

## INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the original text directly from the copy submitted. Thus, some dissertation copies are in typewriter face, while others may be from a computer printer.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyrighted material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each oversize page is available as one exposure on a standard 35 mm slide or as a 17" × 23" black and white photographic print for an additional charge.

Photographs included in the original manuscript have been reproduced xerographically in this copy. 35 mm slides or 6" × 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.



Accessing the World's Information since 1938

300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA



**Order Number 8808810**

**Probability elicitation and the formation of expectations: An  
experimental approach**

**Nelson, Robert Graham, Ph.D.**

**Texas A&M University, 1987**

**U·M·I**  
300 N. Zeeb Rd.  
Ann Arbor, MI 48106



PROBABILITY ELICITATION AND THE FORMATION OF EXPECTATIONS:  
AN EXPERIMENTAL APPROACH

A Dissertation

by

ROBERT GRAHAM NELSON

Submitted to the Graduate College of  
Texas A&M University  
in partial fulfillment of the requirements for the degree  
of

DOCTOR OF PHILOSOPHY

December 1987

Major Subject: Agricultural Economics

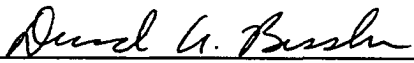
PROBABILITY ELICITATION AND THE FORMATION OF EXPECTATIONS:  
AN EXPERIMENTAL APPROACH

A Dissertation

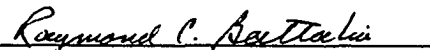
by

ROBERT GRAHAM NELSON

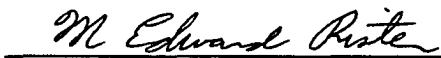
Approved as to style and content by:



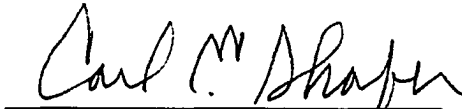
David A. Bessler  
(Chair of Committee)



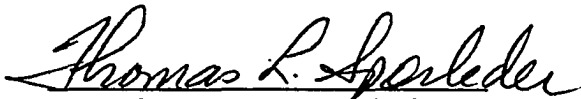
Raymond C. Battalio  
(Member)



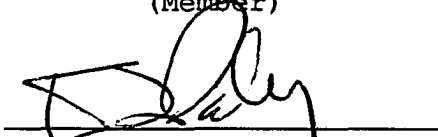
M. Edward Rister  
(Member)



Carl E. Shafer  
(Member)



Thomas L. Spörleder  
(Member)



Daniel I. Padberg  
(Head of Department)

December 1987

## ABSTRACT

Probability Elicitation and the Formation of Expectations:

An Experimental Approach. (December 1987)

Robert Graham Nelson, B.S., Oregon State University;

M.S., Auburn University

Chair of Advisory Committee: Dr. David A. Bessler

This research deals with subjective probabilities and their use in decision-making. Two theories that predict behavior in this context are tested using the methods of experimental economics.

An articulation of the theory of scoring rules leads to weak and strong predictions about behavior under an improper rule--the only kind of predictions that can be directly observed. The weak prediction was tested under controlled laboratory conditions using subjects with linear utility over the range of rewards. One-step-ahead probability forecasts were elicited from eight subjects under a proper (quadratic) scoring rule and from eight subjects under an improper (linear) scoring rule. Using the entire 40 forecast periods, the prediction that subjects under the linear rule will forecast with significantly "tighter" probability distributions was confirmed. However, there was no significant difference in

"tightness" attributable to scoring rules over the first 15 periods, suggesting that a training or feedback effect is required before the predicted behavior is manifested. It is thus possible that, for a limited number of forecasts, the linear scoring rule may be the reward mechanism of choice since it has the advantage (particularly in field elicitation studies) of being easily understood by subjects.

The theory of quasi-rational expectations was tested under controlled conditions of the economics laboratory. Five experiments were conducted with a variety of stochastic processes. In each experiment, subjects produced one-step-ahead forecasts of the variable generated by a Monte Carlo process. Comparisons of the performance of an aggregate of subjects' forecasts versus an ARIMA model showed that for relatively simple series (such as those generated by autoregressive processes of first or second order) the aggregate forecast was indistinguishable from that of the model. These results lend support to the theory that forecasts from an ARIMA model can serve as substitutes for aggregate expectations in macroeconomic policy models.



To Joan, Jenny, and Katie, and my parents.

## ACKNOWLEDGMENTS

A great number of people contributed to this effort and it is a pleasure to count them among my friends and colleagues. The arguments in the production function of this Ph.D. include both indirect and direct inputs. I hope that all those people that I fail to acknowledge by name will know that I am grateful for their support and encouragement. The cheerful comments, the shared anxieties and accomplishments, a doughnut here, a cartoon there; these things I do not forget.

Among those deserving special recognition, my chairman and mentor, Dave Bessler, stands foremost. His eclectic research interests and familiarity with frontier literature and methods made this research possible. His *laissez-faire* management style gave me the freedom to do it my way, and his model of industriousness drove me to finish it. For all this and more, I thank him sincerely.

The members of my committee were encouraging, inquisitive, challenging, and patient. For their formal role in this product I thank Ray Battalio, Ed Rister, Carl Shafer, Tom Sporleder, and my GCRs, Robert Maggio and Ed Funkhouser.

Kusumben Mistry contributed more than her usual competent programming skills. She took a genuine interest

in the experiments, and proved inventive and thorough in designing the FORECAST software.

The experimental design underwent many trials (and errors). For their help in the trials I thank my "volunteers": Mike Glover, Mary McKnight, Wayne Howard, Wes Harris, Dave Dearmont, Komain Jiranyakul, and the members of Dr. Bessler's 1986 Risk class.

Mary McKnight, Teresa Stallings, Gail Love, and H.L. Goodwin allowed me to recruit subjects from their classes and provided much encouragement. Although I cannot mention my subjects by name, I am most appreciative of their participation, patience, and willingness to accomodate me.

In addition to the guidance provided by my chairman, I had many helpful discussions with Ray Battalio, Mike Glover, Komain Jiranyakul, Pete Chamberlain, and Glenn Harrison. The research benefited substantially from their special insight.

Finally I would like to thank Sheila Hawley (particularly for expediting financial arrangements), Linda Crenwelge, Ruth Hicks of the copy center, and Sue Durden of the reference lab, all for doing their jobs with cheerful professionalism. I also appreciate the editing assistance of Nancy Wick and Rhoda Segur.

**TABLE OF CONTENTS**

	Page
CHAPTER I INTRODUCTION . . . . .	1
History of Subjective Expected Utility Theory . . .	1
Philosophy of Subjective Probability Theory . . . .	5
Application to Methods of Decision-making . . . . .	8
Role of Experimental Economics . . . . .	16
Objectives of this Research . . . . .	29
CHAPTER II SUBJECTIVE PROBABILITIES ELICITED UNDER PROPER AND IMPROPER SCORING RULES: A LABORATORY TEST OF PREDICTED RESPONSES . . . . .	31
Introduction . . . . .	31
Theory . . . . .	37
Method . . . . .	45
Background . . . . .	45
Utility Elicitation . . . . .	46
Probability Elicitation . . . . .	51
Results . . . . .	55
Conclusions . . . . .	60
CHAPTER III QUASI-RATIONAL EXPECTATIONS: EXPERIMENTAL EVIDENCE . . . . .	62
Introduction . . . . .	62

	Page
Method . . . . .	69
Monte Carlo Series Generation . . . . .	71
Optimal Statistical Model . . . . .	85
Aggregation of Subjects' Forecasts . . . . .	88
Test Criteria . . . . .	89
Results . . . . .	93
General . . . . .	93
AR1 Experiment . . . . .	95
AR2 Experiment . . . . .	97
RW Experiment . . . . .	97
AE Experiment . . . . .	100
AR4 Experiment . . . . .	100
Conclusions . . . . .	100
CHAPTER IV CONCLUSIONS . . . . .	107
REFERENCES . . . . .	110
APPENDIX A INSTRUCTIONS FOR UTILITY EXPERIMENT . . . . .	119
APPENDIX B INSTRUCTIONS FOR FORECASTING EXPERIMENT . . . . .	127
VITA . . . . .	144

**LIST OF TABLES**

TABLE	Page	
2.1	Significance Level Associated with the F Value ( $PR > F$ ) for Various Combinations of Forecast Periods in ANOVA Model of Effect of Scoring Rule on Number of Zeroes Used in a Forecast . . .	59
3.1	Specification of Monte Carlo Generator and <i>Ex Post</i> ARIMA Model of Data for AR1 Experiment. .	72
3.2	Specification of Monte Carlo Generator and <i>Ex Post</i> ARIMA Model of Data for AR2 Experiment. .	73
3.3	Specification of Monte Carlo Generator and <i>Ex Post</i> ARIMA Model of Data for RW Experiment . .	74
3.4	Specification of Monte Carlo Generator and <i>Ex Post</i> ARIMA Model of Data for AE Experiment . .	75
3.5	Specification of Monte Carlo Generator and <i>Ex Post</i> ARIMA Model of Data for AR4 Experiment. .	76
3.6	Performance of Aggregate Forecast in Five Experiments . . . . .	94
3.7	Individual Subject and Aggregate Results from AR1 Experiment. . . . .	96
3.8	Individual Subject and Aggregate Results from AR2 Experiment. . . . .	98

TABLE	Page
3.9 Individual Subject and Aggregate Results from RW Experiment . . . . .	99
3.10 Individual Subject and Aggregate Results from AE Experiment . . . . .	101
3.11 Individual Subject and Aggregate Results from AR4 Experiment . . . . .	102

## LIST OF FIGURES

FIGURE	Page
1.1 Educational Routes in Extending Methods for Decision Analysis . . . . .	11
2.1 "Historical" Series of 40 Periods Shown to Subjects Prior to Forecasting . . . . .	56
2.2 Graph of Realizations after Completion of 40 Forecast Periods . . . . .	57
3.1 Graph of Series Used in AR1 Experiment . . . . .	77
3.2 Graph of Series Used in AR2 Experiment . . . . .	78
3.3 Graph of Series Used in RW Experiment . . . . .	79
3.4 Graph of Series Used in AE Experiment . . . . .	80
3.5 Graph of Series Used in AR4 Experiment . . . . .	81



CHAPTER I  
INTRODUCTION

**History of Subjective Expected Utility Theory**

The focus of this dissertation is on the use of subjective expected utility (SEU) theory as an aid to decision-making. A brief sketch of the history of probability and utility may serve to orient this study within the larger scheme of statistics, economics, and business.

The earliest notions of probability were undoubtedly motivated by games of chance. Indeed, much of the paraphernalia of chance events--astragali (animal bones), dice, and board games--were in common use at the beginning of this millenium (David). Although randomization, chance and fortune appear to have been accepted facts of life (supporting a lively practice in the arts of divination) two factors conspired to slow the evolution of probability theory for nearly 500 years. First, the concept of "equally likely events" had to await the concept of a perfect solid, such as a cube, for which any side would have the same chance of appearing topmost. The technology

---

This dissertation follows the format and style required for publication in *The American Journal of Agricultural Economics*.

for producing such a cube was not the problem, as some beautifully crafted, true-throwing ceramic dice have been unearthed from settlements of the time. Second, while the idea of counting and enumeration was well established and indeed vital in commerce, the concept of a *number* in the modern sense was lacking. The difficulty in manipulating traditional number systems delayed the systematic examination of combinations of events until contributions from the Hindu and Arab cultures provided a working arithmetic (David).

The "prehistory" (Seneta) of probability is associated with such authors as Fra Luca Pacioli (ca. 1445-ca. 1517), Niccolo Fontana Tartaglia (ca. 1500-1557), Girolamo Cardano (1501-1576), and Galileo Galilei (1564-1642). Cardano, himself an inveterate gambler, is often credited with articulating the concept of probability in gambling. Although the originality of his contributions may be disputed, he seems to be the first author to abstract from the empiricism of dice throwing to the correct calculation of a theoretical probability, the attribution arising from this quotation in his book *Liber de Ludo Alea*: "...the wagers therefore are laid in accordance with this equality [of chances] if the die is honest." (David, p.58)

Seneta credits the beginning of the "history" of probability to the correspondence between Blaise Pascal (1623-1662) and Pierre de Fermat (1601-1665). While their

principal motivation was the correct placement of wagers in games of chance, the prestige of these two men propelled the scholarly development of the theory. Among other names of this time, Christiaan Huygens first proposed the principle of mathematical expectation, which served as the cornerstone of the mathematics of probability and decision-making through the 18<sup>th</sup> century.

Evidence contradicting the notion that people maximize mathematical expectations was provided by James, Daniel, and Nicholas Bernoulli in their famous "St. Petersburg paradox", as well as from everyday examples such as insurance and gambling. Daniel Bernoulli (1700-1782) introduced the idea of utility as "moral worth", and the normative concept of decision-making as maximization of expected utility or "moral expectation". Pierre Simon de Laplace (1749-1827) applied the concept of probability as a "degree of belief or confidence" thereby acknowledging the inherently *subjective* nature of probability in most practical contexts. However, the use of utility and subjective probability as a way of thinking apparently influenced neither mathematical nor practical probabilists until well into the 20th century (Savage 1972).

A period of probability-less ideas about utility seems to be a legacy of Alfred Marshall (Stigler). In the 1920's there was a resurgence of interest in subjective expectations, with contributions by John Maynard Keynes and

Harold Jeffreys in England, and Frank Ramsey in the United States. It is perhaps a result of his untimely death at the age of 26 that Ramsey's decision-theoretic concepts of the duality of subjective probability and utility failed to gain the wide acceptance they deserved at the time.

The modern probability theory of utility was developed by John von Neumann and Oskar Morgenstern. Not without some historical irony, their theory was published in the appendix of the second edition (1947) of their book entitled *Theory of Games and Economic Behavior*. Although the examples used to illustrate their theory dealt only with the canonical variety of probabilities (for example, drawing balls from an urn), the authors remarked in a footnote (p. 19) that if one were to find the frequency interpretation objectionable, then probability and utility could be axiomatized together. This axiomatization was completed by Leonard Savage in his book, *Foundations of Statistics*. There Savage liberally acknowledged the influence of Bruno de Finetti's work on the personalistic view of probability.

Today there may be some argument about the impact of recent contributions to SEU theory, and who are the "fathers" of the field in such rapid growth areas as economics, statistics, psychology, and management science. For purposes of this exposition, von Neumann and Morgenstern are recognized for their complete development

of the utility hypothesis as a consequence of behavioral principles, as well as their unique application of the theory to economic behavior in the context of "games". Savage is credited with the explicit incorporation of subjective probabilities into probability theory, thus giving legitimacy to the Bayesian school of statistics and its use in individual decision-making. Finally, considerable debt is owed to pioneers in the management sciences such as Robert O. Schlaifer, Howard Raiffa, and John W. Pratt for their contributions to practical business statistics and "scientific management".

### **Philosophy of Subjective Probability Theory**

This section is developed to support the logic of SEU theory in decision-making research and to compare the merits of SEU theory with other approaches that are often applied in the literature. It attempts to answer the question: "why should we use SEU theory as an aid to decision-making instead of something else?" The discussion is taken largely from Fine's book, *Theories of Probability*.

Fine labels the major interpretations of probability as "relative frequency", "complexity", "classical", "logical", and "subjective". Among these competing schools of thought, probability theory has been variously claimed to be:

1. an assertion of a physical characteristic of an experiment that will be manifested under prescribed conditions
2. an elaboration of correct inductive reasoning concerned with assessments of degrees of truth
3. a formalization of individual opinion leading to decisions satisfactory to the individual
4. an expression of individual judgments in an interpersonal form
5. a summary description of data
6. a selection of a probabilistic automaton as a model of the data source.

Fine comments:

"The relative frequency, complexity, classical, and logical interpretations of probability are primarily concerned with knowledge and inference. None of these interpretations lend themselves to a ready justification for the use of probability to guide behavior or to facilitate decision-making." (p. 212)

"Of all the methods for decision-making, that based on subjective probability is most likely to satisfy the user. ...he is encouraged to make decisions that agree with his preferences as evaluated to the best of his personal knowledge and belief. In the absence of a unique best set

of instructions for rationally reasoning from knowledge and beliefs to preferences, self-satisfaction is the best that could be expected."

(p. 236)

If we assume that the decisionmaker knows or at least behaves by the model that "subjective probability is combined with utility to beget preferences", then what is to be gained by partitioning these components in order to analyze his decisions in this manner? Fine suggests several reasons why this approach can be helpful. First, a complicated decision problem can be broken into a set of simpler ones. This has the principle advantage of isolating (for separate analysis) the several sources of error that the decisionmaker may introduce. Second, it allows communication of personal judgments about the state of the world. Thus, even though decisionmakers may disagree among themselves or with "experts" on the consequences of various actions, this communication may provide valuable information in selecting a best action for each of them individually. Third, under ideal circumstances we know that the user is assured of self-satisfaction. Fourth, much of classical probability can be subsumed as a special case of subjective probability.

The foregoing discussion about user self-satisfaction has profound implications to the agricultural economics profession, particularly with respect to extension. There

are two ways to extend the benefits of decision analysis to farmers and agribusinessmen. One way is to attempt to convince them that the decisions they make could be improved if they used historical frequencies, "expert opinion", or Outlook forecasts in place of their own judgment. The other way is to provide a systematic method or framework for understanding how they make their decisions and let them choose if and where they want to introduce changes in their information processing methods. It seems that only the second way can secure the objective of user self-satisfaction, and that perhaps no more lofty objective than this can be amenable to the skills of our profession.

#### **Application to Methods of Decision-making**

Decision analysis under SEU theory is characterized by the following steps:

1. identify the available choices of action (e.g. plant corn, sell wheat at harvest, fertilize at pre-plant stage)
2. identify the possible states of nature or "events" (e.g. sorghum price at harvest is \$3.50/bu; no rain at critical growth stage; 200 bu/ac yield)
3. assign probabilities to the states of nature (e.g. 50% chance of cotton price between \$.45-.55/lb; .05



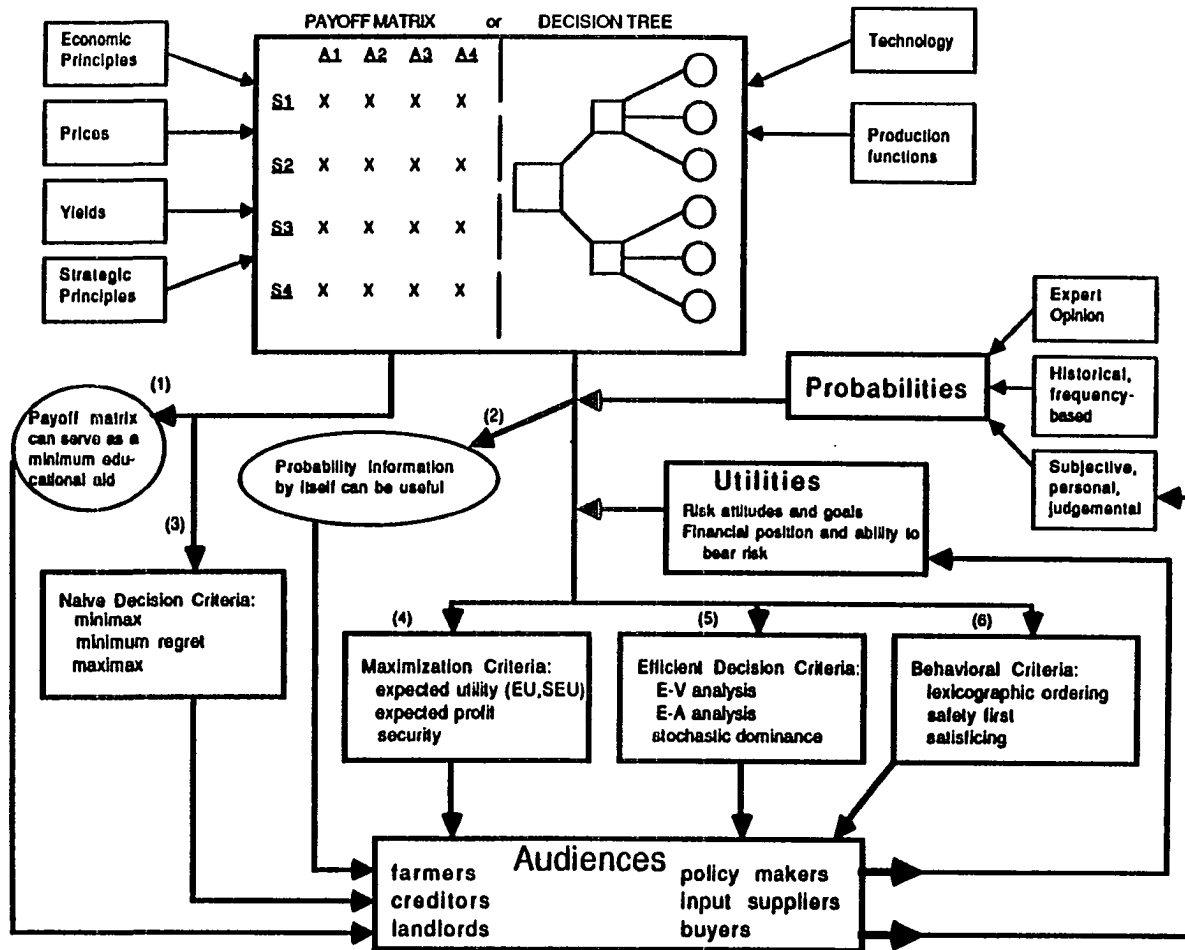
probability of hail damage to half the crop; 1/100 odds of bull being infertile)

4. identify the consequences of each action under each state of nature (e.g. profit on corn if no rain; net present value of tractor if fully utilized; cash outflow per year if combine is leased)
5. assign a utility to each consequence (e.g.  $-\$55/\text{ac} = 3$  utils/ac; entire crop "organically grown" = 8,000 utils; prize bull dies = -50 utils)
6. calculate the expected utility for each action (e.g. plant corn:  $(.2)(-50 \text{ utils/ac}) + (.5)(20 \text{ utils/ac}) + (.3)(100 \text{ utils/ac}) = 30 \text{ utils/ac}$ )
7. choose the action with the largest expected utility (e.g. plant wheat, sell corn at harvest, do not fertilize at pre-plant).

The research comprising this dissertation is concerned solely with step 3: assigning probabilities to states of nature. There is little to be gained here by a more complete development of the necessity or practicability of the other steps except to point out that errors can be introduced at any of the seven steps. For example, the choices of action available or the states of nature possible may not be fully identified; there may be accounting errors in determining the consequences of actions; the utility function may be improperly elicited or incorrectly specified.

Figure 1.1 is adapted from Walker, Nelson, and Olson and is presented as an overview of where SEU fits among the various approaches to making decisions under uncertainty. Items numbered (1) through (6) are described as "educational routes" by Walker et al. to associate them with various research and extension efforts or proposals. In discussing these educational programs, the authors remarked that survey results indicated that farmers find Outlook information in probability form useful. They also commented on the lack of tested approaches for eliciting subjective probabilities, and called for further research in this area. Note that it is the arrows coming out of the bottom box in Figure 1 which qualify the corresponding approaches as "subjective". This provides the logical connection to "self-satisfaction on the part of the user".

Ultimately, the usefulness of SEU theory will be judged by its acceptance by the user, and the tests of its validity will come also from comparing predictions to actual economic behavior. We might therefore expect that a measure of the *rate of acceptance* of the theory would be the proliferation of educational and extension programs promoting and explaining its use. Currently, there are few such programs. Michigan and Minnesota appear to have extension programs which incorporate decision theory. A series of audio-visual instructional materials on such



**Figure 1.1. Educational Routes in Extending Methods for Decision Analysis**

topics as "Using Probabilities in Making Farm Decisions" was developed by Oregon State University (Walker et al.).

It is difficult to imagine much progress in promoting SEU when almost no Outlook reports provide probability forecasts (in place of point forecasts). Some exceptions may be found in Black's study of corn and soybean production, and in King and Lybecker's study of pinto bean marketing in Colorado. It should be noted that in assigning probabilities to states of nature, the use of any probability distribution other than one elicited from the user (such as historical frequencies, or "expert opinion") makes validation of SEU theory via observed behavior impossible. Such validation becomes confounded because it contains two hypotheses: (1) people use SEU in decision-making, and (2) people use historical frequencies or expert opinion exclusively in forming their expectations. This point seems to have been overlooked by a surprising number of authors.

Since a search for large scale applications of SEU theory proved disappointing, attention was turned to individual studies in the research literature. A recent series of papers on empirical estimation and use of risk preferences in the agricultural economics literature focused on three approaches to testing utility theory: observed economic behavior, experimental methods, and validity of axioms (Robison). Rather than paraphrase the

progress made in these three areas, suffice it to say that no agreement on satisfactory methods of validation was reached by the researchers involved. Perhaps the cause of this impasse is the failure to divide a complex problem into a set of simple ones in order to isolate and control possible sources of error. Ironically, this is one of the situations mentioned earlier which recommends the use of SEU in the first place. The seven steps previously enumerated as characterizing the SEU method would be obvious choices for smaller sets in which to look for significant sources of error.

Most of the research attention in risk analysis seems to have been on step #5, "assigning utilities to consequences". Many refinements have been made in elicitation procedures (Officer and Halter, Norris and Kramer). However, since Binswanger's (1980) study in India (in which he demonstrated that utility functions elicited with hypothetical payoffs were significantly different from those elicited with sizeable monetary payoffs) further studies requiring field elicitation of utilities have tapered off. Paradoxically, the suggestion that Binswanger's method be replicated in the U.S. with payoffs significant to U.S. farmers seems repugnant to funding agencies and the profession at large.

Few studies in the agricultural economics literature have sought to critically examine step #3: assigning

probabilities to states of nature. One exception is the work of Grisley and Kellogg (1983). More will be said about their study in the context of methodological problems in elicitation procedures in Chapter II. Hanemann and Farnsworth reported that they found no difference in risk attitudes between farmers who adopted integrated pest management methods and those who used conventional chemical controls. However, they found significant differences in subjective expectations of yields and profits between the two groups (despite the fact that historical data on yields and profits did not support such differences). Many other authors have made passing reference to the importance of considering the subjective element of probabilities but few have considered how to isolate these effects. It is the premise of this dissertation that the appropriate way to examine these subsets of decision theory is through experimental methods. To quote Binswanger (1982):

"The advantage of experimental studies...is that rather than making assumptions about the features [of specification error], one can design experiments where many of the features are under control of the experimenter, and where it is therefore easier to focus on testing subsets of the assumptions of the theory than with alternative approaches." (p. 392)

"...it is important to realize that experiments cannot be expected to provide answers to all questions. Furthermore, the experimental psychology literature presents clear evidence that it is as easy to misspecify an experiment or misinterpret its results as for econometric or programming studies. Nevertheless, the casual treatment of experimental methods in the papers of this session and in the profession as a whole is unwarranted." (p. 393)

Attempts to justify the legitimacy of econometric analysis sometimes contrast economic and experimental sciences in order to distinguish the approach of each to "objective" information. Sims maintains that in the experimental sciences it is the nature of the experiment in which the data is gathered that determines the objectiveness of observations. Thus argument over such observations would likely focus on whether procedures followed in the experiments met certain criteria, such as whether the experimental conditions were realistic in the required sense, or whether the methods for randomizing the choice of samples were adequate. In economics the degree of objectivity of observations arises from the degree of agreement among people with whom the results are shared. Consequently, there can be little focus on "procedures".

In the next section, an effort is made to demonstrate how the procedures of experimental methods can greatly strengthen the foundations of economics, particularly in a microeconomic context.

### **Role of Experimental Economics**

Experimental economics can be defined as the study of economic behavior under controlled and replicable conditions. "Control" is necessary to assure that the variable under examination (the one that constitutes the "treatment") is the only parameter being varied. In this context it is similar to the *ceteris paribus* conditions so prevalent in economic theory. "Replicability" is necessary to assure that later researchers who wish to verify or extend a study will have reasonable success in imitating the conditions under which the original study was conducted.

Economic experiments can be conducted in the field or the laboratory, depending on the nature of the hypothesis. Field studies have included: investigation of peak load pricing of electricity (Battalio et al. 1979); income maintenance and the negative income tax (Kershaw and Fair; Pechman and Timpane); token economies in prisons, psychiatric wards, aircraft carriers and similarly isolated communities (Kagel); and allocation of housing statistics



in Sweden (Bohm). A liberal interpretation would also include the whole class of "marketing research" typified by the marketing departments of business schools and major corporations. Characteristics which distinguish field studies from laboratory studies are largely related to expense, logistics, and manpower requirements. Some field studies cost millions of dollars, last several years, and may employ large staffs of researchers, medical support, and administrative personnel. At this stage of economic inquiry it appears that the role of field experimentation is in validation of results from laboratory studies, i.e. in examining the generalizability of small scale results to situationally richer environments.

The protocol used in laboratory experiments in economics has become fairly standard. Subjects are typically recruited from convenient undergraduate classes (except not from those taught by the experimenter, for obvious reasons). This is satisfactory when the hypothesis being tested does not apply only to a specific population, as is the generally the case in economic theory. A set of instructions which includes the specification of monetary compensation for participation and "good" performance is usually read to the subjects at the start of the experiment. The role of these cash payments is described by Smith (1976) in his theory of "induced value" (more will be said about this feature in a later discussion).

After reading the instructions, subjects then engage in some game-playing or role-playing situation that embodies the essential features of the hypothesis being tested. Although these situations sometimes appear to be gross oversimplifications of reality, they go much farther in approximating authentic conditions than do most of the theories that are being tested. It is in this way that laboratory experiments bridge the gap between theory and observations of the real world.

During the experiment the behavior and responses of the subjects are observed and recorded. These data are then analyzed, typically with simple but robust statistics such as t-tests, F-tests, ANOVA, or their non-parametric counterparts. Since much behavior involves dynamic situations and learning, simple graphs and time charts often tell a revealing story.

Laboratory sessions usually last only two to four hours in order to avoid complications from fatigue or boredom on the part of subjects. Sometimes it is necessary to train subjects beforehand in the mechanics of the game or to pre-select those who exhibit characteristics essential to the research question.

What makes laboratory experiments a valid source of data about the real world? Smith (1982) proposes five sufficient conditions that constitute a valid, controlled microeconomic experiment:

1. Nonsatiation: subjects prefer more goods/money to less.
2. Saliency: the experiment has motivational relevance which links the reward to the task.
3. Dominance: the anticipated rewards in the experimental setting dominate any other costs or benefits which might affect performance of the task.
4. Privacy: subjects are informed only about their own payoffs so as to control for interpersonal utility.
5. Parallelism: propositions derived from laboratory experiments will apply wherever similar *ceteris paribus* conditions hold.

Conditions #1 (Nonsatiation) and #2 (Saliency) are required to create a *microeconomic environment*.

Nonsatiation is a powerful axiom of preference theory which enables us to make predictions about a person's preferences among bundles solely from the observable and measurable quantities of which the bundles are composed. Saliency requires that subjects understand the task to be performed and the rewards to be earned under various conditions, and that these should be obviously related to the performance of that task.

Conditions #3 (Dominance) and #4 (Privacy) are needed for *experimental control*. Subjects whose opportunity costs of time or effort are greater than their expected reward from the experiment may become distracted, or bored,

or may hurry through the task in order to leave sooner. When specific experience or expert knowledge is not a condition of the hypothesis being tested, the concept of dominance is the principal reason for using students in experiments: the cost for subject payments is lower. It is essential to provide privacy since an experiment in which subjects are supposed to behave so as to maximize their absolute payment can become hopelessly complicated if they are also trying to outperform others in their relative payment. In addition, regulations of funding institutions now require strict confidentiality of results where human subjects are concerned:

Conditions #1 through #4 are required to achieve *internal validity*, "the basic minimum without which any experiment is uninterpretable: did in fact the experimental treatments make a difference in this specific experimental instance?" (Campbell and Stanley, p.5). It is argued that internal validity is a necessary (though not sufficient) condition for a laboratory experiment (or any experiment for that matter, be it econometric, simulation, optimization, or otherwise) to be a valid source of data about the real world. This is achieved by what the experimental sciences call "control": the ability to link a specific response to a specific stimulus because all other variables are being held constant. The achievement of control is what allows the experimenter to select the

best explanation of the results from among several competing explanations. This is what enables science to progress.

Condition #5, Parallelism, allows results from the laboratory to be transferred to the many other situations that constitute the "real world". It is a condition of *external validity* or generalizability: "to what populations, settings, treatment variables, and measurement variables can this effect be generalized?" (Campbell and Stanley, p. 5). Often this condition is difficult to satisfy in practice. For example, a straightforward test of the hypothesis that a "nuclear winter" would follow a nuclear war would be unthinkable.

In cases where a critical field validation experiment is impractical it may be possible to expand the number of laboratory tests to include an increasing variety of conditions so as to define the range over which the hypothesis could be expected to hold. Note that a direct field test of a hypothesis which seems to satisfy the condition of parallelism but which has not secured "control" cannot be used to categorically eliminate alternative hypotheses. Sources of error from uncontrolled variables can often produce confidence regions which are large enough to include the results predicted by other hypotheses. Thus, internal and external validity are both required in order to provide the necessary *and* sufficient

conditions for a laboratory experiment to be a valid source of data about the real world (even though each is necessary but neither by itself is sufficient).

What are the advantages of doing laboratory experiments in economics? Plott lists five potential advantages. First, laboratory experiments generate data in a controlled environment. This allows experiments to be replicated independently by anyone skeptical of the original results. Second, most experiments can be executed in a few hours and can be used to simulate transactions which might take days or even years to observe in the real world. Third, experiments in laboratory settings are generally much less expensive than in field settings. Fourth, the environment is flexible and can be changed readily to simulate a variety of conditions, some of which (e.g. "wartime inflation") cannot be observed in the field. This allows investigation of a wider range of parameters. Last, in attempting to subject a theory to experimental validation, users of the theory are required to "operationalize" it: does the theory specify certain conditions required for it to work (such as certain characteristics of agents, maintained functional forms, econometric regularities, stopping rules, equilibrium conditions, etc.)? Quite often these are defined loosely (if at all) in the theory and these conditions come to light in even the simplest of experimental settings.

Having extolled the virtues of laboratory experiments, let us explore some of the ways that such experiments can go wrong. Referring to the previous description of a typical experiment, a common source of problems is that subjects do not understand the task. This may be because the instructions are not clear, or are misleading, or simply because the task is too complicated. Consider an experiment in which subjects are (optimally) required to calculate in their heads the expected values of 35 choices and sum them across five categories in under 30 seconds. Obviously the complexity of the task is far out of proportion to the resources made available to the subjects for the performance of the task. Violations of the conditions of saliency, dominance, and parallelism could all be present in this situation. Most experiments engender, to some degree, a "hide the ball" situation, particularly when examination of the ability to find the ball (i.e. maximize earnings, or behave as theory predicts) is the purpose of the experiment. However, when subjects are not given a reasonable chance of behaving "optimally", generalization to the real world may be questioned. If the 35 choices above were in fact the potential oil reserves located in five regions, then a team of geologists, accountants, and investors who were given several months could probably reach a decision close to the optimal predicted by theory. Some of the popular literature on

subjects' use of "heuristics and biases" in forecasting may arise from unreasonable hide-the-ball experimental settings.

Another problem is that the structure of the reward mechanism may suggest to subjects a strategy which was overlooked by the experimenters. Such a situation might arise when subjects discover that collaboration has not been specifically disallowed by the experimenters and that far more payment can be extracted with this strategy than was anticipated. One of the experiments described in this dissertation addresses another situation where certain payment functions could induce subjects to give answers to questions that differ from what they believe to be the truth, but which happen to maximize their payoffs.

Further problems may arise when rewards fail the dominance criterion. This is most often the case when payments are too low and subjects, perceiving inadequate compensation for their time or effort, behave inconsistently or erratically (note that it is difficult to say categorically that they perform "poorly" or "wrongly" without nesting an assumption about "good" or "right" behavior). Occasionally the combination of monetary and non-monetary rewards may be too high. This would not be a problem if the auxiliary incentive serves only to reinforce whatever behavior the subject would have revealed without it. The problem is that the experimenter cannot always be



sure that this is the case since he may not control the auxiliary reward function. Therefore he surrenders some control gained via the theory of induced value. An example is the classic "princess effect" whereby workers who were singled out for special attention (due to their participation in the experiment) produced more goods in each successive trial regardless of the level of the treatment variable, presumably because they were basking in the positive feedback they received in the form of attention. Another case could arise in attempting to examine the price forecasting ability of farmers by using actual prices as they evolve in "real time". In this case, a few dollars reward for a correct forecast could be overwhelmed (in the sense of the dominance criterion) by the consequences of that forecast in decisions involving thousands of bushels of grain in storage.

As a final comment on sources of errors in experiments, the results of experiments are sometimes questioned on the grounds that the subjects used were not "representative". This is certainly a valid criticism if the theory being tested is conditional on a specific subject pool, e.g. "white males ages 30-39". Most theories in economics do not specify that the economic agents in question must have certain characteristics (physical, behavioral, cultural, socio-economic, or otherwise). However, even where the theory does not specify subject

pool characteristics, it can happen that the task is so designed that it excludes some groups of people from being able to behave "optimally" or "rationally". Thus cases arise where sociologists accuse economists of training people to be free-riders since the only subject pools which displayed that behavior were populated by economics students (Marwell and Ames); or results seem to indicate that statisticians are better at predicting stock prices than stock specialists for banks (Stael von Holstein). Careful examination of the conditions under which the experiment was conducted usually reveals some element which, though originally overlooked, turns out to make a considerable difference when the experiment is replicated independently. In the free-rider experiments all subjects eventually began to free-ride when the experiment was allowed to continue for multiple periods (Isaac, McCue and Plott). In the stock price study, as Stael von Holstein pointed out, the explanation could well be that statisticians understood the rules of the probability calculus which were needed to perform "correctly" while the stock brokers were not familiar with those conditions unique to the experiment. In any case, good experimental design and protocol should obviate problems that might arise concerning the representativeness of subjects. In experimental economics when different subject pools have been used in the same experiment it has proven extremely

rare to find significant differences among subject pools. In the few instances where this has occurred, it has generally not been considered worthwhile to reformulate the theory to account for it, nor to incur the expense of examining it in greater detail.

Let us return now to the question of paying subjects. Why do we pay subjects? Many experiments have been done where subjects were not paid, so why are payments in experimental economics so important? The answer to the first question is that we pay subjects in order to induce monetary value on actions. A number of studies have examined payment vs. non-payment as a treatment variable. Siegel found that when he held the complexity of the task constant and increased the reward, the number of reward-maximizing (salient) choices made by subjects increased. Then when he held the reward constant and increased the complexity of the task the number of reward-maximizing choices decreased. Smith (1976) found that when there were no rewards, or when rewards were chosen randomly, the responses were much less consistent when he tried to replicate them than under conditions where there was a known reward. Phillips and Edwards investigated subjects' ability to incorporate new information in their decisions ("learning"). They found that more learning occurred under payoff conditions than under non-payoff conditions and that there was less variation between subjects' responses in the

payoff group than in the non-payoff group. Since significant differences between treatments are more likely to be found when variance is smaller, these results indicate that experimenters have a better chance of obtaining unambiguous results when they pay subjects.

In a recent study using hypothetical (no pay) and real (pay) gambles, Jiranyakul examined several theories which have been devised to explain results which were inconsistent with the classical von Neumann-Morgenstern theory of expected utility (EU). In almost every case, subjects who gave responses inconsistent with classical EU theory when the reward was hypothetical gave EU-consistent responses when the gamble was associated with real payoffs.

It seems obvious from these studies that, in the absence of unequivocal information that the experiment will meet the conditions of saliency and dominance without payments, one should plan to pay subjects. What seems surprising is not that so many studies have been done without payments but that so many experimenters would jeopardize the significance of their research results because of apparent obstinacy or trepidation about paying subjects.

### Objectives of this Research

The theme of this study is that subjective probabilities are important in all aspects of decision-making under uncertainty. This thesis takes two approaches to extending knowledge under this theme.

The first approach deals with decision-making at the individual level. To study this the researcher may need to elicit subjective probabilities that accurately represent the agent's degree of belief in the likelihood of occurrence of some event--the probabilities that the agent actually incorporates into his decision problem. A monetary reward may be used to induce the agent to reveal these beliefs. The theory of proper scoring rules deals with the selection of such rewards and their effect on agents' responses. The first objective of the research described in this thesis is to articulate this theory so as to predict certain observable behavior, and then to observe whether agents in controlled conditions actually behave as predicted.

The second approach deals with decision-making at the aggregate level. Here a knowledge of aggregate subjective probabilities or expectations becomes important, especially in the areas of forecasting and policy-making. In such applications it is often convenient to use a proxy estimate of aggregate expectations rather than to elicit probabilities from some or all of the individuals involved.

Marc Nerlove has proposed that such a proxy could be obtained in the form of forecasts from a time series model of a simple set of past observations. The second objective of this research is to attempt to validate this theory under certain controlled conditions of a microeconomic laboratory environment. As with any observational science, it is hoped that in the process of validation some useful insight and extension of the theory will be forthcoming.

CHAPTER II  
SUBJECTIVE PROBABILITIES ELICITED UNDER PROPER AND  
IMPROPER SCORING RULES: A LABORATORY TEST OF  
PREDICTED RESPONSES

Introduction

The concept of subjective probability is at the heart of contemporary economics. In microeconomics, subjective probability and utility are the cornerstones of the Subjective Expected Utility (SEU) hypothesis, which is used to explain and predict the behavior of economic agents. In macroeconomics, the idea of subjective probability (more commonly called "expectations") has replaced that of historical frequencies in explaining macroeconomic phenomena since the Keynesian revolution (Sargent). Yet despite our profession's acknowledgement of the centrality of subjective probability very little has been done in the way of eliciting these probabilities from the actual agents to whom they are supposed to apply. Assumptions that encourage this neglect have often been: (a) that the researcher knows, *a priori*, what agents' subjective probabilities are (or should be), or (b) either all agents hold approximately the same expectations, or the extremes cancel each other to produce some average expectation (one

which presumably supports the first assumption). To avoid the difficulties in eliciting subjective probabilities, the substitution of historical frequencies has been the most popular expedient.

In microeconomic studies, when observed economic behavior has differed from that predicted under SEU-plus-assumption-(a), considerable attention has been directed toward the specification of the utility function or the method of eliciting utility, if indeed any was used. However, at least a few authors have suggested that differences in subjective probabilities could alone account for discrepancies. For example, Hanemann and Farnsworth examined the reasons why some farmers adopted integrated pest management practices while others continued to use traditional chemical controls. They found that although there was no significant difference between the risk attitudes of the two groups, expectations about yields and profits did differ significantly. Furthermore, these expectations were not consistent with historical frequencies. The authors suggested that differences in expectations could account for observed economic behavior where differences in utility functions could not.

The foregoing discussion is not meant to understate the difficulties in eliciting subjective probabilities nor to criticize the profession at large for not using them. However, it does suggest that the use of proxies for



agents' actual expectations (such as historical frequencies or "expert opinion") should be explicitly acknowledged and appropriately justified, and that results should be reported as being conditional upon the validity of such proxies.

Having said this, let us suppose that the researcher has in fact elected to elicit subjective probabilities in his study. His first question might then be: "what does it take to get people to reveal the personal probabilities that they are actually going to use in making the decisions of relevance?" Resolution of this question involves several considerations:

1. do people hold precise expectations, or are they quite fuzzy about the probabilities of future events?
2. are there reasons why they might not reveal their true expectations?
3. do they understand what they are being asked to do?
4. is the time interval between eliciting their probabilities and observing their decisions short enough to exclude the possibility that new information was received in the interim which was used to revise their probabilities or their decision problem?

The first consideration, fuzzy expectations, is an empirical question which has received considerable attention in the psychological literature, particularly with regard to the heuristics that people use in forming

expectations and the biases that often result (Tversky and Kahneman; Einhorn and Hogarth). This consideration becomes important when an attempt is made to replicate a study in order to check for consistency in results.

With regard to the second condition, principles of experimental economics can provide some insight as to why people might not always reveal their true expectations. In describing the ideal economic experiment, Smith (1982) has outlined the conditions needed for a microeconomic environment, a controlled experiment, and an externally valid result. The conditions for a microeconomic environment and external validity would presumably be met by the researcher working with human agents involved in "real-world" decision-making. Smith's conditions for a controlled environment, "dominance" and "privacy", would still be of critical importance.

Dominance requires that, in completing the task set before them by the researcher (in this case, revealing their expectations), the motivation presented to and perceived by the agents outweighs any other motivation to perform otherwise. The dominance problem could arise when a farmer receives a 15-page single-spaced questionnaire with terse instructions and ambiguous questions and is inclined to check off the answers haphazardly simply in order to get the task over with.

The privacy condition could be illustrated by a similar questionnaire where the topic deals with political, sexual, or religious predilections and the respondent is suspicious of the legitimacy of the research or confidentiality of the results. In terms of economic theory, the concern here is with the possible interdependence of utility functions, and the inability of the researcher to control for this.

The third consideration, "do they understand what they are being asked to do?", is primarily a subjective assessment on the part of the researcher. Methods that have been used to confirm whether agents understood the task include: checks for consistency in answers; frequency of use of strategies that are indicated by theory to be dominant, or perhaps "pathological", cases; measures of variability in answers; and post-survey interviews and comments. Training and feedback are obvious techniques for assisting agents in understanding the task, particularly when the task is complex.

The fourth consideration, that of the time interval between elicitation and observation of behavior, would not be worthy of elaboration except that researchers frequently ignore the issue by failing to observe the decisions that are subsequently made, and thereby overlook a significant opportunity for validation of their predictions or results.

Returning to the criterion of dominance, it appears that modest payments (e.g. in the range of hourly wage rates that are appropriate to the subject pool) have been used in experimental studies as standard procedure for first-round experiments with human subjects. The principal rationale has been that consistency of response is increased and variability of response is decreased when payments are used. Although the need to continue payments in further studies under the same protocol is quite case-specific, as is the level of payment necessary, the consensus among experimentalists seems to be that unless or until evidence suggests otherwise, subjects should be paid their opportunity cost or at least minimum wage. This was the underlying rationale in a study done in Thailand by Grisley and Kellogg (1983), although they were able to provide relatively substantial payments. They state:

"....motivational biases may be reduced by rewarding the subject for revealing expectations that are factual. In this research farmers were rewarded for such candor.....To our knowledge, this is the first economic study to use a financial reward for eliciting subjective probability distributions." (p.75)

The present study does not attempt to resolve the "pay vs. not pay" question. It will become evident that there are problems with either choice. Let us simply assume that the researcher has decided to elicit subjective

probabilities and to pay his subjects. The stage is now set for an explication of the rationale for conducting the present study.

### Theory

In the development of reward mechanisms (or gain functions) for the elicitation of subjective probabilities, the following condition was accorded considerable attention: such mechanisms should satisfy the criterion that a subject should be able to maximize his expected value of the reward only by revealing the actual probabilities that he believes to be correct. This would seem a trivial requirement except that the most obvious reward mechanism, that of paying a constant amount ( $\$k$ ) multiplied by the probability assigned to the event that occurred ( $p_i^*$ ), that is the gain function ( $\$kp_i^*$ ), does not satisfy the above criterion of "revealing believed probabilities". To demonstrate why this is so, let the following be the maximization problem, constrained by the requirement that the probabilities revealed should sum to unity:

$$\max E[G] = \sum [r_i(kp_i)] \quad \text{s.t. } \sum p_i = 1$$

where  $E[G]$  is the expected gain,  $r_i$  is the *believed* probability of event  $i$  occurring,  $k$  is a constant of reward

(say, \$1),  $p_i$  is the stated probability, and the summation is over  $i=1, \dots, n$  where  $n$  is the number of possible events. Note that in the determination of expected gain the believed probabilities ( $r$ 's) constitute the expectations, but the gain is a function only of the stated probabilities ( $p$ 's). Forming the Lagrangean function:

$$L(p; r, \lambda) = \sum r_i k p_i + \lambda(1 - \sum p_i)$$

to obtain a maximum we require as necessary conditions:

$$\partial L / \partial p_j = k r_j - \lambda = 0 \quad \forall j=1, \dots, n$$

$$\partial L / \partial \lambda = 1 - \sum p_i = 0$$

Now  $\partial L / \partial p_j$  is no longer a function of  $p_j$  (as we might have expected, given a linear objective function) so we cannot solve the system of equations to find the optimum  $p_j$  to use. Since a corner solution evidently exists, the Kuhn-Tucker (K-T) necessary conditions for a local maximum require:

$$\partial L / \partial p_j \leq 0, \quad p_j \geq 0, \quad \text{and } p_j (\partial L / \partial p_j) = 0 \quad \forall j=1, \dots, n$$

Three cases need to be examined:

Case 1:  $\partial L / \partial p_j < 0$  for all  $j=1, \dots, n$

If all of the partial derivatives are negative, this implies that to satisfy the K-T conditions all of the  $p_j$ 's must equal zero. But this violates the requirement:  $\sum p_i = 1$ .

Case 2:  $\partial L / \partial p_j = 0$  for all  $j=1, \dots, n$

If all of the partial derivatives are equal to zero, then all of the  $p_j$ 's could be positive, or only some of them, but at least one must be positive.

Case 3:  $\partial L / \partial p_j < 0$  for some  $j$ 's and  
 $\partial L / \partial p_j = 0$  for the rest

In this case, some of the  $p_j$ 's must equal zero (i.e., those associated with  $\partial L / \partial p_j < 0$ ), but the rest could be positive or zero, with at least one being positive.

To make it easier to examine the dominant strategy, rearrange the maximization problem as follows:

$$\begin{aligned} \max E[G] &= k \sum_{\substack{i=1 \\ i \neq m}}^{n-1} r_i p_i + k r_m \left( 1 - \sum_{\substack{i=1 \\ i \neq m}}^{n-1} p_i \right) \\ &= k r_m + k \sum_{\substack{i=1 \\ i \neq m}}^{n-1} (r_i - r_m) p_i \end{aligned}$$

where the K-T conditions are:

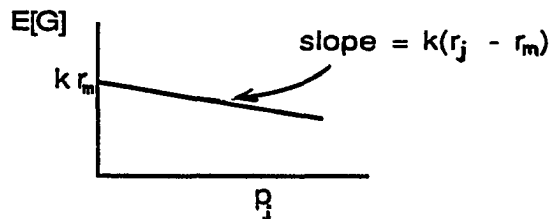
$$\partial E[G] / \partial p_j = k(r_j - r_m) \leq 0,$$

$$p_j \geq 0,$$

$$\text{and } p_j (\partial E[G] / \partial p_j) = 0 \quad \forall j=1, \dots, n$$

Now  $kr_m$  can be interpreted as an intercept, and any one of the pairs in the term  $k(r_i - r_m)p_i$  as a slope of the function  $E[G]$ .

An illustration of one such pair is:



When the partial derivative of the gain function is equal to a constant that is negative (as in the illustration above) then the gain function is maximized at  $(0, kr_m)$  when  $r_m$  is chosen to be the largest believed probability in the suite of  $r$ 's. When the slope is zero, the gain function is still maximized at  $kr_m$  but now it does not matter what  $p_j$  is. Note that this happens when  $r_j = r_m$ , or when the largest believed probability is used more than once.

What this suggests is that when an agent is asked to reveal his probabilities and understands that his primary incentive for doing so is that he will receive  $\$k$  multiplied by the probability he assigned to the event that actually occurred, then in order to maximize his earnings his dominant strategy would be to:

1. "look" at his suite of believed probabilities ( $r$ 's) and pick out the largest one(s),
2. report all his smaller believed probabilities as zeroes



instead of what he really thinks they are,  
 3. report the largest one(s) as anything he wants (as long as the sum of all p's is unity).

For example, under the dominant strategy a suite of r's such as (.2, .2, .4, .2) would be revealed as the suite of p's (0, 0, 1, 0). Or a suite of r's such as (.2, .2, .3, .3) could be revealed as (0, 0, 0, 1) or (0, 0, .8, .2) or any number of other combinations as long as the p's substituting for the largest r's add up to unity. It should be obvious what havoc such revealed expectations would wreak on subsequent analyses of optimal decision-making.

Of course, the key factor that makes this argument a *dominant* strategy is that the agent accepts this reward mechanism as his cardinal incentive to participate, and as such it operates according to the principles discussed previously. In elicitation studies done in the field (rather than in the laboratory) the agent may perceive a greater benefit from cooperating with the researcher and revealing his true r's. If this were thought to be the case, the provision of a subordinate and conflicting reward mechanism would require additional justification.

To deal with the problem of non-revelation of believed probabilities, the concept of a "proper scoring rule" was advanced. A proper scoring rule is simply a reward mechanism or gain function which encourages the respondent to set  $p_i=r_i$  for all  $i=1, \dots, n$ ; that is, to reveal his

believed probabilities. An example of such a scoring rule is the gain function  $\$(k+\ln(p_i))$ , which pays that dollar amount when the  $i^{\text{th}}$  event occurs. To show that this gain function is "proper", we can use the same method as in the previous discussion:

$$\max E[G] = \sum \{r_i [ (k+\ln(p_i)) ] \} \quad \text{s.t.} \quad \sum p_i = 1$$

$$L(p;r,\lambda) = \sum \{r_i [ (k+\ln(p_i)) ] \} + \lambda(1-\sum p_i)$$

The first-order conditions are:

$$\partial L / \partial p_j = r_j / p_j - \lambda = 0 \quad \forall j=1, \dots, n$$

$$\partial L / \partial \lambda = 1 - \sum p_i = 0$$

Summing over all  $i$ 's we have  $\lambda \sum p_i = \sum r_i$ . After honoring the constraint, we have  $\lambda = \sum r_i$ . Now, if the agent is coherent in the sense of the probability calculus (de Finetti, pp. 99-100), we can assume  $\sum r_i = 1$  so therefore  $\lambda = 1$  and the first-order conditions are satisfied when  $p_i = r_i$ , as required for a proper scoring rule.

Some excellent reviews of methods and considerations in eliciting probabilities are provided by Stael von Holstein, Savage (1971), Hampton, Moore, and Thomas, and Norris and Kramer. A discussion of an agricultural application (one in which many of the problems of actual elicitation procedures in the field are raised) can be found in the

article by Grisley and Kellogg (1983), followed by a critical comment on their use of the linear scoring rule by Knight, Johnson and Finley, and the response by Grisley and Kellogg (1985) which provoked this study.

The heart of the issue argued by Grisley and Kellogg vs. Knight et al. has to do with a behavioral assumption that is often made in using the commonly-cited "proper" scoring rules; an assumption which was not explicitly stated earlier in this chapter and which is regularly overlooked in the literature. The assumption which motivates the familiar proper scoring rules such as the logarithmic, quadratic, and spherical scoring rules (see Stael von Holstein, pp.29-30), and which makes them "proper" and the linear rule "improper", is that agents are supposed to have linear (or risk-neutral) utility for the rewards involved. If agents are risk-averse or risk-preferring in the domain of rewards, then the log, quadratic, and spherical rules are no longer proper and we cannot be assured that  $r_i = p_i$ . In fact, if an agent's risk aversion is best specified by a log utility function then the appropriate "proper" scoring rule is the very same linear rule we went to such pains to discredit earlier.

It may be that the familiar rules have enjoyed their popularity because risk neutrality is so routinely assumed in research studies. It has only been in recent literature that optimal strategies for a variety of decision

situations (such as fertilizer rates, pest control, marketing strategies, etc.) have been conditioned over a range of risk aversion parameters (risk aversion is the second most frequently assumed behavior; risk preferring behavior is rarely addressed).

Whatever the reason for the dearth of research on expectations of agents with non-linear utility of rewards, articles by Winkler (1969), and Winkler and Murphy have at least provided the theoretical derivation of proper scoring rules for use in these situations. Since a full explication of their work is not relevant to the present discussion, suffice it to say that if the utility function is known, then a suitably "proper" scoring rule can be derived as a composite function of the inverse of the utility function and any scoring rule that is proper under a linear utility function.

The experiment described in this chapter is an attempt to validate part of the theory of proper scoring rules under controlled laboratory conditions. However, it is not possible to observe the condition  $r_i = p_i$ . Therefore, it is first necessary to determine subjects' utility functions, and then, using a proper scoring rule as a "control" and an improper rule (where certain observable behavior is predicted) as a "treatment", compare the results from the two rules. The hypothesis tested is: subjects demonstrating linear utility for rewards will respond with

"tighter" probability distributions when rewarded by the linear scoring rule (as a result of using  $p_i=0$  more often) than will subjects rewarded by the quadratic scoring rule.

## Method

### Background

Subjects used in the following experiments were all college students chosen from classes in agricultural economics and sociology at Texas A&M University. While no details of biographical data were collected, subjects were drawn from a pool which included all years of undergraduates (and at least one graduate) and a variety of majors including agricultural economics, other agricultural sciences, humanities, business, and engineering. Cost, convenience, and the need for controlled conditions were the primary considerations in using students as subjects. Since the hypothesis being tested applies to all economic agents, it was felt that no additional insight, and considerable extra expense, would be entailed in using so-called "real-world" agents such as farmers.

The first problem faced in this study was the measurement of subjects' utility in the range of payoffs anticipated in the scoring rule experiment. A simple and popular approach to this problem has been to assume that, for such small payments, subjects must be risk neutral (de

Finetti, p.82; Winkler 1967, p.1107). In fact, it appears that no other study in which payments were used to elicit subjective probabilities has ever determined subjects' utility over the range of payoffs in order to select a proper scoring rule. Moreover, Harrison has recently shown that the working hypothesis of risk neutrality cannot be categorically accepted. His method of eliciting subjects' utility of payments was used (with some modifications) in the present study.

To the extent that the utility elicitation method used in this study is but one way of determining utility, the test of the scoring rule hypothesis will be conditional on the validity of the method. However, the variety of utility elicitation methods is not as rich as one might presume from a casual survey of the literature, especially when real monetary payments are involved. Methods such as the "modified Ramsey" (Lin, Dean, and Moore) and the method of Becker, DeGroot, and Marschak (from which Harrison adapted his technique) have imbedded dominant strategies for extracting outrageous payments from the experimenter--strategies which readily become apparent to subjects when monetary payments are a real prospect.

#### **Utility Elicitation**

A total of 113 students from six undergraduate classes participated in the initial experiment. Class members were told that they were participating in an unpaid experiment

which would serve two purposes: to pre-test the method, and to establish a pool of experienced subjects for later, paid experiments. It was explained that the experiment was being conducted to study the behavior of people faced with the choice between a certain gain or a risky prospect (a lottery) when small amounts of payoffs are involved.

Since the purpose of this exercise was to give subjects experience in the lottery game, the protocol was a departure from orthodox experimental procedures in which subjects typically read a set of instructions prior to beginning the experiment. Although the game at first appears simple, previous pilot tests using printed instructions indicated the need for visual aids (overhead transparencies, randomizing devices, record sheets, etc.), careful and consistent explanation of the steps in the game, and a variety of examples.

Each subject received a packet containing a consent form, two cards with their unique code number on each card, a record sheet, and a stack of "tickets" for use in the lotteries. Those who wished to participate in the experiment turned in their signed consent form and one of the cards filled in with their name, address, and phone number.

It was explained that the game would be made up of several trials, and that in each trial a new set of "odds" would be posted. The odds represent the chance of winning

the lottery. For example, if the odds posted for a certain trial were 84:16 then there would be an 84% chance of winning \$1.00 and a 16% chance of winning \$0. Each ticket represented the costless right to play one lottery.

To determine the outcome of a lottery, a number was drawn from a bingo cage filled with wooden balls numbered 1-100. If the number drawn was larger than the second number in the odds pair (e.g. 16, above) then subjects would record a win of \$1 on their record sheet; otherwise, they recorded \$0. However, prior to drawing the number that determined the lottery outcome, subjects were given an opportunity to sell their tickets to the experimenter. To do this, they indicated on a "value scale" (provided with the ticket) the amount of money which best represented the level of compensation at which they would be indifferent to receiving that amount (and surrendering their ticket), or keeping the ticket and being entitled to play the lottery under the odds posted for that trial.

To determine who would sell their ticket and who would play the lottery, a number was drawn from the bingo cage. If this number was larger than the "indifference level" marked on the value scale, then it was supposed that the subject preferred the amount drawn and was therefore willing to sell his ticket for that amount. If the number drawn was less than the indifference level, then it was supposed that the subject would rather keep his ticket and



take a chance at winning \$1 by playing the lottery. If the number drawn was exactly equal, then the toss of a fair coin would determine whether the subject kept his ticket or sold it. Hence, at each trial the new odds were posted, subjects filled in the value of indifference on their ticket, a draw from the bingo cage determined those who would sell for that amount, and (for the remaining subjects) a second draw determined the outcome of the lottery.

The theory and use of the value scale as a demand-revealing mechanism is discussed in Becker, DeGroot, and Marschak. In these experiments, subjects were simply shown an example of the importance of being "accurate" in revealing their true point of indifference (a written explanation was given in the instructions used in later experiments; see Appendix A).

After the game was explained to the subjects, one of the code cards was drawn randomly from a box and it was announced that the person holding that code number would receive actual payment for his winnings at the end of the session. This was done simply to help subjects think in terms of real, rather than hypothetical, money.

Class periods lasted 50 minutes, so by the time the explanation of the game was completed there was only enough time left for 10-12 trials. Due to this small number of data points, statistical tests for linearity of the utility

function were generally inconclusive. As a result, subjective criteria based on examination of a graph of each subject's utility function were used to select subjects to participate in subsequent experiments (in which all subjects were paid their winnings). In general, these criteria were: that a polynomial of the first order should fit the points about as well as second or third orders, and that there should be few outliers which were grossly inconsistent with the expected shapes of the function (linear, concave, or convex). From these results, approximately 40% of the subjects (45) in the pre-test were determined to have understood the directions of the utility game and seemed to have good prospects for displaying linear utility in a paid experiment. These subjects were contacted and 27 were able to participate in a paid experiment at one of the three dates offered.

In the paid experiments the instructions provided in Appendix A were read *verbatim* to the subjects. Sessions required about 80 minutes to complete 24 trials. Aside from the use of printed instructions and actual payments, the procedures were the same as in the pre-test. Subjects earned \$15-\$18 each in the paid tests.

It should be noted that subjects accumulated their earnings with each successive trial. Harrison (p.8, footnote 1) comments on the rationale for this procedure. It differs from the procedure of Becker, et al. in which

subjects were paid for only one trial (with that trial chosen at random) in order to avoid the "wealth effect". This effect is thought to influence the interpretation of what is being measured at each trial: is it simply the utility of a 25:75 chance in a \$1:\$0 lottery compared to a certain amount (say \$0.25), or is it the utility of that amount as an increment to the wealth that has accumulated to that point? For purposes of this study, Harrison's method was congruent with conditions that subjects later faced in the probability elicitation experiments, since wealth accumulated in those experiments too.

The responses in the paid test were analyzed using the standard F-test described in Harrison. Of the subjects participating in the paid utility test, 68% demonstrated linear utility functions up to a level of significance of 0.12 and one was subjectively included by overruling the F-test as inappropriate (the subject was perfectly linear except at the 99:1 odds). These 19 people formed the pool from which subjects were contacted to participate in the probability elicitation experiments.

### **Probability Elicitation**

The instrument used to elicit subjective probabilities of future events was an interactive computer game called FORECAST which was specially developed for this research. Only the general features of the program relevant to the experiment are described here.

Subjects logged on to a terminal connected to a PRIME computer and provided personal information, including the code number assigned in the utility experiments, which uniquely identified their input. They viewed a "historical series" of 40 positive and negative integers ranging from about -2000 to +2000. From this information they were asked to provide a forecast, in probability form, of the likelihood of the next number (the "event") falling in one or the other of eight ranges. The actual outcome was then revealed and the subjects received a score, actually a monetary payment, which reflected the accuracy of their forecast as determined by the scoring rule being used with each group. They continued to make such forecasts for 40 periods, receiving summary information about their forecast performance as well as the accumulating historical series as they progressed. At the end of the session they were paid the sum of their earnings from each period.

A representative set of instructions is presented in Appendix B with appropriate tables for the different treatments. Except for the table entitled "Payments Possible Under Different Probabilities and Outcomes", and examples in the text where payments are mentioned, the two instructions were identical. The differences reflect the two scoring rules used as "control" and "treatment" in testing the hypothesis in question: quadratic (proper) in the first case, and linear (improper) in the second.

The quadratic rule was chosen as the proper scoring rule because it allows the use of zeroes as probabilities. In this respect it differs from the logarithmic scoring rule which must be adjusted so as not to penalize unduly for assigning a zero probability to an event that subsequently occurs. The quadratic rule was also appealing because it can be decomposed into three components which offer some insight into the way subjects organize information to produce their forecasts (Murphy).

The actual payment function incorporating the quadratic rule was:

$$\$G = 0.2 \left[ 2 - \sum_{i=1}^n (p_i - d_i)^2 \right]$$

where  $\$G$  is gain (payment) for each forecast period,  $p_i$  is the probability assigned to the  $i^{\text{th}}$  event,  $d_i$  is a binary variable equal to unity if the  $i^{\text{th}}$  event is the one that actually occurs and equal to zero otherwise, and  $n$  is the number of events possible (eight, in this study). Note that the quadratic rule includes compensation for probabilities on events that did not occur, unlike the logarithmic and linear scoring rules which are functions only of the probability placed on the event that occurs.

The payment function using the linear rule was:

$$\$G = 0.75 (p_i)$$

The scaling factors for the two payment functions were determined from pilot experiments and were designed to yield average earnings of approximately \$4/hr. Each session lasted about two and a half hours. Subjects were not told how many forecast periods there would be, although they were assured that the session would last no longer than three hours. Eight subjects were randomly chosen to play the FORECAST game under the quadratic rule, and eight other subjects played under the linear rule.

The data series used in all experiments was generated using the time-series software TIMESLAB (Newton). The generating process was specified to be a univariate, autoregressive time series of order one [AR(1)] with a coefficient of 0.85 and an error variance of 99,900. This specification produced a series of numbers positively correlated with their own first lags, centered approximately on a mean of zero, with one standard deviation of realizations (about two-thirds of the series) included between the tail ranges ( $\pm 600$ ). The numbers generated were allowed to be either positive or negative so that subjects were not led to believe that it could be an actual price series (Smith 1976, p.278).

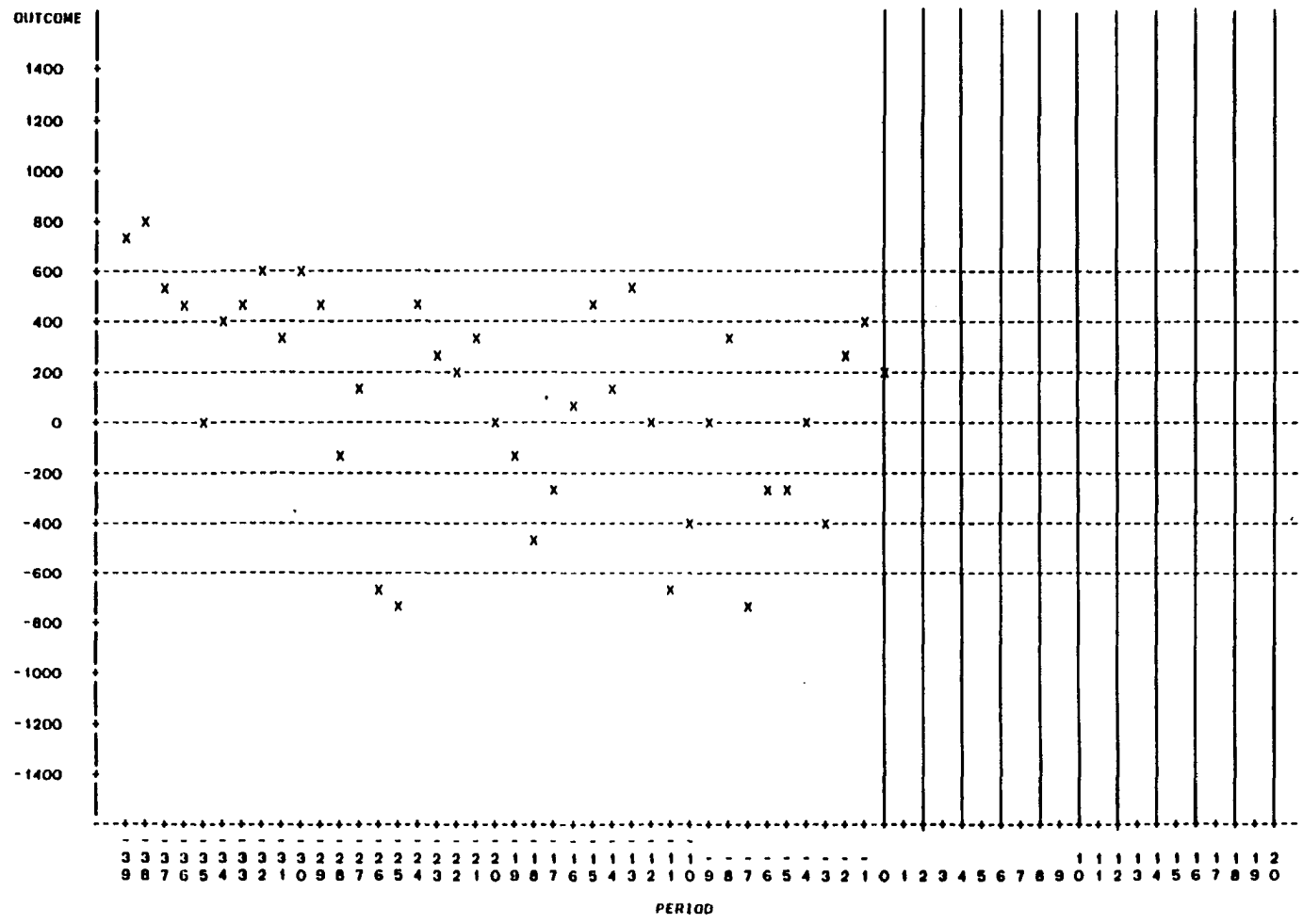
The first 40 numbers in the series (those seen as "historical data" by the subjects) were confirmed to be fit best by an AR(1) process, using Box-Jenkins methods and Akaike's criterion of Final Prediction Error (FPE). The

same confirmation was obtained from all 80 points as a check that those particular points were not artifacts of an unusual "run" in the original series. These 80 points then constituted the input file for the FORECAST program. The output files from the program were: the "Final Report" for each subject which included the probability distribution used in each forecast; the actual outcome; current-period and accumulated earnings; and (for the quadratic rule) the decomposition into the three components of "inherent uncertainty of the event", "calibration", and "resolution" (Murphy). These partitions were not relevant to this experiment and thus are not reported here. Another output file captured all the screen images produced in real time from each terminal during the course of each subject's session.

Figure 2.1 is a graph of the series shown to subjects before they began forecasting. Figure 2.2 is a graph of the series after the subjects had completed 40 forecast periods. The horizontal dotted lines divide the event space into the eight ranges mentioned earlier. These were numbered consecutively from the bottom up.

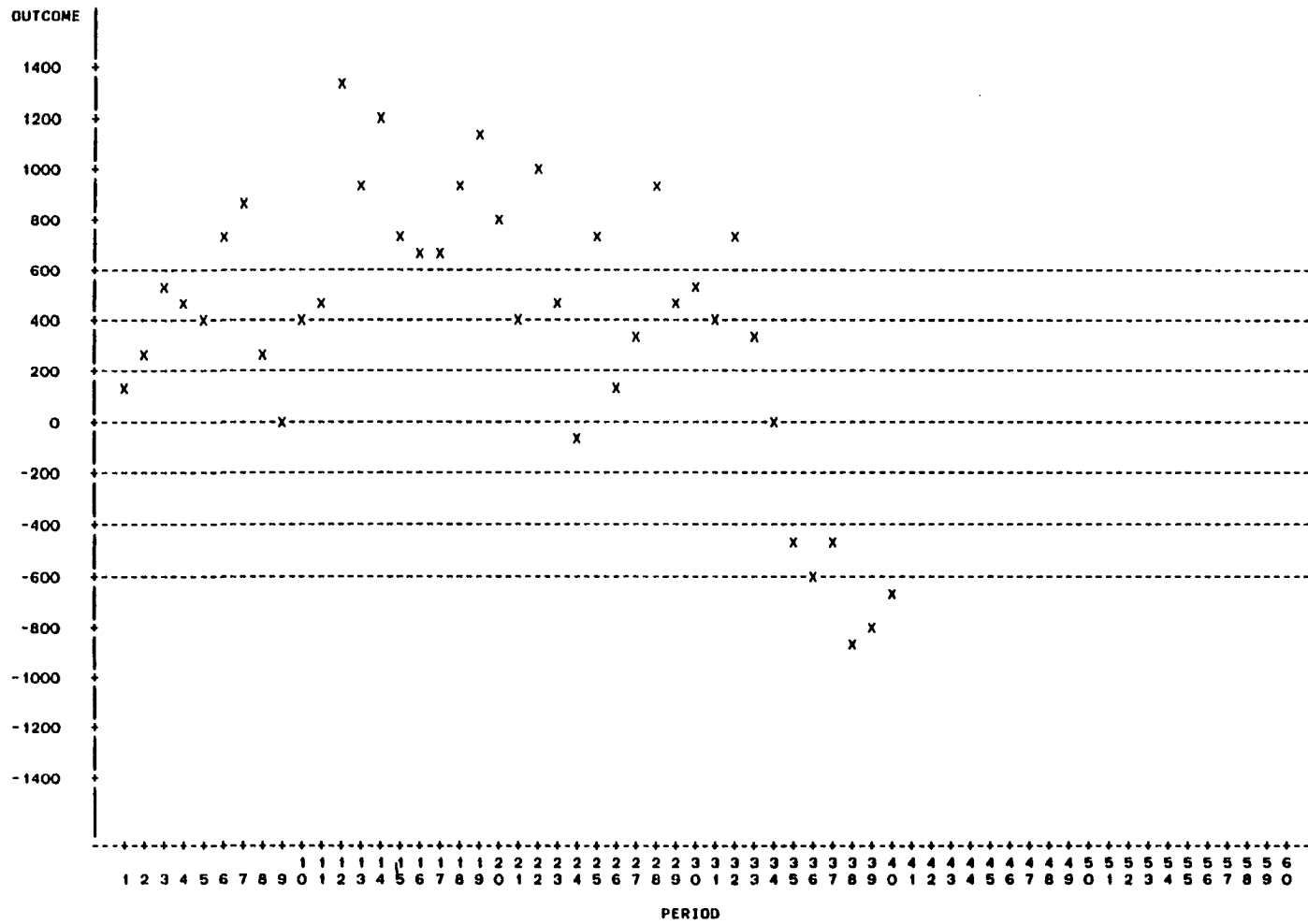
### Results

The hypothesis was that subjects use "tighter" distributions under the linear scoring rule than under the



**Figure 2.1. "Historical" Series of 40 Periods Shown to Subjects Prior to Forecasting**





**Figure 2.2. Graph of Realizations after Completion of 40 Forecast Periods**

quadratic. The model used to test this hypothesis was that the number of zeroes used by a subject in a forecast is a function of the scoring rule used. An analysis of variance showed that the scoring rule used (linear or quadratic) had a highly significant effect ( $p < 0.0001$ ) on the number of zeroes used in a forecast when the observations from eight subjects in each treatment over all 40 forecast periods were used. However, when observations over the first 5, 10, and 15 forecast periods were used there was no significant difference between scoring rules. Table 2.1 shows the progression of significance values over various partitions of the series. Scoring rule only begins to have a significant effect (at conventional levels) between periods 17 and 19. Reference to Figure 2.2 indicates that this particular series of realizations enters the upper tail range at period 12 and remains there for the next eight periods. It appears that while subjects under both scoring rules began to use tighter distributions when the series was consistently in this tail range, subjects under the linear rule *continued* to use relatively tighter distributions for the remainder of the session. This conclusion was supported by results from observations over periods 21-40 and 31-40 which showed that the scoring rule effect persisted after period 20 even in smaller samples.

**Table 2.1. Significance Level Associated with the F Value (PR>F) for Various Combinations of Forecast Periods in ANOVA Model of Effect of Scoring Rule on Number of Zeroes Used in a Forecast**

Forecast Periods Included in Model	PR>F
1 - 5	0.9500
1 - 10	0.2555
1 - 15	0.3095
1 - 16	0.2054
1 - 17	0.1249
1 - 18	0.0867
1 - 19	0.0549
1 - 20	0.0644
1 - 25	0.0044
1 - 30	0.0002
1 - 40	0.0001
- - - - -	
21 - 40	0.0001
31 - 40	0.0041

### Conclusions

The subjective evaluation of these results suggests that some distinctive training or feedback effect resulted from subjects' experience in the tail-range sequence, and this brought about the expression of scoring rule-related behavior which otherwise might have taken considerably longer to manifest itself. An alternative explanation is that learning occurred gradually and at approximately the same rate over the whole series. Since the learning behavior observed in this experiment is unique to the particular series used, further experimentation would be required to distinguish between these competing explanations.

Nevertheless, a number of conclusions can be drawn which may help direct further investigations. First, it is evident that learning to respond according to scoring rule theory takes training and feedback, but it can eventually occur in a fashion which is predicted by the theory, at least in terms of the weak prediction of "tightness" of distributions. Second, it seems likely that subjects who are not trained in the quadratic scoring rule will respond with probability assessments that are not influenced significantly by the choice of a scoring rule when they are asked to provide a *limited* number of one-step-ahead forecasts. Thus, in situations where researchers elicit

forecasts of prices or yields from farmers using a monetary reward (as in the Grisley and Kellogg vs. Knight, Johnson and Finley debate), it may not matter which scoring rule is used. This information may be helpful to researchers who are contemplating a field elicitation study with small monetary rewards, since the linear payment rule is considerably easier to explain to subjects than are any of the proper scoring rules. The validity of these conclusions with respect to much larger ranges of rewards (including losses) and other scoring rules is open to further research.

**CHAPTER III****QUASI-RATIONAL EXPECTATIONS: EXPERIMENTAL EVIDENCE****Introduction**

The specification of the expectations of economic agents has had that peculiar genesis which is best described by George Santayana's epigram: "one learns to itch where one can scratch". Model sophistication has progressed from the cobweb, to extrapolative expectations, to adaptive expectations, to distributed lags, and finally to rational expectations. However, with the exception of rational expectations, rigorous economic or behavioral underpinnings were noticeably lacking.

Apparently discontented with the minimal economic content of distributed lag models and the omniscience required for fully rational expectations, Nerlove (1967) developed a synthesis which has considerable economic content and yet can be readily applied in practical contexts. This model later became known as "quasi-rational expectations" (Nerlove, Grether, and Carvalho) to emphasize its close correspondence with fully rational expectations in many situations.

Nerlove's theory incorporates a central premise of rational expectations: in forming their expectations,

rational agents extract much of the relevant information about stochastic processes by examining historical data on the variables generated by those processes. Nerlove operationalized his theory by suggesting that under fairly reasonable conditions the relevant information about a particular stochastic process could be fully extracted from an examination of a univariate (or small multivariate) set of data, rather than requiring that agents know the structural parameters for the whole model. Consequently, he proposed that the forecasts from an optimal statistical predictor such as an autoregressive integrated moving average (ARIMA) or simple vector autoregression (VAR) model could be substituted for expectations from quasi-rational agents (Nerlove 1979).

Our interest in his theory stems from a desire to obtain an appropriate proxy for agents' expectations in order to avoid the inconvenience of actually eliciting these expectations from individuals. There are two ways to supply data on agents' expectations. The first, and by far the most popular, is to assume that such expectations are formed by some stochastic (or even deterministic) process which is known by the researcher. Such assumed processes have included the cobweb model, random walk, distributed lags, and certain behavioral mechanisms such as adaptive expectations and "anchoring and adjustment" (Tversky and Kahneman; Einhorn and Hogarth). These approaches often

appeal to aggregation phenomena and the concept of the "efficient market" for justification. This case has been argued for both the random walk and commodity futures market prices as substitutes for aggregate expectations (Fama; Gardner).

The popularity of these approaches is warranted when one contemplates the alternative approach to obtaining agents' expectations, that is to ask them. The concept is simple but the mechanics are often staggering. Survey methodology is well established and is in fact employed regularly to elicit statistically relevant samples of people's expectations on everything from Presidential elections to nuclear reactor safety. Nevertheless, surveys are rarely used to elicit time series data for routine use in econometric modeling because the cost in money and manpower can be substantial and there is perhaps some suspicion that the rule of one-man-one-"vote" may not apply in an economic environment.

Whatever the reason, the use of some device to substitute for surveys is the preferred method of the professional economist. This proxy may be used for expectations of exogenous variables in simultaneous equation systems, or simply for forecasting the variable of interest. The emphasis that Nerlove placed on examination of the stochastic environment makes the theory of quasi-rational expectations an objective and replicable method of



specifying expectation models which does not suffer from the ad hoc nature of distributed lag models nor the severe informational requirements of rational expectations models. Nerlove, Grether and Carvalho discuss the theoretical conditions under which quasi-rational and fully rational expectations are identical. However, it is important to note that this study does not test whether quasi-rational expectations are reasonable approximations of fully rational expectations. Rather, this study attempts to test whether quasi-rational expectations are a reasonable approximation of actual, observed behavior. To that extent, the results should be of benefit to the applied economist in search of a practical alternative to direct elicitation of agents' expectations.

The decision to conduct a "laboratory" experiment was due to two problems that arise in experimenting under so-called "real-world" conditions. The first problem is the issue of whether agents should be exposed to data on a real variable or on a hypothetical variable before asking them to formulate their expectations of future realizations of that variable. Other things being equal, the use of a real variable such as the price of corn would make the results of the study of immediate interest to participants in the corn market. But it is well known that a minimum of 30-40 "realizations" (e.g., forecasts in real time) is necessary for reliable time series analysis (Newbold and

Granger). Only daily or weekly data would make such a time frame feasible. Such daily or weekly periods would have to be equally relevant to the agents involved: incentives to form and reveal expectations would have to be the same from one period to the next. Furthermore, the variable of interest would have to be generated by a stochastic process with at least some subjective prospect of "stability": a change in government policy which drastically affects the information set could make the results ambiguous, irrelevant, or misleading.

The second problem, closely linked to choice of variable, is choice of "agent". Economic theory is rarely specific about the characteristics of the "economic agent" which are relevant to the situation. Indeed, it is perhaps a strong point of the discipline that predictions based on theory frequently hold over a vast range of environments, some of which are not even populated by humans (e.g. animal subjects in experiments by Battalio et al. 1981). The minimum requirements for a microeconomic environment, one in which the subjects presumably are "economic agents" by definition, have been proposed by Smith (1982). These requirements are surprisingly undemanding and include only the conditions of Nonsatiation (subjects prefer more goods to less) and Saliency (the reward is clearly associated with the task).

The problems associated with "real-world" agents are admittedly logistical. It can be expensive and time-consuming to elicit expectations from large numbers of busy people in distant places. But the standards of the profession do not exist for the purpose of alleviating the researcher's logistical problems, so it is necessary to justify systematically the trade-off of "real-world agents" (e.g., agricultural producers) for "agents in a microeconomic environment" (e.g., students).

Besides logistics, the primary advantage is that a greater degree of experimental control can be exercised by using non-specific agents. When used in combination with a hypothetical data series, these agents can be assumed to have no *a priori* knowledge of the stochastic process, nor to have access to any other information about the process other than that provided by the experimenter. In addition, incentives to participate can be more closely matched with the difficulty of the task and the opportunity costs of participation.

Another advantage is that the expected reward for revealing a subjective probability formed in period  $i$  about the outcome likely to occur in period  $i+1$  is the same for all agents as that formed in some other period  $j$  about the outcome in period  $j+1$ . Contrast this with the day-to-day relevance of corn prices to producers for whom the choice of marketing strategies may include forward contracting,

hedging, cash sale at harvest, and outright speculation. Furthermore, the method of aggregating results becomes an important consideration. Should larger producers have more weight attached to their expectations than smaller producers?

Finally, the issue of inductive inference, or the ability to generalize from specific results to more general cases, is still unsettled when one considers the decisions that have to be made concerning the relevant population and how to sample it. Do results from one region or occupation generalize to all others? How large a sample is necessary? Should the sample be random or stratified?

The purpose of this study was to conduct a "first-pass" test of the theory of quasi-rational expectations over a variety of stochastic processes and with as much experimental control as possible. It was hoped that this approach would at least yield some confirmation or refutation of the theory under specific, replicable conditions and that in either case some further articulation of the theory in an empirical context would be possible. The method of experimental economics was perceived to be an efficient way to gain experience with the decision problem and its dynamic environment--one with good prospects for directing further research.

The objectives of this research were to test Nerlove's theory under laboratory conditions which controlled the

stochastic and behavioral environment in such a way that agents' responses to the treatment series could be interpreted unambiguously. In this analysis, an aggregate of agents' forecasts was subjected to three statistical tests which constitute Nerlove's criteria for quasi-rationality. An alternative hypothesis is that agents form expectations using some heuristic behavioral mechanism (such as adaptive expectations) in apparent disregard or ignorance of the stochastic structure of the series.

### **Method**

Conditions of this study required that subjects display linear utility for money in the same range of rewards as would be used to elicit their forecasts. The method described by Harrison, which uses a binary lottery and the demand-revealing value scale developed by Becker, DeGroot and Marschak, was used to screen subjects on the basis of their utility functions. From an initial pool of 113 students recruited from six undergraduate classes in agricultural economics and sociology, 19 were identified as risk neutral using Harrison's method.

An interactive computer game called FORECAST was developed to elicit subjects' probability assessments of future events. Elicitation of the entire probability distribution differs from most previous studies of forecast

performance which typically ask only for point forecasts. The general features of the program relevant to this study are described in the following paragraphs.

Subjects assembled in a computer laboratory and were read a set of instructions describing the forecasting task, the reward mechanism, and the operation of their computer terminal (Appendix B). They then reviewed a "historical series" of positive and negative integers which was displayed on their computer screen and on tables and graphs. From this information alone, each subject was asked to produce a forecast, in probability form, of the likelihood that the next number would fall in one or the other of eight "ranges" (identified by the dotted lines in the graph of the series, see Appendix B). For example, a subject might produce the following forecast in the first period: no chance of the number falling in ranges 1, 2, 3, or 8; a 20% chance in each of the ranges 4, 5, and 6; and a 40% chance of the number falling in range 7.

After the probability forecast was entered into the computer, the actual outcome was revealed and each subject received a monetary reward (tallied on the screen) which reflected the accuracy of his forecast according to the quadratic scoring rule:

$$\$G = \$0.2 \left[ 2 - \sum_{i=1}^n (p_i - d_i)^2 \right]$$

where  $\$G$  is gain (payment) for each forecast period,  $p_i$  is the probability assigned by the subject to the  $i^{\text{th}}$  event,  $d_i$  is a binary variable equal to unity if the  $i^{\text{th}}$  event is the one that actually occurs (and equal to zero otherwise), and  $n$  is the number of ranges. For subjects with linear utility of rewards this scoring rule is one of the family of "proper" scoring rules which have the property that they do not provide subjects with any dominant strategy other than that of revealing the probabilities they actually believe (Stael von Holstein).

Subjects continued to produce one-step-ahead forecasts in this way for up to 58 periods, after which they were paid the sum of their earnings. An experiment typically lasted about 140 minutes and subjects earned \$9-17 each, depending on the experiment.

#### **Monte Carlo Series Generation**

Five stochastic processes were chosen for study. In general, these processes were selected so as to include simple processes (i.e., of low order), autoregressive and moving average processes, "classic" processes often invoked in the literature on expectations, and a process with cyclical components. Characteristics of the stochastic process and experimental conditions are summarized in Tables 3.1-3.5. Graphs of the actual data are displayed in Figures 3.1-3.5. The processes used, their rationale, and the experimental protocol are described below in the order

**Table 3.1. Specification of Monte Carlo Generator  
and Ex Post ARIMA Model of Data for AR1 Experiment**

	Generating Process	Model of Data	
Observations	----	1-40	1-80
Representation	AR	AR <sup>a</sup>	AR <sup>a</sup>
Order	1	1 <sup>b</sup>	1 <sup>b</sup>
Coefficients:			
constant (t)	0	44.00 (0.73)	63.27 (1.33)
first lag (t)	0.85	0.45 (3.19)	0.69 (8.01)

<sup>a</sup>Determined by Box-Jenkins identification methods.

<sup>b</sup>Determined by FPE criterion (Akaike).



**Table 3.2. Specification of Monte Carlo Generator and Ex Post ARIMA Model of Data for AR2 Experiment**

	Generating Process	Model of Data	
Observations	----	1-60	1-110
Representation	AR	AR <sup>a</sup>	AR <sup>a</sup>
Order	2	2 <sup>b</sup>	2 <sup>c</sup>
<b>Coefficients:</b>			
constant (t)	0	22.29 (0.66)	31.21 (1.30)
first lag (t)	1.3	1.11 (8.44)	1.06 (11.09)
second lag (t)	-0.4	-0.23 (-1.73)	-0.21 (-2.20)

<sup>a</sup>Determined by Box-Jenkins identification methods.

<sup>b</sup>Restricted to be of order 2.

<sup>c</sup>Determined by FPE criterion (Akaike).

**Table 3.3. Specification of Monte Carlo Generator  
and *Ex Post* ARIMA Model of Data for RW Experiment**

	Generating Process	Model of Data	
Observations	----	1-60	1-115
Representation	White Noise	AR <sup>a</sup>	AR <sup>a</sup>
Order of AR Process	0	1 <sup>b</sup>	1 <sup>b</sup>
Order of Integration	1	0	0
Coefficients:			
constant (t)	-	-22.61 (-0.26)	-16.15 (-0.31)
first lag (t)	-	0.90 (13.72)	0.91 (25.55)

<sup>a</sup>Determined by Box-Jenkins identification methods.

<sup>b</sup>Determined by FPE criterion (Akaike).

**Table 3.4. Specification of Monte Carlo Generator  
and Ex Post ARIMA Model of Data for AE Experiment**

	Generating Process	Model of Data	
Observations	----	1-60	1-110
Representation	MA	MA <sup>a</sup>	MA <sup>a</sup>
Order of MA Process	1	1 <sup>b</sup>	1 <sup>b</sup>
Order of Integration	1	1	1
Coefficients:			
constant (t)	0	20.28 (0.64)	21.34 (0.90)
MA first lag (t)	0.6 <sup>c</sup>	0.56 <sup>c</sup> (5.02)	0.50 <sup>c</sup> (5.91)

<sup>a</sup>Determined by Box-Jenkins identification methods.

<sup>b</sup>Determined by FPE criterion (Akaike).

<sup>c</sup>Sign of coefficient follows notation convention of Box and Jenkins.

**Table 3.5. Specification of Monte Carlo Generator  
and Ex Post ARIMA Model of Data for AR4 Experiment**

	Generating Process	Model of Data	
Observations	----	1-60	1-118
Representation	AR	AR <sup>a</sup>	AR <sup>a</sup>
Order of AR Process	4	4 <sup>b</sup>	4 <sup>b</sup>
Coefficients:			
constant (t)	0	-53.84 (-0.97)	22.30 (0.72)
first lag (t)	0.3	0.13 (1.18)	0.30 (4.43)
second lag	0.0	0.0 <sup>b</sup>	0.0 <sup>b</sup>
third lag	0.0	0.0 <sup>b</sup>	0.0 <sup>b</sup>
fourth lag (t)	0.6	0.64 (5.63)	0.63 (9.11)

<sup>a</sup>Determined by Box-Jenkins identification methods.

<sup>b</sup>Restricted to be a subset model by FPE criterion (Akaike).

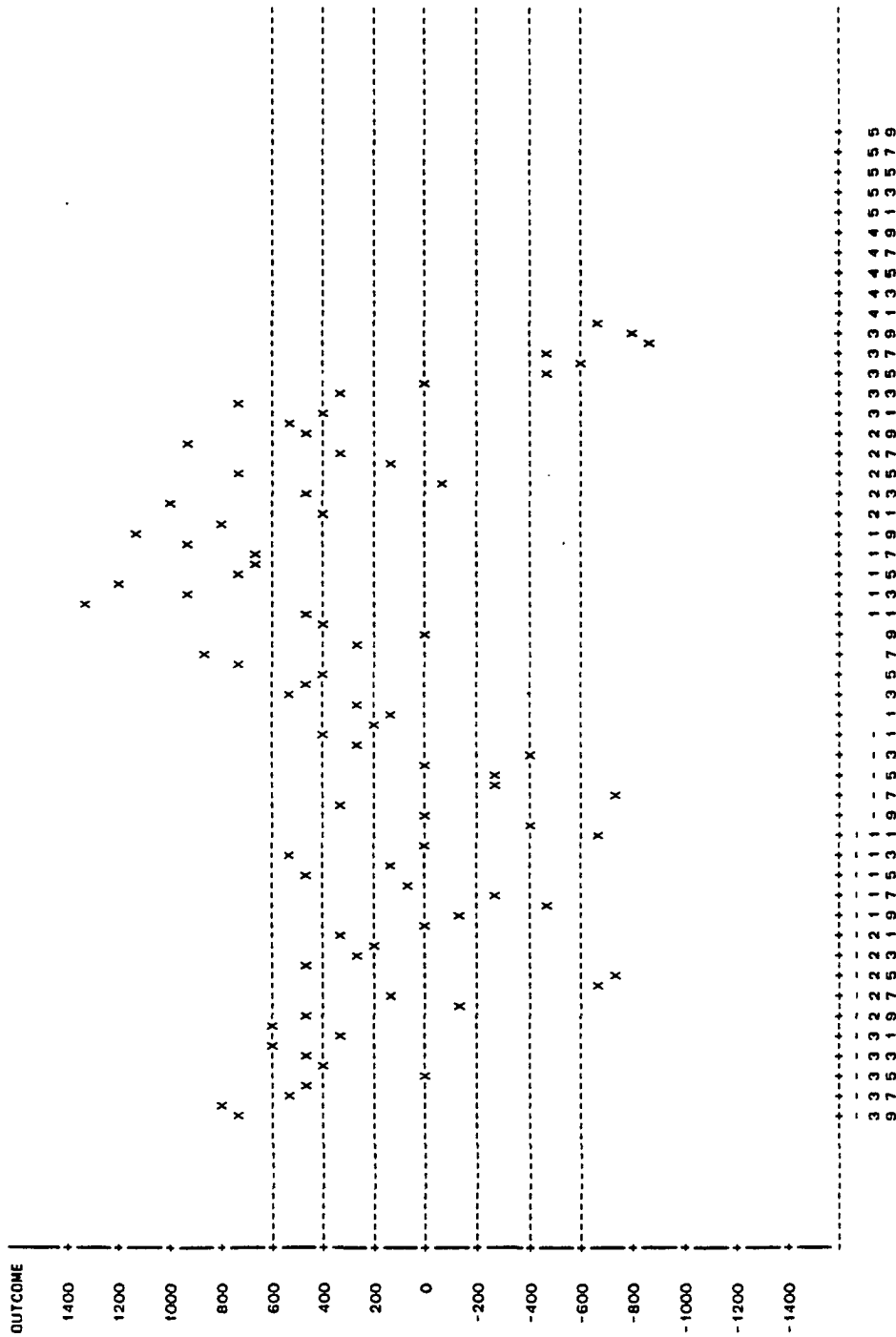


Figure 3.1. Graph of Series Used in AR1 Experiment

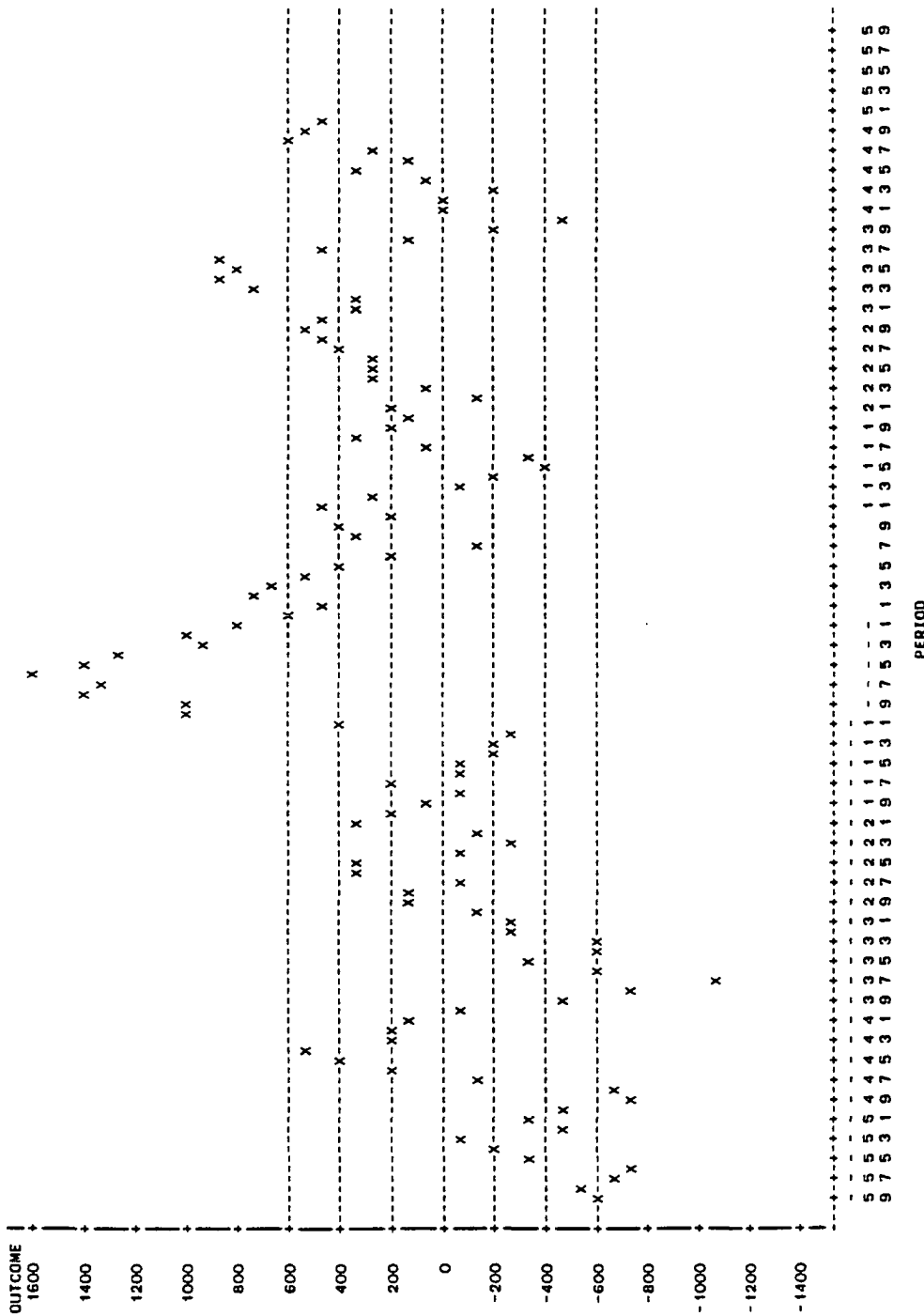


Figure 3.2. Graph of Series Used in AR2 Experiment

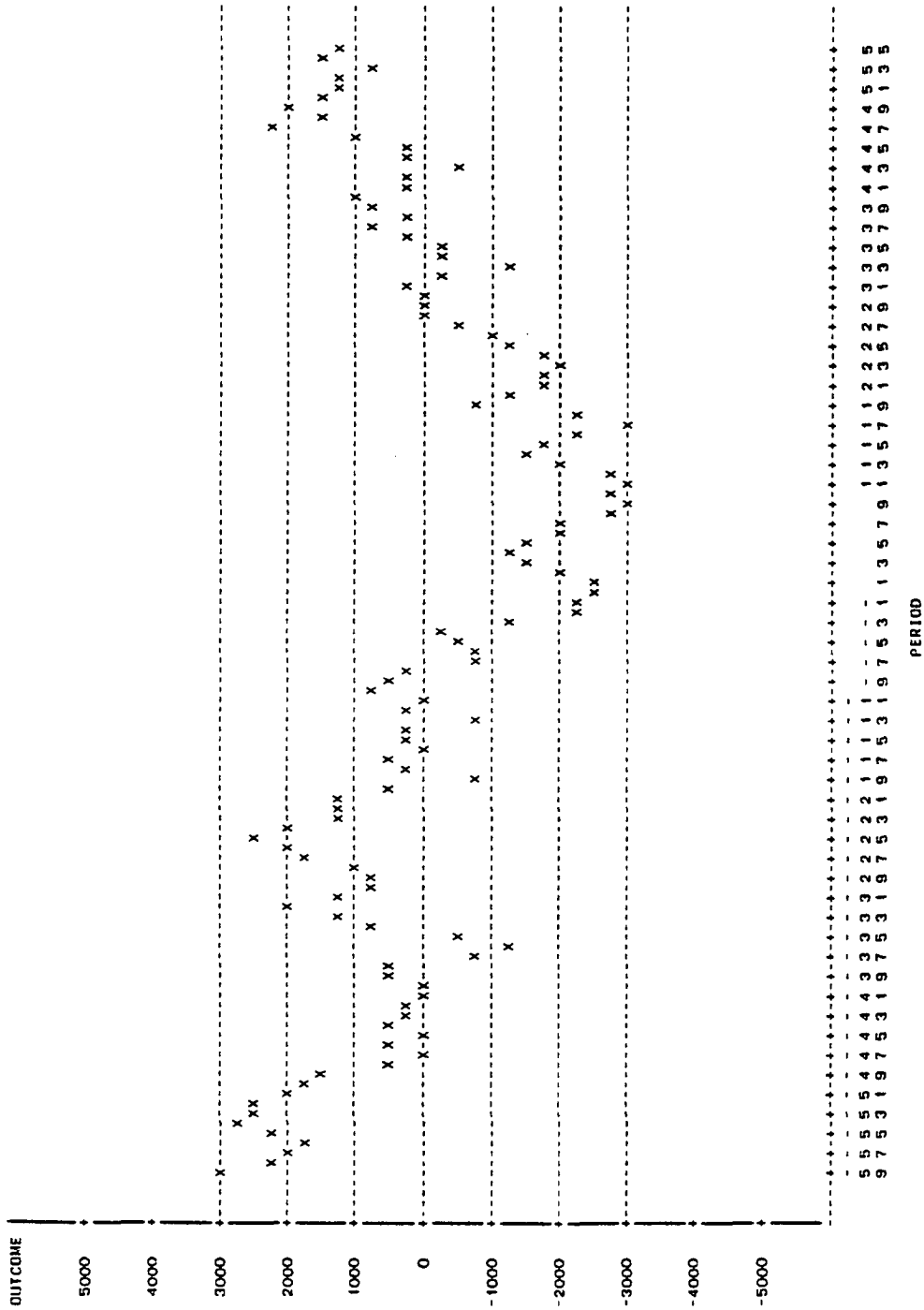


Figure 3.3. Graph of Series Used in RW Experiment

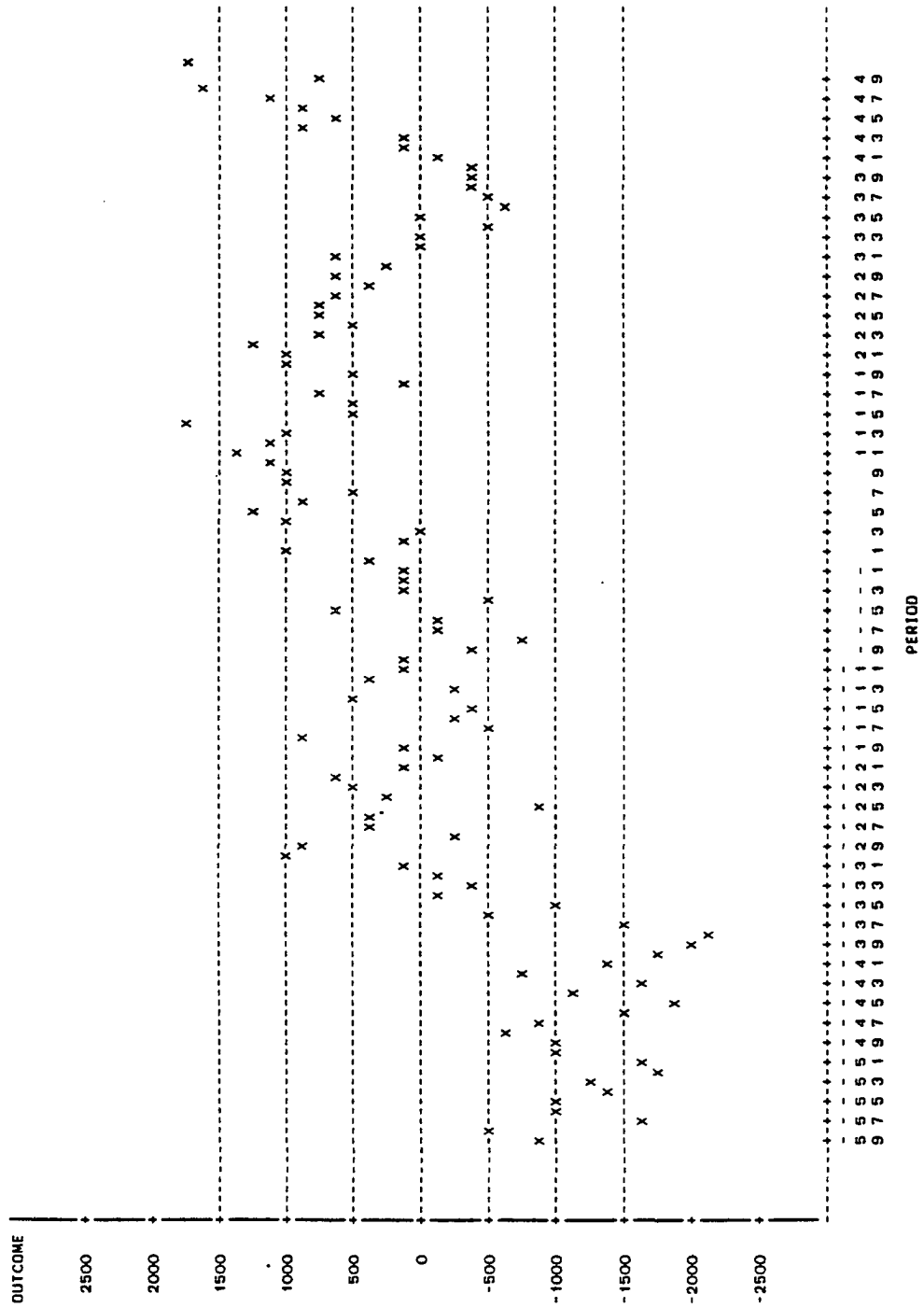


Figure 3.4. Graph of Series Used in AE Experiment





they were used in experiments. Rather than use the more cumbersome notation ARIMA(p,d,q) of Box and Jenkins, the processes are designated by a shorthand label for convenience.

The first experiment used an autoregressive process of order one, designated "AR1", as the treatment effect. The standard formula for this process is:

$$P_t = \phi P_{t-1} + e_t$$

where  $P_t$  is a realization in period  $t$ , and  $e_t$  is distributed  $N(0, \sigma)$ . Subjects reviewed 40 historical data points and forecasted 40 more. The AR1 process is a relatively simple one which describes many variables of economic interest. It is the parent process or general case of several historically popular specifications such as the cobweb model, the random walk, and the "naive" forecast (Theil).

The second experiment used an autoregressive process of order two, and is designated "AR2". The formula is:

$$P_t = \phi_1 P_{t-1} + \phi_2 P_{t-2} + e_t$$

This process was chosen to resemble one of empirical relevance to the overall research program, that of monthly grain sorghum prices in the Corpus Christi region of Texas. The signs of the coefficients were the same, as were the approximate magnitudes, except that the coefficient on the second lag in the hypothetical series was made slightly larger so as to increase the likelihood of its estimate

being significantly different from zero in small samples. Due to concern about having an inadequate number of realizations available for reliable time series analysis, the number of historical data points was increased to 60 and the number forecasted was increased to 50.

The third experiment was designed to be a random walk and is therefore designated "RW". The formula is:

$$P_t = P_{t-1} + e_t$$

The random walk is commonly observed in price series and is associated with the "efficient market hypothesis" (Fama). It has the characteristic that the best predictor of what will happen in the coming period is the outcome that occurred in this period. Thus, from a modeling standpoint, if last period's outcome is subtracted from the outcome in this period, the only statistical property remaining in that "differenced" series is purely random error, or in engineering jargon, "white noise". In the RW experiment the historical data consisted of 60 points, and 55 points were forecasted.

The fourth experiment represented the classic specification of adaptive expectations and is thus designated "AE" rather than the more general or parent case of the exponentially weighted moving average, or ARIMA(0,1,1). The formula is:

$$P_t = P_{t-1} + \theta e_{t-1} + e_t$$

What made this a "classic" specification in a historical sense was that the coefficient on the moving average term was restricted to lie between zero and unity in order to satisfy some notion of behavioral regularity based on a convex combination of observed value and forecast. The general case allows the coefficient to range between -1 and +1 and is concerned only with meeting invertibility conditions (Bessler). This specification also served as an indirect test of the behavioral hypothesis that agents always formulate adaptive expectations regardless of the regularities in the historical data.

The fifth and final experiment attempted to incorporate a "quarterly" or "seasonal" component in the process by using an autoregressive process with a non-zero coefficient on lags one and four. The formula is:

$$P_t = \phi_1 P_{t-1} + \phi_4 P_{t-4} + e_t$$

Even though this process is designated "AR4" and would be described as an ARIMA(4,0,0), it is more correctly termed a "subset AR4" since the coefficients on the second and third lags were restricted to be zero in the generating function. The actual data series was expected to have a weak first order component and a strong fourth order component, behaving in the same fashion as the "peaks and troughs" found in monthly data with a strong quarterly component.

Along with the characteristics of the five Monte Carlo generating functions, Tables 3.1-3.5 show the statistical

models subsequently derived from a finite number of realizations.

In the three autoregressive processes (AR1, AR2, and AR4) the actual series was confirmed to be covariance stationary. Also, an effort was made to control the amount of random variation in the series by specifying a "target variance" for the generating function such that one standard deviation around the mean ("target mean" = 0) was equivalent to having about two-thirds of the realized points lie between the two end ranges (i.e. between ranges 1 and 8 in the graphs). The "real-world" counterpart of this concept is that, while in principle the ranges of many series may be very large (possibly infinite), in practice there is a subdividable range of decision-making among relevant alternatives, and there are two end ranges where only one choice exists (an example might be "less than one inch of annual rainfall" at one end, and "more than six feet" at the other).

Neither the random walk (RW) nor the adaptive expectations (AE) series was constrained to be covariance stationary in the levels, although invertibility conditions were met for the AE series.

#### **Optimal Statistical Model**

Tables 3.1-3.5 show only the best ARIMA model for each series, based on the data available. In the construction of each model, standard time series methods were applied.

Autocorrelation and partial autocorrelation analyses were used in addition to Akaike's criterion of Final Prediction Error (FPE) in order to identify the best model of the data. Forecasts from these models were compared to those of actual subjects in the tests of quasi-rationality.

The methods described by Box and Jenkins for "identifying" the appropriate model of univariate time series data can often be quite subjective when used on relatively small amounts of data. In this study it was necessary to evaluate more than one possible ARIMA model in four of the five experiments. In the AR1 experiment an ARIMA(1,0,0) was the only process clearly identified. In the AR2 experiment, because of the "weak" second order coefficient in the generating function, ARIMA(1,0,0) and (2,0,0) models were both possible candidates for the optimal statistical model. Both models were tested against subjects' forecasts and against each other. Neither model significantly outperformed the other in the test of mean squared error of forecast (described later) and the conclusions regarding quasi-rationality of subjects' forecasts were the same for both models, so only the results from the (2,0,0) model are reported here.

The identification of the RW series was ambiguous as to whether it was an ARIMA(0,1,0) or a (1,0,0). Using the FPE criterion, the (0,1,0) model was preferred on the basis of in-sample fit. But the test of mean squared error (MSE)

indicated that the (1,0,0) model significantly outperformed the (0,1,0) model in out-of-sample forecasts. Since the conclusions regarding quasi-rationality were the same regardless of the model chosen, only the results from the (1,0,0) model are reported here.

The AE series was strongly identified as an ARIMA(0,1,1) but there seemed to be some possibility of a (2,0,0) specification as an alternative. The latter was evaluated and rejected on the basis of the FPE criterion and the MSE test. Furthermore, the forecasts from the (2,0,0) model were biased while those from the (0,1,1) were not.

As mentioned earlier, the AR4 series was generated as a subset ARIMA(4,0,0) with non-zero coefficients only on the first and fourth lags. The actual data series was strongly identified as an ARIMA(4,0,0) and the FPE criterion favored the subset model over a "full model" (i.e. one with non-zero coefficients at all four lags). The subset model was also the significantly better forecaster. Neither model was unbiased in this small (apparently *too small*) sample.

After "identification", the next step in the traditional modeling procedure of Box and Jenkins is "estimation". In the case of each of the series described above, model identification was based on the same set of historical data given to subjects prior to forecasting.

The order of the process and degree of differencing were also fixed at this point. So while the model was re-estimated with each new realization (and the next forecast was made using new coefficients based on all available data) the representation (AR or MA), degree of differencing, and order of the process remained the same over the forecasting period.

The final step in Box-Jenkins modeling is "diagnostic checking". This is done in order to establish that no patterns of regularity remain in the forecast errors-- patterns which might be used to improve forecast performance. The Q-statistic developed by Box and Pierce was used as the test criterion. All of the models passed this test.

#### **Aggregation of Subjects' Forecasts**

The theory of quasi-rational expectations specifically assumes that, at least to a first approximation, economic agents "respond to conditional expectations of the variables rather than to higher moments" (Nerlove 1972, p.231). Although Nerlove acknowledged the possibility of having to obtain conditional second moments as well as conditional means in order to obtain fully optimal solutions (1972, p.233), he did not articulate the theory further to include this as a condition for quasi-rationality. So, even though the elicitation method used in this study provides the entire distribution of each



forecast from each subject, the tests of the theory were conducted using just the mean of each distribution.

Moreover, from a practical standpoint the mean of any individual's forecast is not generally as relevant to the end user of expectations data as some aggregate of all agents' forecasts. For this aggregation a simple average of the means of the individual distributions was chosen.

The mean of each subject's distribution was derived by multiplying the probability assigned to that range by the midpoint of the range and then summing over the eight ranges. For purposes of calculating the midpoints of ranges 1 and 8, the ends of these ranges were taken to be respectively the smallest and largest numbers listed on the ordinates of Figures 3.1-3.5. The results that follow are based on calculations using a simple average of these means at each of the forecast periods.

### **Test Criteria**

The real tests of Nerlove's theory are found in an examination of the period-by-period error between the data and the two forecasters: model and man. The criteria for quasi-rationality (Nerlove 1981) require that forecasts be: (1) unbiased (mean error=0), (2) have no systematic components in the forecast errors (white noise residuals), and (3) produce forecasts that are indistinguishable in mean squared error from those produced by a minimum-MSE predictor (such as the ARIMA model in this case). All

three tests are based on an examination of forecasts errors, defined as:

$$E_{it} = A_t - F_{it}$$

where  $E_{it}$  is the error made by the  $i^{\text{th}}$  forecaster (individual subject, aggregate, or ARIMA model) in period  $t$ ,  $A_t$  is the actual number realized in period  $t$ , and  $F_{it}$  is the point forecast (represented by the mean of the subjective probability distribution) made by the forecaster in the previous period that represents the number expected to occur in period  $t$ .

The null hypothesis of the test for bias is that the mean of all errors made by a forecaster is not significantly different from zero, based on a standard  $t$ -test.

A common test for white noise in the errors is based on the  $Q$ -statistic (Box and Jenkins, p. 291-292). For this analysis the  $Q$ -statistic was calculated at the 24<sup>th</sup> sample autocorrelation. The null hypothesis was that no significant regularities exist in the errors at the 24<sup>th</sup> autocorrelation. The significance level was determined approximately from a chi-square table.

The mean squared error test is described by Ashley, Granger, and Schmalensee (see also Ashley; and Brandt and Bessler). The details of the methodology need not be repeated here, but some elaboration with respect to issues

encountered in application is necessary. Ashley et al. define the following terms:

$$\Delta_t = (E_{it} - E_{jt})$$

$$\Sigma_t = (E_{it} + E_{jt})$$

The regression equation upon which the test is based is:

$$\Delta_t = \beta_0 + \beta_1 [\Sigma_t - m(\Sigma_t)] + u_t$$

where  $m(\Sigma_t) = 1/T \sum_{t=1}^T (\Sigma_t)$  and  $u_t$  is random error with

classical assumptions. In this study, the  $i^{\text{th}}$  forecaster was always a human subject or aggregate of subjects, while the  $j^{\text{th}}$  forecaster was an ARIMA model. Therefore the alternative to the null hypothesis was always that the model was a better forecaster than human subjects.

In all cases the mean error of the  $j^{\text{th}}$  model,  $1/T \sum_{t=1}^T (\Sigma_{jt})$ ,

was positive. In the cases where the mean error of the  $i^{\text{th}}$

subject or aggregate,  $1/T \sum_{t=1}^T (\Sigma_{it})$ , was negative the

correction described in Brandt and Bessler (p.247) was applied.

In comparing each subject in each experiment against competing ARIMA models, the MSE test was applied 74 times. Frequently, significant first order autocorrelation in the residuals was indicated by the Durbin-Watson statistic. In

a few cases, possible additional time series properties were indicated by large sample autocorrelations and partial autocorrelations. To deal with first order autocorrelation in a systematic way, the maximum likelihood correction procedure of Beach and MacKinnon, (as programmed on RATS, the time series analysis software by Doan and Litterman), was applied in every case. Consequently, the coefficients from the MSE tests are all derived from this estimation procedure. In a few cases higher order autocorrelation corrections using autoregressive models were applied but these did not affect the conclusions, and so the results have been omitted.

There is one technical point that needs elaboration. The null hypothesis of the MSE test is that there is no significant difference in MSE between forecasters. In Ashley et al. the procedure for accepting or rejecting the hypothesis is based on the signs and t-statistics of  $\beta_0$  and  $\beta_1$  in the previous regression equation. Not surprisingly, the sign of a coefficient that is not significantly different from zero can be sensitive to various estimation procedures used to correct for first order autocorrelation, and this affects any hypothesis test which is based on the signs of those coefficients. Therefore, the joint test of significance afforded by the F-statistic is probably more appropriate. In the various tables and in the conclusions, the p-value of the one-

tailed F-test was taken to be the statistic of preference in testing the null hypothesis, especially in cases where the method of Ashley et al. gave ambiguous conclusions.

## Results

### General

Table 3.6 lists the results for the aggregate of subjects' forecasts from each of the five experiments. Examination of the table suggests that by fairly conservative standards of significance the aggregate forecast was indistinguishable from that of the ARIMA model in the AR1, AR2, and RW experiments. These results support the hypothesis of quasi-rationality. Conversely, the ARIMA model was a significantly better forecaster than the aggregate in the AE and AR4 experiments and thus would have been a poor substitute for subjects' aggregate expectations for those series.

The case was made earlier that only the aggregate results are of practical importance since we are generally not concerned with finding substitutes for a single agent's expectations--it makes more sense just to ask him. In the subchapters that follow, the results from individual subjects are briefly discussed. The coded individual results are presented primarily to provide some insight as to how the phenomenon of aggregation might operate. No

**Table 3.6. Performance of Aggregate Forecast in Five Experiments**

	<b>AR1</b>	<b>AR2</b>	<b>RW</b>	<b>AE</b>	<b>AR4</b>
<b>Bias Test:</b>					
Mean	90.03	4.76	54.03	71.87	124.08
[ p-value]	[0.133]	[0.891]	[0.494]	[0.264]	[0.020]
<b>White Noise Test:</b>					
Q(24)-statistic	13.7+	28.3?	20.4+	25.0+	28.3?
<b>MSE test:</b>					
B(0)	29.30	-3.35	-11.07	66.61	38.96
[t-statistic]	[ 1.13]	[ -0.20]	[ -0.26]	[ 1.57]	[ 0.92]
B(1)	-0.005	0.033	0.022	0.060	0.134
[t-statistic]	[ -.27]	[ 1.03]	[ 1.40]	[ 2.56]	[ 2.46]
<b>1-tail F-test:</b>					
p-value	0.128	0.145	0.092	0.004	0.010

+ = p>0.25

? = 0.25>p>0.10

attempt is made to explain or predict the behavior of individuals in these experiments. Indeed, information on factors that might be of relevance in such an endeavor, such as socio-economic characteristics, is unavailable due to the confidentiality conditions required under the "Guidelines for Human Subjects in Research" (Texas A&M University).

For convenience, remarks in the following discussion are made about "failing" or "passing" one or more of the tests of quasi-rationality. These interpretations are based loosely on a significance level of around 0.10. The tables of results list the exact statistics and p-values. Also, to avoid confusion about "failing" the MSE test it should be noted that in no case did a subject ever significantly outperform the ARIMA model. The subject numbers used in the tables (e.g. S3) do not necessarily designate the same person between one experiment and another.

#### **AR1 Experiment**

Of the eight subjects participating in this experiment, three subjects (S1, S5, and S8) passed all tests for quasi-rationality (Table 3.7). All five of the remaining subjects had significantly higher mean squared forecast error than the ARIMA(1,0,0) model, plus four subjects among this group produced forecasts that were biased.

**Table 3.7. Individual Subject and Aggregate Results from AR1 Experiment**

	SUBJECT								Aggregate
	S1	S2	S3	S4	S5	S6	S7	S8	
<b>Bias Test:</b>									
Mean	7.08	134.05	131.15	111.45	33.40	201.95	114.00	1.38	90.03
[ p-value]	[0.906]	[0.064]	[0.025]	[0.101]	[0.669]	[0.006]	[0.084]	[0.980]	[0.133]
<b>White Noise Test:</b>									
Q(24)-statistic	11.2+	18.3+	11.2+	10.4+	32.0?	18.4+	25.5+	8.5+	13.7+
<b>MSE test:</b>									
B(0)	-56.08	73.30	71.52	60.33	-47.92	137.86	50.10	-41.47	29.30
[t-statistic]	[-0.55]	[ 1.70]	[ 2.18]	[ 1.06]	[-0.73]	[ 4.96]	[ 2.10]	[-0.41]	[ 1.13]
B(1)	-0.273	0.599	-0.016	0.048	0.025	0.065	0.022	0.008	-0.005
[t-statistic]	[-0.57]	[2.03]	[-0.66]	[+1.69]	[0.83]	[ 2.07]	[0.73]	[0.16]	[-0.265]
<b>1-tail F-test:</b>									
p-value	0.183	0.010	0.022	0.038	0.135	6E-06	0.025	0.228	0.128

+ = p>0.25  
 ? = 0.25>p>0.10



**AR2 Experiment**

Ten subjects participated in this experiment but not one passed all three tests for quasi-rationality (Table 3.8). Seven were unbiased (S1, S2, S6, S7, S8, S9 and S10), but two of these (S1 and S8) failed the test for white noise (along with S4, who was also biased). All subjects were significantly poorer forecasters than the model. Nevertheless, the aggregate of these ten subjects was well represented by the optimal statistical model.

**RW Experiment**

Table 3.9 shows that six of the eight subjects in this experiment (S1, S2, S3, S4, S5, and S7) passed the test for bias. Five of these passed the test for white noise residuals as well; S5 failed it. One subject (S6) failed the bias test but passed the white noise test. The only subject who passed the MSE test (S8) failed the bias and white noise tests. So while none of the subjects passed all three tests, the aggregate did.

Although this experiment was conceived to be a test of a random walk, or ARIMA(0,1,0) process, the small number of realizations reflected an ARIMA(1,0,0) process and this was the specification used for the forecasting model. Consequently, this experiment serves as a "replication" of the AR1 experiment and it is gratifying to note that while the individual results concerning quasi-rationality were weaker, the aggregate results were replicated. A check of

**Table 3.8. Individual Subject and Aggregate Results from AR2 Experiment**

	SUBJECT										Aggregate
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
<b>Blas Test:</b>											
Mean	7.20	5.82	191.24	171.42	80.08	5.34	21.94	60.30	3.44	18.26	4.76
[ p-value]	[0.875]	[0.878]	[8E-5]	[0.002]	[0.031]	[0.917]	[0.644]	[0.216]	[0.932]	[0.632]	[0.891]
<b>White Noise Test:</b>											
Q(24)-statistic	33.1*	25.7+	17.1+	32.2*	24.4+	27.4?	16.4+	36.2**	18.2+	23.1+	28.3?
<b>MSE test:</b>											
B(0)	-1.66	-3.41	182.42	163.77	67.68	4.81	13.22	58.05	-4.54	9.18	-3.35
[t-statistic]	[-0.03]	[-0.18]	[ 3.09]	[ 3.77]	[ 1.61]	[0.11]	[0.41]	[ 1.00]	-[ .16]	[0.50]	[-.20]
B(1)	1.100	0.115	1.220	0.309	-0.034	0.201	0.245	1.605	0.107	0.096	0.033
[t-statistic]	[3.51]	[ 2.57]	[ 3.96]	[ 3.93]	[-0.54]	[ 3.60]	[ 3.93]	[ 5.06]	[ 3.01]	[ 2.98]	[ 1.03]
<b>1-tail F-test:</b>											
p-value	0.001	0.011	0.00001	2E-06	0.061	0.0008	0.0003	6E-06	0.004	0.004	0.145

+ =  $p > 0.25$   
 ? =  $0.25 > p > 0.10$   
 \* =  $0.10 > p > 0.05$   
 \*\* =  $0.05 > p$

**Table 3.9. Individual Subject and Aggregate Results from RW Experiment**

	SUBJECT								Aggregate
	S1	S2	S3	S4	S5	S6	S7	S8	
<b>Bias Test:</b>									
Mean	16.60	48.20	98.64	66.02	109.40	161.38	53.80	204.56	54.03
[ p-value]	[0.856]	[0.576]	[0.253]	[0.458]	[0.269]	[0.058]	[0.651]	[0.036]	[0.494]
<b>White Noise Test:</b>									
Q(24)-statistic	21.5+	19.8+	16.6+	24.9+	39.1**	24.8+	8.9+	38.8**	20.4+
<b>MSE test:</b>									
B(0)	-50.02	7.76	46.76	24.15	55.61	132.74	1.19	133.10	-11.07
[t-statistic]	[ -0.74]	[ 0.10]	[ 1.06]	[ 0.28]	[ 0.40]	[ 1.75]	[ 0.01]	[ 0.99]	[ -0.26]
B(1)	0.087	0.085	0.08	0.142	1.325	0.066	1.246	0.007	0.022
[t-statistic]	[ 2.11]	[ 4.36]	[ 2.19]	[ 4.61]	[ 3.64]	[ 2.32]	[ 5.27]	[ 0.15]	[ 1.40]
<b>1-tail F-test:</b>									
p-value	0.022	0.00007	0.015	0.00003	0.001	0.005	4E-06	0.152	0.092

+ = p>0.25  
 \*\* = 0.05>p

an alternative ARIMA(0,1,0) forecasting model gave the same results except that the individual who passed the MSE test in Table 3.9 failed it in the latter case.

#### **AE Experiment**

Seven of the nine subjects in this experiment passed the bias test; S7 and S8 failed it (Table 3.10). All subjects passed the white noise test. Two subjects (S1 and S6) passed the MSE test (as well as the other two tests). However, the aggregate did not pass the MSE test. So, while quasi-rationality was an adequate description of two individuals, it was not suitable for the aggregate.

#### **AR4 Experiment**

Table 3.11 shows that only three out of six subjects were unbiased (S1, S2, and S6), and one of these (S6) did not pass the white noise test. None of the subjects passed the MSE test, nor did the aggregate. Quasi-rationality was on all counts a poor description of behavior in this experiment.

### **Conclusions**

With respect to the conditions of these experiments it seems quite clear from the results that the description of quasi-rationality closely approximates aggregate behavior when the series of interest is relatively "simple", i.e. a low order autoregressive process. This conclusion is

**Table 3.10. Individual Subject and Aggregate Results from AE Experiment**

	SUBJECT									Aggregate
	S1	S2	S3	S4	S5	S6	S7	S8	S9	
<b>Bias Test:</b>										
Mean	14.98	59.8	5.56	46.86	63.36	8.46	219.88	202.56	53.76	71.87
[ p-value]	[0.806]	[0.407]	[0.939]	[0.540]	[0.338]	[0.897]	[0.012]	[0.008]	[0.442]	[0.264]
<b>White Noise Test:</b>										
Q(24)-statistic	11.6+	20.5+	28.5?	18.5+	19.1+	16.4+	26.0+	28.7?	31.8?	25.0+
<b>MSE test:</b>										
B(0)	10.66	52.16	-0.96	41.12	58.89	7.24	204.01	203.65	52.16	66.61
[t-statistic]	[0.47]	[0.85]	[-0.01]	[0.52]	[1.36]	[0.06]	[2.12]	[5.57]	[0.96]	[1.57]
B(1)	-0.018	0.120	1.094	0.151	0.069	0.150	0.225	0.055	0.103	0.060
[t-statistic]	[-0.87]	[ 3.64]	[ 1.58]	[ 3.63]	[ 2.67]	[0.31]	[ 3.95]	[ 2.41]	[ 2.56]	[ 2.56]
<b>1-tail F-test:</b>										
p-value	0.155	0.001	0.074	0.001	0.004	0.238	0.00005	2E-07	0.007	0.004

+ = p>0.25  
 ? = 0.25>p>0.10

**Table 3.11. Individual Subject and Aggregate Results from AR4 Experiment**

	SUBJECT						Aggregate
	S1	S2	S3	S4	S5	S6	
<b>Bias Test:</b>							
Mean	44.17	69.96	102.09	168.66	295.07	64.88	124.08
[ p-value]	[0.436]	[0.209]	[0.054]	[0.006]	[0.0001]	[0.223]	[0.020]
<b>White Noise Test:</b>							
Q(24)-statistic	18.7+	22.5+	16.4+	38.5**	44.9**	32.2**	28.3?
<b>MSE test:</b>							
B(0)	-42.09	-11.57	17.83	83.94	209.33	-18.81	38.96
[t-statistic]	[ -1.07]	[ -0.19]	[ 0.37]	[ 1.80]	[ 3.41]	[ -0.36]	[ 0.92]
B(1)	0.166	0.195	0.143	0.169	0.347	0.154	0.134
[t-statistic]	[ 3.03]	[ 3.11]	[ 2.56]	[ 3.46]	[ 5.03]	[ 2.58]	[ 2.46]
1-tail F-test:							
p-value	0.002	0.003	0.011	0.0003	2E-07	0.010	0.010

+ = p>0.25  
 \*\* = 0.05>p

supported by evidence from the AR1, AR2, and RW experiments.

Unfortunately, not much can be said about more complex series like the AR4 without further experimentation. Perhaps a larger sample of subjects might produce an aggregate forecast vector more closely approximating that of the ARIMA model. Or perhaps a longer series is necessary for unequivocal results to become apparent. After all, the model itself was not unbiased even after 118 realizations. Also, any pattern of "peaks and troughs" is at best reduced to a quarter of the actual realizations in a fourth order process. Furthermore, the perception of "seasons" or "cycles" seems to connote some additional structural information, such as meteorological or biological phenomena. Such knowledge was not available to subjects in the rarefied information environment of the AR4 experiment.

The AE experiment is unusual in that while little can be said about moving average processes and differenced series in general, it is apparent that the behavior characterized as "adaptive expectations" clearly is not a satisfactory description of actual behavior in the aggregate. If it were, we would most expect it to be manifested when the stochastic process is of that type. Since that behavior was not observed in conjunction with its counterpart process, it suggests that such behavior

does not come "naturally" to most people, or perhaps that it is reserved for special circumstances which were not apparent in this research.

Overall conclusions highlight the fact that quasi-rational behavior in some individuals is not sufficient to guarantee it in the aggregate, as the AE results show. On the other hand, the absence of such behavior in all individuals is not sufficient to guarantee its absence in the aggregate, as demonstrated by the AR2 and RW results. Clearly, the phenomenon of aggregation warrants careful examination.

This research would be of limited interest if nothing could be said about generalizing the results to a larger economic setting. Perhaps the most practical and least controversial result is that surprisingly small numbers of agents are required to demonstrate quasi-rationality in the aggregate. It is likely that field validation of the results would require far fewer subjects than are typically surveyed for other research purposes.

Field validation of the results involves two interrelated factors: the series, and the agents. Generalization of the series to a real-world variable introduces what might best be described as a "multivariate information set". Much more information would be available--some of it vital and some of it spurious. Compounding sources of uncertainty would become important, as would a



*priori* information, opinion, experience, and conjecture. Even with all this additional information the basic question is still whether agents forecast *differently* from an ARIMA or small vector time series model. If they forecast differently they may do so because they are *better* forecasters, in which case an appropriate substitute for their expectations may involve larger multi-variate econometric models which incorporate prior information that is not of a time series nature. But agents may also be *worse* forecasters, in which case the forecasting accuracy of econometric models is no longer of relevance--the effectiveness of policy based on a poorly specified model could actually be decreased with an increase in the accuracy of the statistical model. If agents are not strictly quasi-rational, but are at least consistent and predictable in their expectation formation behavior, then a case can still be made for substituting some prediction of their behavior, perhaps based on a lower order process. However, if they are inconsistent, or especially if the stochastic environment is profoundly unstable, then periodic elicitation via survey methods may be the only recourse.

Generalization of the agents is intimately connected to the information set. That subjects in these experiments were unfamiliar with the "variable" may in fact strengthen the conclusions by the argument that, with plentiful

supplementary relevant information, subjects might be even more likely to display quasi-rational behavior. Also a minor case could be made that the signal-to-noise ratio in the series used in these experiments was low in comparison to most real-life situations. That is, most real-life variables do not contain as much random error as was artificially introduced in the generating function. To rigorously pursue this argument requires some assumptions about the decision situation and the relevant "ranges" (counterparts to the range slots in the FORECAST program) across which distinct choices become viable. Suffice it to say that quasi-rational behavior under conditions of relatively low signal-to-noise ratios in the laboratory would again seem to strengthen the conclusions inasmuch as an increase in signal strength should produce an increase in forecast accuracy.

Finally, it is perhaps trivial to point out that the whole concept of quasi-rational expectations, being based so solidly on historical frequencies, is most applicable in situations where markets have some semblance of "efficiency"--where the institutions encourage free exchange of information, low transactions costs and ready arbitrage opportunities. Highly unstable processes which lack historical regularities would not be good candidates for application of the time series methods inherent in the theory of quasi-rational expectations.

**CHAPTER IV**  
**CONCLUSIONS**

The stated objectives of this research were: (1) to articulate the theory of proper scoring rules and test its ability to predict behavior, and (2) to test Nerlove's theory of quasi-rational expectations under certain controlled conditions. The degree to which these objectives have been achieved is left to the reader to decide. The fascination of human behavior science is in the *process*--the intuition from introspection and casual observation, the experience from controlled observation.

Introspection and casual observation suggested that the dominant strategies of improper scoring rules would not be readily apparent to untrained or unsophisticated subjects. Controlled observation suggested that this was indeed the case at least over the first few forecasts. This information can be used to advantage in rewarding agents in field elicitation studies with easily understandable, though possibly improper, reward mechanisms. On the one hand, such generalization of the results is technically complicated by the utility function: the results were conditional on subjects having linear utility functions. On the other hand, experience gained from these experiments encourages the inductive inference

that if subjects with linear utility failed to exploit the dominant strategy of the linear scoring rule, then subjects with log utility functions probably would also overlook a similar strategy with the log scoring rule. The next logical step would then be to propose that the same behavioral mechanism--call it ingenuousness or "unsophisticated gamesmanship"--would operate for any utility function and any reasonable scoring rule over a limited number of forecasts. Consequently the reward mechanism of preference for elicitation of individual expectations could be the linear scoring rule. Much of this speculation is, fortunately, testable.

Introspection and casual observation suggested that complicated time series processes such as an ARIMA(11,0,0) would not be readily perceived by subjects with no *a priori* information about the generating process or its environment. It also seemed too good to be true that all human beings should always use adaptive expectations or the random walk in producing their forecasts. But this kind of inference could be characterized as the setting up of one "straw man" and the knocking down of another. Controlled observation suggested a range of processes over which an ARIMA model might make a ready substitute for the aggregate expectations of agents. Further work along these lines could profitably investigate: (1) the generalizability of the results to more "realistic" settings, and (2) whether a

larger sample of subjects (through the phenomenon of aggregation) could overcome the problem of extracting regularities from more complex series.

## REFERENCES

- Akaike, H. "Fitting Autoregressive Models for Prediction."  
*Ann. Instit. Stat. Math.* 21(1969):243-47.
- Ashley, R. "Inflation and the Distribution of Price  
Changes: A Causal Analysis." *Economic Inquiry*  
19(1981):650-60.
- Ashley, R., C.W.J. Granger, and R. Schmalensee.  
"Advertising and Aggregate Consumption: An Analysis of  
Causality." *Econometrica* 48(1980):1149-67.
- Battalio, R.C., J.H. Kagel, H. Rachlin, and L. Green.  
"Commodity Choice Behavior with Pigeons as Subjects."  
*J. Polit. Econ.* 89(1981):67-91.
- Battalio, R.C., J.H. Kagel, R.C. Winkler, and R.A. Winett.  
"Residential Electricity Demand: An Experimental  
Study." *Rev. Econ. and Stat.* 61(1979):180-89.
- Beach, C.M. and J.G. MacKinnon. "A Maximum Likelihood  
Procedure for Regression with Autocorrelated Errors."  
*Econometrica* 46(1978):51-58.
- Becker, G.M., M.H. DeGroot, and J. Marschak. "Measuring  
Utility by a Single-Response Sequential Method."  
*Behav. Sci.* 9(July, 1964):226-32.

- Bessler, D.A. "Adaptive Expectations, the Exponentially Weighted Forecast, and Optimal Statistical Predictors: A Revisit." *Agr. Econ. Research* 34(1982):16-23.
- Binswanger, H.P. "Attitudes toward Risk: Experimental Measurement in Rural India." *Amer. J. Agr. Econ.* 62(1980):395-407.
- "Empirical Estimation and Use of Risk Preferences: Discussion." *Amer. J. Agr. Econ.* 64(1982):391-93.
- Black, R. "Weather Variation as a Cost-Price Uncertainty Factor as it Affects Corn and Soybean Production." *Amer. J. Agr. Econ.* 57(1975):44-47.
- Bohm, P. "Revealing Demand for an Actual Public Good." *J. Public Econ.* 24(1984):135-51.
- Box, G.E.P., and G.M. Jenkins. *Time Series Analysis: Forecasting and Control* (rev. ed.). Oakland CA: Holden-Day, 1976.
- Box, G.E.P., and D.A. Pierce. "Distribution of Residual Autocorrelations in Autoregressive Integrated Moving Average Time Series Models." *J. Amer. Stat. Assoc.* 65(1970):1509-26.
- Brandt, J.A., and D.A. Bessler. "Price Forecasting and Evaluation: An Application in Agriculture." *J. Forecasting* 2(1983):237-48.

- Campbell, D.T., and J.C. Stanley. *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin Co., 1963.
- David, F.N. *Games, Gods, and Gambling*. New York: Hafner Publishing Co., 1962.
- de Finetti, B. *Theory of Probability, Vol. I*. London: John Wiley & Sons, 1974.
- Doan, T.A., and R.B. Litterman. *User's Manual RATS Version 4.30*. Minneapolis, MN: VAR Econometrics, 1984.
- Einhorn, H.J., and R.M. Hogarth. "Decision Making under Ambiguity." *J. Business* 59(1986):S225-50.
- Fama, E.F. "Efficient Capital Markets: A Review of Theory and Empirical Work." *J. Finance* 25(1970):383-417.
- Fine, T.L. *Theories of Probability*. New York: Academic Press, 1973.
- Gardner, B.L. "Futures Prices in Supply Analysis." *Amer. J. Agr. Econ.* 58(1976):81-84.
- Grisley, W., and E.D. Kellogg. "Farmers' Subjective Probabilities in Northern Thailand: An Elicitation Analysis." *Amer. J. Agr. Econ.* 65(1983):74-82.
- . "Farmers' Subjective Probabilities in Northern Thailand: Reply." *Amer. J. Agr. Econ.* 67(1985):149-52.
- Hampton, J.M., P.G. Moore, and H. Thomas. "Subjective Probability and its Measurement." *J. Royal Stat. Soc. A* 136(1973):21-42.



- Hanemann, W.M., and R.L. Farnsworth. "The Roles of Risk Preferences and Perceptions in the Adoption of Integrated Pest Management." Unpublished paper, Dept. Agr. Econ., Univ. Calif., Berkeley, 1981.
- Harrison, G.W. "An Experimental Test for Risk Aversion." *Economics Letters* 21(1986):7-11.
- Isaac, R.M., K.F. McCue, and C.R. Plott. "Public Goods Provision in an Experimental Environment." *J. Public Econ.* 26(1985):51-74.
- Jiranyakul, K. "Utility Function in the Domains of Gains and Losses: An Experimental Study." Ph.D. thesis, Texas A&M Univ., 1986.
- Kagel, J.H. "Token Economies and Experimental Economics." *J. Polit. Econ.* 80(1972):779-85.
- Kershaw, D., and J. Fair. *The New Jersey Income Maintenance Experiment*. New York: Academic Press, 1976.
- King, R.P., and D.W. Lybecker. "Flexible Risk-Oriented Marketing Strategies for Pinto Bean Producers." *West. J. Agr. Econ.* 8(1983):124-33.
- Knight, T., S.R. Johnson, and R.M. Finley. "Farmers' Subjective Probabilities in Northern Thailand: Comment." *Amer. J. Agr. Econ.* 67(1985):147-48.
- Lin, W., G.W. Dean, and C.V. Moore. "An Empirical Test of Utility vs. Profit Maximization in Agricultural Production." *Amer. J. Agr. Econ.* 56(1974):497-508.

- Marwell, G., and R.E. Ames. "Economists Free Ride: Does Anyone Else? Experiments on the Provision of Public Goods. IV." *J. Public Econ.* 15(1981):295-310.
- Murphy, A.H. "A New Vector Partition of the Probability Score." *J. App. Meteorology* 12(1973):595-600.
- Nerlove, M. "Distributed Lags and Unobserved Components of Economic Time Series." *Ten Economic Studies in the Tradition of Irving Fisher*, ed. W. Fellner and others, pp. 127-69. New York: John Wiley & Sons, 1967.
- . "Lags in Economic Behavior." *Econometrica* 40(1972):221-51.
- . "The Dynamics of Supply: Retrospect and Prospect." *Amer. J. Agr. Econ.* 61(1979):874-88.
- . "Expectations, Plans and Realizations: In Theory and Practice." Discussion Paper No. 511, Northwestern Univ., Evanston, IL, 1981.
- Nerlove, M., D. Grether, and J.L. Carvalho. *Analysis of Economic Time Series: A Synthesis*. New York: Academic Press, 1979.
- Newbold, P., and C.W.J. Granger. "Experience with Forecasting Univariate Time Series and the Combination of Forecasts." *J. Royal Stat. Soc. A* 137(1974):131-146.
- Newton, H.J. *TIMESLAB: A Time Series Computing Laboratory*. Preliminary version, Texas A&M Univ., 1985.

- Norris, P.E., and R. A. Kramer. "The Elicitation of Subjective Probabilities with Applications in Agricultural Economics: A Survey of the Literature." Dept. of Agr. Econ., Virginia Polytechnic Institute and State Univ., Blacksburg, VA, Publication A.E. 56, 1986.
- Officer, R.R., and A.N. Halter. "Utility Analysis in a Practical Setting." *Amer. J. Agr. Econ.* 50(1968):257-77.
- Pechman, J.A., and P.M. Timpane, eds. *Work Incentives and Income Guarantees: The New Jersey Negative Income Tax Experiment*. Washington, DC: The Brookings Institution, 1975.
- Phillips, L., and W. Edwards. "Conservatism in a Simple Probability Inference Task." *J. Exper. Psych.* 72(1966):346-54.
- Plott, C.R. "Experimental Methods in Political Economy: A Tool for Regulatory Research." *Attacking Regulatory Problems: An Agenda for Research in the 1980's*, ed. A.R. Ferguson, pp.117-43. Cambridge, MA: Ballinger Publishing Co., 1981.
- Robison, L.J. "An Appraisal of Expected Utility Hypothesis Tests Constructed from Responses to Hypothetical Questions and Experimental Choices." *Amer. J. Agr. Econ.* 64(1982):367-75.

- Sargent, T.J. "An Economist's Foreward to *Prediction and Regulation by Linear Least-square Methods*." *Prediction and Regulation by Linear Least-square Methods*, P. Whittle, pp. v-vii. Minneapolis, MN: Univ. Minnesota Press, 1983.
- Savage, L.J. "Elicitation of Personal Probabilities and Expectations." *J. Amer. Stat. Assoc.* 66(1971):783-801.
- . *The Foundations of Statistics*. Second Revised Edition, New York: Dover Publications, Inc., 1972.
- Seneta, E. "Probability, History of (Outline)." *Encyclopedia of Statistical Sciences*, Vol. 7, ed. S. Kotz and N.L. Johnson, pp.218-22. New York: John Wiley & Sons, 1986.
- Siegel, S. "Decision Making and Learning under Varying Conditions of Reinforcement." *Ann. New York Acad. Sci.* 89(1961):766-83.
- Sims, C.A. "Scientific Standards in Econometric Modeling." Discussion Paper No. 82-160, Center for Economic Research, Dept. of Econ., Univ. Minnesota, Minneapolis, MN, 1982.
- Smith, V. "Experimental Economics: Induced Value Theory." *Amer. Econ. Rev.* 66(1976):274-79.
- . "Microeconomic Systems as an Experimental Science." *Amer. Econ. Rev.* 72(1982):923-55.

- Stael von Holstein, C-A.S. *Assessment and Evaluation of Subjective Probability Distributions*. Stockholm: The Economic Research Institute, 1970.
- Stigler, G.J. "The Development of Utility Theory." *J. Polit. Econ.*, Part I, 58(1950):307-27; Part II, 58(1950):373-96.
- Texas A&M University, Office of University Research Services. "Guidelines for Human Subjects in Research." College Station, Texas, 1984.
- Theil, H. *Applied Economic Forecasting*. Amsterdam: North-Holland Publishing Co., 1966.
- Tversky, A., and D. Kahneman. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185(1974):1124-31.
- von Neumann, J., and O. Morgenstern. *Theory of Games and Economic Behavior*. 2nd ed. Princeton NJ: Princeton Univ. Press, 1947.
- Walker, O.L., A.G. Nelson, and C.E. Olson. "Educational Programs for Risky Decision Making." *Risk Management in Agriculture*, ed. P.J. Barry, pp.166-79. Ames: Iowa State Univ. Press, 1984.
- Winkler, R.L. "The Quantification of Judgment: Some Methodological Suggestions." *J. Amer. Stat. Assoc.* 62(1967):1105-1120.
- "Scoring Rules and the Evaluation of Probability Assessors." *J. Amer. Stat. Assoc.* 64(1969):1073-78.

Winkler, R.L., and A.H. Murphy. "Nonlinear Utility and the Probability Score." *J. App. Meteorology* 9(1970):143-48.

**APPENDIX A**  
**INSTRUCTIONS FOR UTILITY EXPERIMENT**

## INSTRUCTIONS--LOTTERY

This is an experiment in the economics of decision making. The experimenters are studying how people value small amounts of money in a lottery setting. It will not cost you any money to participate, and if you follow the instructions carefully you may earn a *CONSIDERABLE AMOUNT OF MONEY* which will be *PAID TO YOU IN CASH* at the end of the experiment.

### LIST OF STEPS IN THE EXPERIMENT

#### Step 1: How the Lottery Works

In this game you will play a series of lotteries; we'll call each one a "trial". In each trial you will be shown a lottery with only two possible payoffs, \$1.00 and \$0. The chances of winning a lottery (that is, of winning \$1.00) will change from one trial to the next, so you need to pay attention to what those chances are in each trial. The chance of winning the lottery in any given trial is determined by the "WIN:LOSE" odds which we will write on the blackboard at the beginning of each trial. To decide the outcome of the lottery, we will draw a ball from a bingo cage that has 100 balls in it that are numbered from 1 to 100. On the blackboard we will write the "breakpoint" for that lottery (this is always the number on the LOSE side of the WIN:LOSE ratio). If the number on the ball that is drawn is *larger* than the breakpoint on the blackboard, you win \$1.00. If the number on the ball is *smaller* than the breakpoint on the blackboard, or if it is *equal* to that number, then you do not win anything.



Now, to make the game more interesting, we will ask you to write down the amount of money you would have to receive *with certainty* in order to make you indifferent between getting that amount of money or playing the lottery with the odds listed for that trial. What we mean by this is that there is probably some amount of money which you would prefer to have right now rather than play the lottery. And there is also some other amount of money which, if you were offered it right now, would not be enough for you to give up your chance at winning the lottery. So it stands to reason that there would be some in-between amount of money that we could offer you right now and you would not care one way or the other whether you got the "sure thing" or played the lottery with the given odds. Obviously the amount that represents your "value of indifference" would depend on what the WIN:LOSE odds were for that trial and so it would change with every trial. For example, a 99:1 ratio (that is, a 99% chance of winning the lottery) would make the opportunity to play that trial's lottery considerably more valuable than, say, a ratio of 5:95 (that is, only a 5% chance of winning the lottery).

To help you decide on this in-between amount of money we will give you a "ticket" in each trial that represents your right to play that trial's lottery. This ticket now has some value to you since you may keep it and play the lottery and take a chance at winning \$1.00, or you may sell it to the experimenters for some amount of money which you will be certain of receiving. But since the "exact value of indifference" that we ask you to write down should be the one which is exactly in-between caring whether you keep the ticket or sell it, then it should not matter to you which one happens. The next section describes how we will do this.

## Step 2: The Tickets and the Value Scale

The tickets you will be using are attached to the instructions. You will use a new ticket at the beginning of each trial. Printed on each ticket is a "value scale" for you to use in indicating how much you think each ticket is worth to you. Refer to one of these tickets for the following discussion.

To indicate the value of a ticket to you, put your finger at the bottom of the scale and ask yourself which you would prefer to have--the dollar amount shown on the scale, or the ticket. We assume that you would prefer the ticket to \$0.00. Now move your finger up the scale toward the top continuing to ask the same question. At the very top of the scale is an amount of money equal to the largest amount that could be earned by keeping the ticket. The scales used in this experiment are constructed so that for some of the numbers at the bottom you will prefer to keep the ticket, and for some at the top you will prefer to have the money.

As you move your finger up the scale, stop when you have reached the point at which you are indifferent between keeping the ticket and receiving the amount on the scale. What we would like to know is this: What is the exact dollar amount at which you are indifferent between keeping the ticket and the amount of money on the scale. Mark this amount with an "X" on the scale. Since X's are not always easy to read, and as the scale may not be fine enough for you, we also ask that you write the amount you marked in the space provided. This number may be written in any amount of cents between \$0.00 and \$1.00. Please also write your code number and the chance of winning the lottery in the appropriate spaces on the ticket.

In order to provide you with an incentive to be as accurate as possible we will do the following: after your choice on the scale has been made, one of the dollar amounts on the scale will be randomly chosen from the bingo cage. If the amount drawn from the cage is greater than the amount you marked on the value scale, *you will receive the amount chosen* and give up your ticket to the experimenters. If the amount drawn is less than the amount you marked, you will keep the ticket and go on to play the lottery. If the amount you marked exactly matches the number randomly picked in the draw, the toss of a fair coin will determine whether you get the amount drawn or the ticket.

Notice that your interest is best served by accurately representing your preference. If the mark you place on the scale is too high or too low, you will be passing up opportunities that you prefer. For example, suppose you are indifferent between taking \$0.40 and keeping the ticket. In this case you should have marked \$0.40 on the value scale, but instead you marked \$0.60. If the amount picked at random on the scale is anything between \$0.40 and \$0.60 (perhaps \$0.50), you would be forced to keep the ticket, even though you would rather have the amount the experimenter would pay you. Suppose you put a mark on the scale that was too low? Your point of indifference was really \$0.40, but you marked \$0.20. If the amount chosen at random is greater than \$0.20 and less than or equal to \$0.40, then you would be forced to take the money even though you would rather keep the ticket and play the lottery.

When you have filled in the blanks on your ticket, transfer the "exact value of indifference" to your Record Sheet (attached) and set your completed ticket aside. After the bingo cage is turned and the first ball

is drawn you may not change your value of indifference on your ticket or record sheet for that trial.

### Step 3: Playing the Lottery

If you sell your ticket at the first draw stage, then write down the amount of money you are to collect in the first payment column of the Record Sheet.

If you did not sell your ticket at the first draw stage then you will play the lottery as described in Step 1. The first ball is returned to the bingo cage and a second draw from the bingo cage will determine whether those of you who are playing the lottery win \$1.00 or \$0. Write down the outcome (whether you won \$1 or \$0) in the second payment column. You will receive a payment in either the first draw or the second draw but not both. When you have recorded your payment we will continue on to the next trial where there will be a new WIN:LOSE ratio, ticket, and lottery.

**PLEASE DO NOT SPEAK TO ANY OF THE OTHER PARTICIPANTS DURING THE EXPERIMENT. THIS IS IMPORTANT TO THE VALIDITY OF THE EXPERIMENT AND WILL NOT BE TOLERATED.**

If you have a question that you feel was not adequately answered in the instructions, please raise your hand and ask the experimenters at this time. Your earnings may suffer if you proceed into the experiment without understanding the instructions!!!

# TICKET

**VALUE SCALE**

\$1.00	_____
.95	_____
.90	_____
.85	_____
.80	_____
.75	_____
.70	_____
.65	_____
.60	_____
.55	_____
.50	_____
.45	_____
.40	_____
.35	_____
.30	_____
.25	_____
.20	_____
.15	_____
.10	_____
.05	_____
\$0.00	_____

Your code number \_\_\_\_\_

Chance of winning  
the lottery (\$) = \_\_\_\_\_

Exact Value of  
Indifference = \_\_\_\_\_

## RECORD SHEET

TRIAL	WIN:LOSE RATIO	EXACT VALUE	FIRST DRAW	PAYMENT (IF ANY)	BREAK- POINT	SECOND DRAW	PAYMENT (IF ANY)	CUMULATIVE EARNINGS
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								
25								
26								
27								
28								
29								
30								
31								
32								
33								
34								
35								
36								
37								
38								
39								
40								

**APPENDIX B**  
**INSTRUCTIONS FOR FORECASTING EXPERIMENT**

## INSTRUCTIONS -- FORECASTING GAME

This is an experiment in the economics of decision making. The experimenters are studying how people make forecasts.

If you follow the instructions carefully and make good decisions you may earn a *CONSIDERABLE AMOUNT OF MONEY* which will be *PAID TO YOU IN CASH* at the end of the experiment.

In this experiment you will be shown a series of numbers and then you will be asked to forecast what the next number in the sequence might be. After you have made your forecast you will be shown the actual outcome of the number and you will be given a "score" which reflects how accurate your forecast was. This score is actually a cash payment, so you will be earning money according to how good your forecasts are.

You will be using a computer terminal to enter your forecast and receive information. The computer is completely "passive" in the sense that it is used only to record your entries, calculate your earnings, and display information. *Neither the computer nor you have any influence on what number actually occurs in any period.* You do not need to know anything about computers to participate in this experiment. You will be given instructions later that will tell you how to enter information into the terminal.



## LIST OF STEPS IN THE EXPERIMENT

The following is a brief outline of the experiment to give you an overall picture. Detailed descriptions of each step will be given in the next section.

### *STEP 1: FORECASTING THE NEXT NUMBER*

You will forecast the next number in a series of numbers, over many periods, with one forecast being made and one outcome revealed each period. You will use a "probability line" to make your forecast in terms of what you think the chances are that the next number will fall in one or the other of eight ranges.

### *STEP 2: THE PAYMENT RULE*

You will be paid for each forecast according to a rule that depends on how close you come to the correct answer.

### *STEP 3: INFORMATION AVAILABLE*

Information in the form of tables and graphs which summarize or display the outcomes from past periods will be provided to help you in your forecasting task.

### *STEP 4: YOUR EARNINGS*

You will be paid the sum of your earnings from all your forecasts at the end of the experiment.



Anything to the left of the center line represents negative numbers, and anything to the right represents positive numbers. The numbers lined up vertically are the beginning and ending numbers in each range of numbers. For example, the range identified as "4" includes numbers between -001 and -200. The range identified as "6" includes numbers between +200 and +399. The range identified as "1" includes any number equal to or less than -601. The range identified as "8" includes any number equal to or greater than +600.

In each of the range slots marked "1", "2", "3", "4", "5", "6", "7", and "8" you will be asked to enter a probability (i.e. a number from 0 to 1, inclusive) which represents what you think the chances are that the next number will fall in the range of numbers signified by that slot. For example, if you think that there is a 20% chance that the next number will be a number between 400 and 599 then you would put a ".2" in the slot marked "7". An example of this task when completed might look like the following:

```

FOR FORECAST      1 PLEASE ENTER YOUR PROBABILITIES:
1. : 0
2. : 0
3. : 0
4. : .167
5. : .333
6. : .333
7. : .167
8. : 0

```





ally occurs, the possible payment from using that particular set of probabilities is either \$0.15 or \$0.00.

Case 2 illustrates the case where all eight slots are assigned the same probability, ".125". In this case, no matter what the outcome, the payment will be \$.0937.

Case 3 illustrates the payments possible from using just 0's and 1 as probabilities for all eight possible outcomes: the payment is \$0.00 if you are completely wrong (i.e. the outcome is not one to which you assigned a probability of "1"); whereas it is \$0.75 if you are perfectly right.

Cases 4 through 8 illustrate the possible payments from five other sets of probabilities. Note that the possible number of sets you might use is huge; these eight cases are only used to give you a feeling for how the payment rule works. As the experiment progresses you will be getting continuous feedback from the rule in terms of the payment that follows each of your choices of a forecast probability set.

### Step 3: Information Available

The experimenters have attempted to make information summarizing what is known about the series of numbers being forecasted readily available to you so as to make your task as easy as possible. This information comes in the form of tables and graphs and may be displayed on your screen or on handouts.

As explained earlier, your terminal displays the "HISTORICAL SERIES OF DATA" and adds to this screen each new number as it occurs. You can also access your own "PERFORMANCE SUMMARY" screen to look back at your

.

past forecasts, the actual outcomes (identified by the range slot they fell in), and the resulting payments. An *example* of this screen, using the eight cases illustrated earlier, is shown below:

FOR. PER.	PROBABILITIES USED								ACT. OUT.	CURR. SCORE	CUMUL. EARN.
	1	2	3	4	5	6	7	8			
1	0.000	0.000	0.200	0.200	0.200	0.200	0.200	0.000	4	0.1500	0.15
2	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	6	0.0937	0.24
3	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	5	0.0000	0.24
4	0.000	0.000	0.333	0.334	0.333	0.000	0.000	0.000	3	0.2497	0.49
5	0.000	0.000	0.000	0.167	0.333	0.333	0.167	0.000	4	0.1252	0.62
6	0.020	0.080	0.100	0.300	0.300	0.100	0.080	0.020	2	0.0600	0.68
7	0.000	0.000	0.000	0.500	0.500	0.000	0.000	0.000	1	0.0000	0.68
8	0.100	0.100	0.200	0.000	0.500	0.100	0.000	0.000	6	0.0750	0.75

In this experiment you will not be able to access the "EXPERT'S PERFORMANCE SUMMARY".

Handouts such as that illustrated in Figure 1 (attached at back) may also be helpful. This sheet can be used to keep track of the probabilities you have used. Figure 2 (attached at back) is an *example* of the kind of graph of past outcomes which will be handed out.

#### Step 4: Your Earnings

At the end of the experiment you will be paid in cash the amount of your "CUMULATIVE EARNINGS".

## USING THE COMPUTER TERMINAL

If you look at your terminal you will see the following message:

*ENTER YOUR CODE NAME:*

Here you will type in the code number you have used in past experiments, and then press the "RETURN" key. In a few seconds the following message will appear:

*WELCOME! PLEASE BE SEATED AND WAIT FOR YOUR INSTRUCTIONS. WHEN YOU UNDERSTAND ALL THE INSTRUCTIONS, PRESS THE "RETURN" KEY SO WE MAY BEGIN.*

Any time you are asked to enter information into your terminal you will do so by typing what you want to enter and then pressing the "RETURN" key. Now press "RETURN" so that we can begin. You are now asked to enter your name. Use your code number again here. After you have typed in your "name" and pressed the "RETURN" key you will be asked to enter your social security number (SSN). Do this in the same way as before. Next you will be asked to enter your special ID (not your University ID!). Your special ID number is printed on the card given to you. Then your "name", SSN, and ID will reappear together with the following message:

*DO YOU WANT TO CHANGE ANY OF THE ABOVE (Y/N) ?*

If you are not satisfied with either your "name", SSN, or ID as entered, type the letter "Y" and you will be given an opportunity to correct it. Otherwise, type "N" to continue.

Notice that any time you are asked to enter something and the terminal is waiting for your response, the "cursor" (i.e. the little white flashing box) is on. This will be true throughout the experiment. That



is, any time the cursor is on, the computer is waiting for you to enter something so that it may continue. REMEMBER THAT WHENEVER YOU ENTER ANY INFORMATION ON THE TERMINAL YOU MUST ALWAYS PRESS THE "RETURN" KEY AFTER TYPING IN YOUR INFORMATION. If you do not press the "RETURN" key the computer will not receive your information. Also be careful that you do not type the letter "o" in place of the number "0", nor the letter "l" for the number "1". The top row of keys on your keyboard should be used for numbers, or you may use the keypad on the right side of the keyboard. The period can be used for a decimal point. To backspace, use the left-pointing arrow at the top of the keyboard.

Please do not play with the keyboard while you are waiting for the experiment to continue. That is, you should only be entering information from the keyboard of your terminal when the cursor is on.

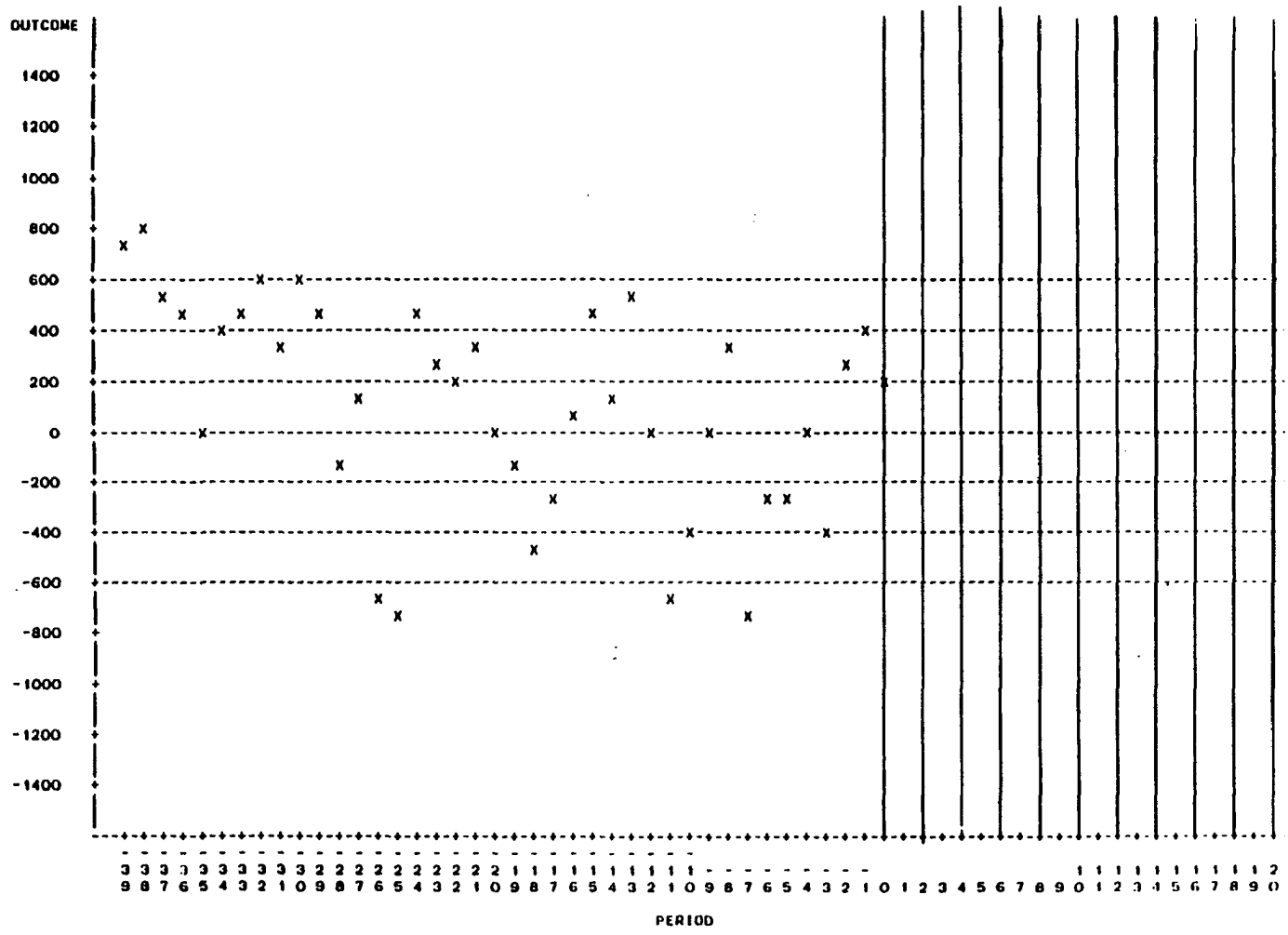
Your identity will remain confidential and will not be used for any purposes other than to account for our expenditures to the funding agencies.

PLEASE DO NOT SPEAK TO ANY OF THE OTHER PARTICIPANTS OR LOOK AT THEIR COMPUTER SCREENS DURING THE EXPERIMENT. THIS IS IMPORTANT TO THE VALIDITY OF THE EXPERIMENT AND WILL NOT BE TOLERATED.

If you have a question that you feel was not adequately answered in the instructions, please raise your hand and ask the classroom monitor at this time. Your earnings may suffer if you proceed into the experiment without understanding the instructions!!!



**FIGURE 2**  
**Graph of Past Outcomes by Period (-39 to Present)**



**Supplement 1**  
**Payment Table Shown to Subjects under Quadratic**  
**Scoring Rule**



**Supplement 2****Payment Table Shown to Subjects under Linear Scoring  
Rule**



**VITA**

Name: Robert Graham Nelson

Education: M.S., 1977  
Auburn University  
Major: Fisheries

B.S., 1974  
Oregon State University  
Major: General Science (Biology)

Mailing Address: c/o G. A. Nelson  
83 Stillwaters Dr.  
Dadeville, AL 36853