.

McCarthy, J., 1956: Measures of the value of information. Proceedings of the National Academy of Sciences, 42, 654-655.

Meglis, A. J., 1960: Annotated bibliography on forecast verification. Meteorological and Geoastrophysical Abstracts, 11, 1129-1174.

Miller, R. G., 1962: Statistical prediction by discriminant analysis. Meteorological Monographs, 4, No. 25, 54 pp.

——, 1964: Regression estimation of event probabilities. Hartford, Conn., Travelers Research Center, Inc., Technical Report No. 1 (Contract No. Cwb-10704), 153 pp.

Murphy, A. H., 1966: A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio decision situation. Journal of Applied Meteorology, 5, 534-537.

——, 1969a: Measures of the utility of probabilistic predictions in cost-loss ratio decision situations in which knowledge of the cost-loss ratios is incomplete. Journal of Applied Meteorology, 8, 863-873.

——, 1969b: On the "ranked probability score." Journal of Applied Meteorology, 8, 988-989.

——, 1970: The ranked probability score and the probability score: a comparison. Monthly Weather Review, 98, 917-924.

——, 1971: A note on the ranked probability score. Journal of Applied Meteorology, 10, 155-156.

——, 1972a: Scalar and vector partitions of the probability score: part I. Two-state situation. Journal of Applied Meteorology, 11, 273-282.

——, 1972b: Scalar and vector partitions of the probability score: part II. N-state situation. Journal of Applied Meteorology, 11, 1183-1192.

——, 1972c: Scalar and vector partitions of the ranked probability score. Monthly Weather Review, 100, 701-708.

——, 1972d: Ordinal relationships between measures of the "accuracy" and "value" of probability forecasts: preliminary results. Tellus, 24, 531-542.

——, 1972e: Probability forecasting in meteorology: a review of recent developments. Boulder, Colo., National Center for Atmospheric Research, unpublished manuscript, 19 pp.

——, 1973a: Hedging and skill scores for probability forecasts. Journal of Applied Meteorology, 12, 215-233.

——, 1973b: A new vector partition of the probability score. Journal of Applied Meteorology, 12, 595-600.

——, 1974: A sample skill score for probability forecasts. Monthly Weather Review, 102, 48-55.

——, and R. A. Allen, 1970: Probabilistic prediction in meteorology: a bibliography. Silver Spring, Md., ESSA, Weather Bureau, Technical Memorandum WBTM-35, 60 pp.

——, and E. S. Epstein, 1967a: Verification of probabilistic predictions: a brief review. Journal of Applied Meteorology, 6, 748-755.

——, and ——, 1967b: A note on probability forecasts and "hedging." Journal of Applied Meteorology, 6, 1002-1004.

——, and C.-A. Stael von Holstein, 1974: The family of quadratic scoring rules. Boulder, Colo., National Center for Atmospheric Research, and Stockholm, Sweden, Stanford Research Institute - Scandinavia, unpublished manuscript, in preparation.

——, and ——, 1975: A geometrical framework for the ranked probability score. Monthly Weather Review, 103, in press.

——, and R. L. Winker, 1970: Scoring rules in probability assessment and evaluation. Acta Psychologica, 34, 273-286.

——, and ——, 1971: Forecasters and probability forecasts: some current problems. Bulletin of the American Meteorological Society, 52, 239-247.

——, and ——, 1974a: Subjective probability forecasts experiments in meteorology: some preliminary results. Bulletin of the American Meteorological Society, 55, 1206-1216.

——, and ——, 1974b: Credible interval temperature forecasting: some experimental results. Monthly Weather Review, 102, 784-794.

von Myrbach, O., 1913: Verlasslichkeit der wetterprognosen (ein vorschlag). Meteorologische Zeitschrift, 30, 496-497.

van Naerssen, R. F., 1962: A scale for the measurement of subjective probability. Acta Psychologica, 20, 159-166.

Nelson, R. R., and S. G. Winter, 1964: A case study in the economics of information and coordination: the weather forecasting system. Quarterly Journal of Economics, 78, 420-441.

Penn, S., 1948: An objective method for forecasting precipitation amounts from winter coastal storms for Boston. Monthly Weather Review, 76, 149-161.

Peterson, C. R., and L. R. Beach, 1967: Man as an intuitive statistician. Psychological Bulletin, 68, 29-46.

——, K. J. Snapper, and A. H. Murphy, 1972: Credible interval temperature forecasts. Bulletin of the American Meteorological Society, 53, 966-970.

Pontryagin, L. S., 1952: Foundations of Combinatorial Topology. Rochester, N.Y., Graylock Press, 99 pp.

Raiffa, H., 1968: Decision Analysis. Reading, Mass., Addison Wesley, 309 pp.

——, 1969: Assessment of probabilities. Cambridge, Mass., Harvard University, School of Business, unpublished manuscript, 23 pp.

——, and R. Schlaifer, 1961: Applied Statistical Decision Theory. Boston, Mass., Harvard University, Graduate School of Business Administration, 356 pp.

Rapaport, A., and T. S. Wallsten, 1972: Individual decision behavior. Annual Review of Psychology, 23, 131-176.

Roberts, H. V., 1968: On the meaning of the probability of rain. Proceedings of the First National Conference on Statistical Meteorology (Hartford, 1968). Boston, Mass., American Meteorological Society, pp. 133-141.

Roby, T. B., 1965: Belief states and the uses of evidence. Behavioral Science, 10, 255-270.

Rolf, B., 1917: Probabilite et pronostics des pluies d'ete. Uppsala, Sweden, University of Uppsala, Ph.D. Dissertation, 51 pp.

Root, H. E., 1962: Probability statements in weather forecasting. Journal of Applied Meteorology, 1, 163-168.

Sadowski, A. F., and G. F. Cobb, 1973: National Weather Service May 1971 - April 1972 public forecast verification summary. Silver Spring, Md., NOAA, National Weather Service, Technical Memorandum NWS FCST-19, 26 pp.

Sanders, F., 1958: The evaluation of subjective probability forecasts. Cambridge, Massachusetts Institute of Technology, Department of Meteorology, Scientific Report No. 5 [Contract No. AF19(604)-1305], 60 pp.

——, 1963: On subjective probability forecasting. Journal of Applied Meteorology, 2, 191-201.

——, F., 1973: Skill in forecasting daily temperature and precipitation. Bulletin of the American Meteorological Society, 54, 1171-1179.

Savage, L. J., 1954: The Foundations of Statistics. New York, N.Y., John Wiley and Sons, 294 pp.

——, 1971: Elicitation of personal probabilities and expectations. Journal of the American Statistical Association, 66, 783-801.

Schmidt, R. C., 1951: A method of forecasting occurrence of winter precipitation two days in advance. Monthly Weather Review, 79, 81-95.

Schmitt, S. A., 1969: Measuring Uncertainty: An Elementary Introduction to Bayesian Statistics. Reading, Mass., Addison-Wesley, 400 pp.

Shannon, C. E., and W. Weaver, 1949: The Mathematical Theory of Communication. Urbana, University of Illinois Press, 117 pp.

Shuford, E. H., 1969: Confidence testing: a new tool for measurement. Lexington, Mass., The Shuford-Massengill Corporation.

——, A. Albert, and H. E. Massengill, 1966: Admissible probability measurement procedures. Psychometrika, 31, 125-145.

——, and H. E. Massengill, 1967: How to shorten a test and increase its reliability and validity. Lexington, Mass., The Shuford-Massengill Corporation.

Slovic, P., and S. Lichtenstein, 1971: Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organizational Behavior and Human Performance, 6, 649-744.

Spetzler, C. S., and C.-A.S. Stael von Holstein, 1974: Probability encoding and decision analysis. Management Science, 21, in press.

Stael von Holstein, C.-A.S., 1970a: Assessment and Evaluation of Subjective Probability Distributions. Stockholm, Sweden, Stockholm School of Economics, Economic Research Institute, 235 pp.

——, 1970b: A family of strictly proper scoring rules which are sensitive to distance. Journal of Applied Meteorology, 9, 360-364.

——, 1971: An experiment in probabilistic weather forecasting. Journal of Applied Meteorology, 10, 635-645.

Thompson, J. C., 1950: A numerical method for forecasting rainfall in the Los Angeles area. Monthly Weather Review, 78, 113-124.

——, 1952: On the operational deficiencies in categorical weather forecasts. Bulletin of the American Meteorological Society, 33, 223-226.

——, 1972: The potential economic benefits of improvements in weather forecasting. San Jose, California State University, Department of Meteorology, Final Report (Grant No. NGR 05-046-005), 80 pp.

——, and G. W. Brier, 1955: The economic utility of weather forecasts. Monthly Weather Review, 83, 249-254.

Toda, M., 1963: Measurement of subjective probability distribution. State College, Pennsylvania State University, Institute for Research, Division of Mathematical Psychology, Report No. 3, 32 pp.

Tversky, A., and D. Kahneman, 1973: Judgment under uncertainty: heuristics and biases. Eugene, Oregon Research Institute, ORI Research Bulletin, Vol. 13, No. 1, 33 pp.

U.S. Army Signal Crops, 1919: The Signal Corps meteorological service, A.E.F. (Excerpts from the Annual Report of the Chief Signal Officer). Monthly Weather Review, 47, 870-876.

Vernon, E. M., 1947: An objective method of forecasting precipitation 24-48 hours in advance at San Francisco, California. Monthly Weather Review, 75, 211-219.

Webster's Third New International Dictionary, 1966: Webster's Third New International Dictionary of the English Language Unabridged. Springfield, Mass., G. and C. Merriam Company, 2662 pp.

Williams, P. W., 1951: The use of confidence factors in forecasting. Bulletin of the American Meteorological Society, 32, 279-281.

Winkler, R. L., 1967a: The assessment of prior distributions in Bayesian analysis. Journal of the American Statistical Association, 62, 776-800.

——, 1967b: The quantification of judgment: some methodological suggestions. Journal of the American Statistical Association, 62, 1105-1120.

——, 1969: Scoring rules and the evaluation of probability assessors. Journal of the American Statistical Association, 64, 1073-1078.

——, 1971: Probabilistic prediction: some experimental results. Journal of the American Statistical Association, 66, 675-685.

——, 1972: Introduction to Bayesian Inference and Decision. New York, N.Y., Holt, Rinehart and Winston, 563 pp.

——, and A. H. Murphy, 1968: "Good" probability assessors. Journal of Applied Meteorology, 7, 751-758.

——, and ——, 1970: Nonlinear utility and the probability score. Journal of Applied Meteorology, 9, 143-148.

——, and ——, 1973: Experiments in the laboratory and the real world. Organizational Behavior and Human Performance, 10, 252-270.

Woodall, M. P., 1967: A suggested modification of the Brier and Allen score for probability forecasts. Journal of Applied Meteorology, 6, 580.

World Meteorological Organization, 1966: Statistical analysis and prognosis in meteorology. Geneva, Switzerland, World Meteorological Organization, Technical Note No. 71, 197 pp.