Decision Costs and Economic Rationality: Some Experimental Design Problems

Rick K. Wilson

Rice University
and
California Institute of Technology

I wish to thank my long suffering colleagues, Bobbi Herzberg and Cliff Morgan, each of whom experienced the fatal design flaws discussed in this paper. While I wish I could pin the blame on them, they are blameless. This paper was prepared for delivery at the annual meeting of the American Political Science Association, Atlanta, Georgia, August 31 - September 3, 1989.

Abstract

When are controlled laboratory experiments valid tests of abstract political models? This paper touches on the general question of internal validity in the design of laboratory experiments. Using a number of (failed) examples from the author's own work, key threats to internal validity in experimentation are explored. While the focus of the paper is with testing models of rational choice, the concerns are equally applicable to any controlled laboratory environment testing an abstract theoretical model.

Introduction

Laboratory experimental methods have not been widely used in political science. Only in the past ten years have laboratory experiments gained a foothold in the discipline, although that foothold remains precarious. Unlike the widespread acceptance of laboratory experimental methods in a cognate discipline like psychology or the growing use of experiments in economics, such an empirical method lags behind in political science. That political scientists have not embraced laboratory experiments is odd since much of politics concerns experimentation. Certainly the reform of political institutions represents a form of experimentation, whether characterized by changes in voting rules wrought by the Progressive movement throughout the United States at the turn of the 20th Century or by reforms in the structure of Congress taken in the early 1970's. These "field experiments" are based on conjectures about the fundamental relationship between institutions and outcomes. Changes in the structure of the institution are expected to produce changes in political outcomes.

A niche for laboratory experimentation in political science is being carved out. This is so for two reasons. First, laboratory experiments offer an exceptional means for gaining precision over empirical instruments. This is accomplished both through the experimenter imposing controls over the settings being tested and by developing careful measurement instruments. Clearly many of the empirical instruments traditionally used in political science carry with them sources of bias, both in terms of reliability and validity. Survey instruments, government expenditures, or roll call votes all carry with them measurement problems. These problems are exacerbated when relying on such instruments to test highly sensitive theoretical constructs. A second reason is that as political scientists turn to increasingly abstract and complex models, it has become increasingly difficult to find natural settings that provide unambiguous tests of those models. Laboratory experiments allow researchers to build settings

capable of testing the robustness of those theoretical models. As well, such experiments can provide insights for the redevelopment and extension of those theoretical models.

Laboratory experimental methods are not without drawbacks. Given the simplicity of inducing controls over an experiment, it is equally simple to allow the experimental design to take over for theorizing. Rather than thinking through and modeling the relationship between several variables, this amounts to little more than "hunting" for relationships. Conducting many different experiments, each with minor changes across treatments, looking for measurable differences, without any theoretical guidance, is no different than searching an m by m correlation matrix for "the big ones." In both instances relationships will be uncovered. However, in neither case will they be interpretable without a theoretical construct predicting those relationships. The temptation to let experiments develop theory should be suppressed. Instead, experiments should remain the tool of theory.[1]

This paper explicitly focuses on laboratory experiments. Fundamental to the discussion is the linkage between theory and experiment. Part of my contention is that in the absence of the former, the latter will surely flounder. The structure of the theoretical model largely dictates the design of an experiment. Design problems, however, haunt laboratory experiments. Especially troublesome are experiments in political science aimed at examining collective choice settings. Unlike classical experimental settings concerned with treatments administered to an individual, most settings of interest to political scientists involve the interaction of many individuals. While the usual problems of external and internal validity must be attended to in the design of a laboratory experiment, I single out for discussion two additional problems related to internal validity. The first problem focuses on task complexity while the second focuses on subject motivation. To illustrate the serious nature of those threats to internal validity, I provide two examples of my own recent experimental designs.

1 What constitutes a "theory" in political science, unfortunately, is often quite loose. Everything from "hunches" to "correlations" between variables to formal, mathematical constructs are considered theories by their various proponents. In this paper I reserve the use of theory for deductive models providing causal explanation and yielding predictions. For an extended discussion on this point, see Fiorina (1975).

External and Internal Validity

Laboratory experiments, unlike field or natural experiments, are appealing due to the ability of the researcher to impose control over the institution he has created. As Leon Festinger writes

> A laboratory experiment may be defined as one in which the investigator creates a situations with the exact conditions he wants to have and in which he controls some, and manipulates other, variables. He is then able to observe and measure the effect of the manipulation of the independent variables in a situation in which the operation of other relevant factors is held to a minimum. (1953, p. 137)

However, the ability of a researcher to create and change a laboratory setting also carries with it the danger that the experiment is used to search for relationships, rather than test hypothesized relationships. Only theoretical models can yield consistent predictions. As such theories should be regarded as fundamental for experimental design.[2] As I argue below, laboratory experiments face two serious threats — threats of external and internal validity. If experiments are linked with theoretical models, then problems of external validity are endemic and little can be done to compensate for such a threat. However, the more closely theory and experiment are linked, the less damaging are threats to external validity. Instead researchers must focus their attention on problems of internal validity

Laboratory experiments, like field experiments and quasi-experiments, are useful only if properly designed. A proper design implies that the experiment satisfies basic cannons regarding internal and external validity. Part of the argument in this paper, however, is that external validity is problematic for most laboratory experiments. However, it need not be considered as a serious criticism of experimental method. Satisfying basic criteria for internal validity, then, becomes critical for assessing laboratory experiments.

---

2 Festinger (1953) provides an opposing view on this point. Instead, he argues for a powerful link between laboratory experiments and "real-life situations" (p. 140). However, such a perspective goes hand-in-hand with a search for relationship from which theoretical constructs can then be developed. Unfortunately political scientists have an abundance of theoretical constructs, but lack sufficient data to test those constructs.

The Problem of External Validity.

External validity is concerned with how well events in an experiment generalize to a broader (e.g. natural) setting. At heart is the isomorphism between the structure of the experiment and the natural setting. So, for example, in the study of committee experiments, where participants build agendas, form coalitions, and vote over an abstract policy space, questions of external validity are concerned with how well the results obtained in this setting translate to committee voting in a legislature. External validity's fundamental question is with whether an observed phenomenon generates the same phenomenon in the more general setting. If the experimental design is isomorphic with the natural setting, then the same effects should occur in both.

It is rarely the case that laboratory experiments are designed to exactly replicate natural settings. Why reconstruct a complex environment that already exists? Indeed, the rationale behind most experimentation is to provide for controls over the variation (noise) found in most natural settings. This typically means simplifying the setting and inducing controls such that the laboratory experimental setting may be quite different from the natural setting. The same is true for theoretical models, which try to capture crucial linkages for complex settings. This is typically achieved through abstraction and simplification of the setting under study, rather than replicating that setting in toto.

As experiments seek to simplify the environment, degrees of freedom are lost in linking the experimental setting to the natural setting. As these degrees of freedom are lost, so too is external validity. The rationale behind laboratory experimentation is with testing specific relationships, not with recreating existing institutional settings. If experiments do not aim at replicating existing environments, but instead aim at simplification, then what dictates the proper design of an experiment? Theory should direct experimental design such that laboratory experiments should focus on theoretical concerns, not on recreating the world.

Problems of Internal Validity.

While problems of external validity can be set aside if the experimenter is not aiming at reconstructing a particular natural setting, problems of internal validity cannot be ignored. Quite simply, problems associated with internal validity are concerned with the correspondence between theory and experiment. While external validity focuses on the linkage between experimental results and generalizations to a natural setting, internal validity is concerned with the linkage between what the experimenter intends to test and what the experiment in fact tests.

One claim forwarded here is that laboratory experiments should be theoretically motivated. If they are, then ensuring that the experiment meshes with the theory is crucial. Theories are a means of capturing important causal linkages of natural settings and are not intended to fully explain the world. As abstractions of more complicated natural settings, theories isolate central elements of natural processes. But such simplifications make generalizations to more complex natural settings quite difficult. Likewise the lack of control over a natural setting and its complexity makes testing abstract theories quite difficult. At best, quasi-field experiments provide only a rough approximate test of theoretical concepts. Laboratory experiments, however, allow researchers substantial control over the testing process. Abstract theoretical structures can be replicated in a laboratory setting, establishing a correspondence between theory and empirical test that is necessary in order to test the robustness and utility of theoretical models. As Plott (1979) argues, experimenters need not be tied to existing natural settings, but within the laboratory can design completely artificial settings if necessary to test theoretical conjectures.

Laboratory experiments, then provide considerable flexibility with which to test abstract theories. Experimenters, then, must focus their attention on problems linking theory and experiment. These problems fall under the rubric of internal validity, since they involve the linkage between what the experimenter intends and what the experiment measures. In some sense these problems cut to the heart of operationalizing theoretical concepts into experimental conditions and measurements.

In this paper I differentiate between two general types of internal validity problems. The first pertains to the usual experimenter effects and other unplanned contaminants that undermine the validity of an experiment. The second set of problems pertain to the isomorphism between theory and experiment. It is on these problems that this paper focuses.

Internal validity problems due to experimenter effects have been thoroughly discussed and documented by Campbell and Stanley (1963) and Cook and Campbell (1984). In experimental settings the researcher must be concerned with a variety of effects that affect the interpretability of the experimental results. Campbell and Stanley identify eight threats to internal validity in these instances. The first is the threat of history, in which exogenous events during the course of the experiment exercise an independent effect over outcomes. The second is a maturation effect, in which observed experimental differences are a function of time processes rather than experimental processes. Testing effects represent a third threat to internal validity when an initial test has an affect on a subsequent test. Changing the instruments that measure experimental differences also threatens interpretability of results. The fifth threat is that of statistical regression, in which subject groups are selected based on extreme differences. A more general threat is posed by using other means for differential selection of subject groups. If experiments extend over time, mortality rates in experiments present yet another threat, since it is never certain whether these losses are randomly distributed or the effect of particular subject groups. The eighth and final threat to internal validity involves an interaction between maturation and selection of subject groups in complex experimental designs with multiple groups. All of these effects are detailed in Campbell and Stanley (1963) and are readily treated under a number of experimental designs.

A second set of threats to internal validity are concerned with the operationalization of theoretical concepts. These are threats tied directly to interpreting whether the experiment tests what the experimenter intends. Here we focus on two types of problems: problems with interpreting the task and problems with inducing motivation. Problems of task interpretation are concerned with whether or not subjects understand the task they are to accomplish in the

experiment. If the experimental conditions do not adequately differentiate for subjects what it is that they are to do during the experiment, results from such experiments will be uninterpretable. Fundamentally this questions whether all participants reach the same conclusions about the structure of the experiment and whether the conclusions participants reach are the same as those the researcher intends. Problems of inducing motivation stem from concerns that the motivations driving the theoretical model are not linked to those relied on by subjects in an experiment. If there is no overlap between motivations suggested by the model and what motivates subjects in an experiment, then the data will say nothing concerning the theory. In the next two sections I treat each of these threats to internal validity separately and provide examples of how each affects experimental results.

Task Interpretation

As noted any experiment is a representation of some environment. When designing an experiment it is incumbent on the experimenter to define a complete, self-contained environment. This environment must include complete rules defining admissible actions by participants (strategies) and rules that map actions onto outcomes (payoffs). If an experiment is concerned with predictions of voting, then the experimental environment must characterize the scope of the issues being voted on, the sequencing of those issues, the aggregation rule employed to determine winners and losers, the weights assigned to different votes, levels of information about the issues, the extent to which actors may communicate, and the content of any communication. In addition the matter of individual motivations must be considered, though we reserve discussion of this point to the next section. In essence, the experimenter is charged with defining a complete environment within which subjects are free to act. This environment is no less "real" for the actors than are natural settings, since each must learn the rules defining possible actions, accomplish tasks proscribed by the institution, and derive rewards from implementing some set of strategies. Above all, subject must be clear as to their task in the experimental setting.

When subjects fail to conceptualize an experiment in the same manner intended by the experimenter confusion will erupt over interpreting outcomes from the experiment. As an analogy, imagine the confusion in a football game when some of the players interpret the game under the rules of rugby, others as soccer, and a third subset has no idea how the rules work. While an outcome could be observed — a final score — still the activity (carnage) on the field leading to the outcome would not resemble the game played by the NFL. Moreover, it is quite likely that the final score would fail to resemble that of other football games played while following the rules.

Subject misinterpretation comes under three different guises. While each may have different consequences for the process of an experiment, all have similar effects for interpreting what is measured. The data from the experiment ordinarily will not match what is predicted by theory, leading the researcher to reject the theory (committing a Type I error). Worse still, an incorrect theory might be supported under an experimental test, when subjects misunderstanding the experiment yield results consistent with the incorrect theory's predictions (thereby committing a Type II error). In either instance, such threats to internal validity are a function of subjects misinterpreting the task.

One source of subject misinterpretation stems from a subject simply not understanding the task. This could come from unclear instructions, confusing tasks to perform, or too many tasks for the subject. If the subject does not understand what it is that he should do during the experiment, then whatever he does will only accidently be linked to what the researcher intends to measure. For instance, what can be inferred from a subject's behavior, if confused by what it is he should do, that subject randomly selects his actions?

A second source of misinterpretation comes from a subject interpreting the rules of the experiment differently from that intended by the experimenter. This has two consequences. As with the case where the subject does not understand the task at hand, a subject's choices (actions) will not be predicted by the theory. In a second sense, in experiments where subjects interact with one another (as is the case with most experiments concerned with political

behavior), a single subject playing under different rules may send misleading signals to other players, resulting in them switching their strategies. In either instance, this creates an important disjuncture between theory and experiment and undermines the interpretability of data generated by the experiment.

A third source of misinterpretation stems from a subject correctly interpreting the task of the experiment, dealing with other subjects who are confused about the experiment, and having that subject revise his interpretation to conform to that of those who misunderstand or misinterpret the experimental task. In this sense, the subject is simply mislead, succumbing to examples set by other subjects in the experiment.

How, then, can one ensure that all subjects interpret the experiment's task in the manner intended by the researcher? First, and foremost, the researcher must ensure that the rules and procedures imposed bring closure to the experiment. As Smith (1982) argues, experimental design involves the construction of institutions within which actors make choices. As with all institutions, if the rules are not well defined, then actors are free to act in ways quite different from that expected (or predicted). A well designed experiment closes off alternative behavioral strategies not envisioned by the experimenter or behavioral strategies that are not part of the theoretical model. Closure eliminates unplanned effects due to ambiguity or looseness in the rules. These rules may define admissible strategies, property rights, levels of communication, or collective choice rules. Whatever the rules, they must clearly delimit for subjects, what they are able to do.

A second, crucial design element to eliminate subject misinterpretation is constructing a proper instruction set. Instructions are critical for experimental design, since they provide important cues for subjects. Instructions also provide the means for guaranteeing that subjects comprehend the details of the experiment. By including tests of comprehension at key points in instructions, an experimenter can observe whether or not each subject understands that part of the instructions and reinforces the point of the instruction. Tests of comprehension are simple

and largely unobstrusive when included in instructions. Moreover, such tests provide a point at which subjects can ask for clarification if confused over an instruction.

A third design element is to substitute abstract for concrete symbols in the experiment. This ensures that subjects do not adopt a form of transference, where their admiration or disdain for the symbol overrides their motivation for performing the task or their interpretation of the task. For example, in the context of a majority rule committee game in which individuals vote for agenda changes across issue dimensions, those issues should remain featureless. If characterized as guns and butter, a subject, who is a hawk, might reconceptualize the game as one between substantively important dimensions, ignoring his own (induced) utility for tradeoffs across arbitrary dimensions. A model of collective choice, predicting some set of outcomes, presumes that member's preferences are fixed (and known). While it would be relatively easy to represent a "hawk" or a "dove's" preferences on such issues, these must be controlled by the experimenter. A subject setting his own utility functions based on an unintended element of experimental design is misinterpreting the task of the experiment. Avoiding symbols with substantive meaning for subjects can control such effects.

Avoiding deception in experiments is another means of assuring that subjects do not misinterpret the task of the experiment. Unfortunately, subjects have often been exposed to deception in order to hide experimental treatments. Consequently, subjects build up expectations about what the experimenter intends, imagining that the task at hand is really nothing more than a placebo hiding the real purpose of the experiment. To this end, subjects attempt to reinterpret the experiment in light of what they imagine the experimenter is testing. This poses a serious threat to interpretatibility of results. Assuring subjects that they will not be deceived goes a short way toward minimizing this problem. Unfortunately the use of deception in experimentation produces a real public goods problem, since the subject pool (ordinarily small even at large state universities) quickly becomes polluted. Although, deception is often a quick and convenient method for building an experimental design it should be eschewed.

Finally, reducing task complexity also contributes to reducing subject misinterpretation. The fewer tasks that subjects must deal with means fewer opportunities for confusion. The introduction of computers into experimental methods is one step in the direction of reducing subject confusion, especially if computers minimize relatively trivial bookkeeping tasks. Any experimental design should be re-examined to see if some feature of the experiment can be eliminated and if so whether doing so would remove a potential source of confusion for a subject.

While the point is rather obvious that an experimental design should not confuse a subject, still experiments are conducted in which confusion is so serious as to undermine their results. To show the nature of this problem, I provide an example of a committee experiment in which subject confusion and misinterpretation resulted in data points at some variance with the predicted theory.

Committee Experiments on Agenda Costs.

The theoretical model motivating these committee experiments is a commonly used spatial model of a strong simple game. The question motivating the theory concerns the kinds of transaction costs that yield an equilibria in a multidimensional policy space when building an agenda. Most spatial models of voting show that under a typical forward-moving agenda process, outcomes will be distributed throughout the policy space if no one is assigned specialized agenda setting powers (McKelvey, 1976; Schofield, 1978). Other models show that outcomes may be somewhat more constrained in their distribution in the policy space, but will still be widely distributed (Shepsle and Weingast, 1984; McKelvey, 1986). Taking a similar model, Herzberg and Wilson (1989) show that when costs are imposed on the agenda process in reasonably trivial ways, an equilibrium is induced on the policy space. This equilibrium, then provides a set of predictions which can be tested in a laboratory experimental setting.

The experimental design is based on committee experiments conducted by Fiorina and Plott (1978), McKelvey, Ordeshook and Winer (1978) and Wilson (1986). Only "naive"

participants were allowed in the experiment --individuals who had not previously participated in a decision making experiment. Participants self-selected the time at which they wished to participate, choosing from a variety of time slots. Since player identities were randomized and kept anonymous during the experiments, there is little danger that groups of players successfully colluded using pre-arranged coalition strategies.

Upon showing up for the experiment, individuals were seated at micro-computers which were physically separated from one another by partitions. Participants were given computerized instructions designed to familiarize them with the experiment.[3] The instructions usually lasted no more than 15 minutes and were self-paced. The idea behind self-paced, computerized instructions was that subjects gained hands-on experience at the terminals, subjects could focus on those parts of the experiment they thought confusing, and experimenter effects stemming from differences in the presentation of instructions were removed. Also, to ensure subjects understood the instructions, they were quizzed at various points and had to provide correct answers before they could proceed. Upon completing these instructions, individuals participated in a practice period for which they were not paid. Participants were urged to try all the options until they were familiar with the experiment. Participants were cautioned that once they completed the practice session, they would begin the experiment, and that their earnings depended on the collective choice that was reached. Subjects, then, were given a number of opportunities to familiarize themselves with the experiment. It was thought that the instructions, quizes, and practice session would remove any subject confusion.

In the experiment, participants were to collectively choose an alternative from a two-dimensional policy space. Alternatives were represented as Cartesian coordinates from orthogonal dimensions labelled X and Y. Each individual was assigned a specific point in this two-dimensional space as an ideal point and was given a preference function. In these experiments, member preferences were represented as circles, with utility linearly decreasing with distance from the member's ideal point. These ideal points and associated utility functions

3 These instructions are available from the author upon request.

are given at the bottom of figure 1. By using an abstract policy space (made up of X and Y axes) and through inducing player's valuation for points in the space, we sought to avoid problems associated with participants adopting a different, subjective valuation for the policy space. Since a participant's payoff could become negative, each was given a $5.00 credit before beginning the experiment. The computer terminal displayed the alternative space, the member's ideal point, representative indifference curves, and the ideal points of all other members (but not their utility functions). The current status quo, as well as all proposals currently on the floor were also represented on this alternative space. In addition, members had before them a menu from which they could select a number of actions. This screen is displayed in figure 1.

<Figure 1 about here>

The choice of an outcome from the alternative space was controlled by strict forward moving agenda rules. Participants could bring any proposal to the floor. However, no proposal could be offered as an amendment to the status quo unless "seconded" by another member. This prevented a large number of "nuisance" votes from being counted and fits well with standard parliamentary procedure. A seconded amendment would be paired with the current status quo, with a majority vote for the amendment making it the status quo. Otherwise the status quo remained unchanged. From the menu participants could choose to propose an alternative, second another alternative, or choose to adjourn the experiment. As with a vote to amend the status quo, a vote to adjourn required a majority. If a vote to adjourn was successful, the period was over and members were paid the value of the current status quo.

Two treatments were used in these experiments. Each is identical, except for costs assessed at each agenda step. The first setting has no costs imposed. In the second treatment only those who vote for a successful amendment to the status quo were assessed a cost of $0.75. As member's accumulated costs in the experiment, these amounts were represented as a decrease in their utility function. So, for this cost treatment, a member's ideal point was initially worth $14.50. After the first successful amendment for which the member voted the ideal point was worth (and displayed as) $13.75. This continued with every successful amendment.

Operationalizing these agenda access costs as a function of the number of changes to the agenda directly stems from a model of costs developed in Herzberg and Wilson (1989). Throughout the experiment member's were fully informed as to their transaction costs.

In the absence of decision costs, the preference configuration used in these experiments lacks a majority rule equilibrium. However, when agenda costs are imposed a retentive equilibrium is induced. The model predicts:

> Experimental outcomes under a cost treatment will appear in the cost equilibrium (in the shaded area of Figure 2), while experiments under the no-cost treatment will not concentrate in this equilibrium set.

<Figure 2 about hero

The outcomes from five no-cost and five cost experiments are plotted on figure 3. As seen from figure 3, none of the cost-treatment outcomes are in the predicted equilibrium set. Meanwhile, two of the no-cost outcomes were at the outer boundary of the predicted equilibrium set. Nonparametric tests based on the euclidean distance of each outcome from the cost-set do not allow us to reject the null hypothesis that the outcomes are taken from the same distribution. In short, these results do not provide compelling evidence that agenda costs induce any empirical equilibrium.

<Figure 3 About Hero

Two pieces of evidence, however, indicate that agenda costs exert some effect on the agenda process. First, even though subjects take about the same number of votes in each of the treatments (an average of 20.2 votes for the cost experiments and 26.4 for the no-cost experiments) there are substantial differences as to the number of successful changes to the status quo by treatment. On average just over two amendments were passed in the cost induced experiments, while an average of nine amendments were successful in the no-cost experiments. This can easily be seen from the agenda trajectories in the experiments. Figure 4 plots successful amendments for the cost experiments, while Figure 5 plots successful amendments for a representative no-cost experiment. By and large under the cost treatment, the agenda

trajectories converge to the equilibrium. Under the no-cost treatment the trajectory cycles throughout the interior of the convex hull of member ideal points.

<Figures 4,5 About Hero

A second piece of evidence is that under the cost treatment there were almost no proposals made that could defeat the final outcome. In three of the experiments no proposals were made which could have defeated the final outcome. In another of the experiments a single proposal was made which could have defeated the final outcome. Finally, in the remaining cost experiment four proposals were on the floor that could have defeated the final outcome (though these proposals could just as well been considered two similar pairs since there was less than three units difference between alternatives constituting each "pair"). By comparison there was an average of 11.6 proposals that could defeat the final outcome in the no-cost experiments. One interpretation for these results is that subjects treat the agenda cost experiments differently, interpreting the time costs differently than intended, and as a consequence fail to search for winning amendments to the status quo. A second interpretation is that some subjects are simply confused by the structure of the experiment and as a consequence, those subjects, who are pivotal, never uncover winning moves into the cost equilibrium. Under the no-cost treatment, confused subjects are not a serious problem given the many possible agenda paths.

There are real opportunities for subject misinterpretation and confusion in these experiments. First, there is the obvious problem that the experimental design is confusing. Subjects must not only be able to cast a ballot, but also understand the structure of their preferences, learn to make proposals, and then second the proposals of others. Although matters are simplified since the computer performs all calculations for individuals and subjects can view their indifference curves, still the task in this experiment is demanding. The data from the experiment supports such a point. In three of five no-cost and in all five cost experiments at least one subject made less than ten percent of the proposals. In some cases those subjects made fewer than 5 percent of the proposals. Likewise, in all ten of the experiments at least one

subject called a vote less than 10 percent of the time. Subject inactivity is taken as an indicator that they incompletely understood the structure of the experiment.

The second problem is that not all subjects may understand the treatment. Although they are explicitly told that in the experiment they are assessed costs only if they vote for an amendment and if it passes, subjects can be confused that they will be charged for changes to the status quo regardless of their vote. This seems plausible given that an average of 17.2 percent of the votes taken between the status quo and an amendment were "wrong" in the agenda cost experiments. That is an individual voted contrary to what self-interest would predict. By contrast, in the no-cost experiments, on average, individuals voted against their self-interest only 1.1 percent of the time.

Given the apparent confusion by subjects in these experiments and their incomplete understanding of the treatment, it is difficult to draw conclusions from this data. Does a model of transaction costs have any empirical validity? The answer to this question is certainly not clear given the experimental design. But, that is the point. Subject misinterpretation can present a serious threat to the internal validity of an experiment.

Motivation

The motivation of subjects poses a second major threat to internal validity in laboratory experimental settings. Models of behavior always assume some form of individual motivation that serves as the "glue" for the model. Regardless of one's subject of study, whether it is institutional constraints on collective behavior or bargaining strategies pointing to the sources of war, all models impose some form of behavioral motivation on the actors. Over the past 30 years formal models of rational choice have made tremendous inroads in political science by relying on the simple motivational principle that actors will select strategies which maximize their utility. Even models which are not explicitly rational choice models contain motivational principles holding together the model. Models of cognition, for instance, hold that actors are motivated to make choices which are consistent, while learning models hold that actors adopt

strategies in patterned ways based on past experience and inferences from the experiences of others. All causal models of individual behavior contain a motivational assumption tying strategies to outcomes. The causal glue holding each model together is the motivation leading actors to choose a particular set of strategies.

Motivational assumptions are critical for a model. As such, these principles are also critical for laboratory experiments. If a model and the experiment purporting to test it are at variance over the motivations of actors, then it is unclear what the experiment is testing. In order to determine the robustness of a theoretical model, it is necessary for an experimenter to focus on the isomorphism between the model and the experiment. This is especially so when ensuring that the motivation of actors, which serves as the glue for a model, is the same in both theory and experiment.

Vernon Smith (1982) proposes four different threats to microeconomic experiments that focus on subject's motivations. While Smith's concern is with microeconomic experiments, these points translate well to any laboratory experiment concerned with models of human actors in collective choice settings. Smith's general point is well taken, since he argues that individuals participating in experiments are expected to perform tasks. Theoretical models predict how actors will perform those tasks, which typically means that subjects have to choose among several strategies in carrying out a task. Models predict, based on an assumption about motivation, which strategies will be used, and consequently, which outcomes will result. As such, Smith places a powerful emphasis on rewards to subjects in experimental settings.

Smith's first point is that laboratory experiments must build on the principle of <u>non-satiation</u> (1982, p. 931). Quite simply this means that given a <u>costless</u> choice between two alternatives and where choice A yields more of the reward medium than choice B, autonomous actors choose A.[4] While a fundamental principle of utility maximization, non-satiation

4 What constitutes a "choice" in this setting is left ambiguous. A choice might be nothing more than a selection between two rewards, in which case the interpretation of choice is as used in its common sense. However, a choice could also be the selection of a strategies, which incorporates a series of choices leading to an outcome. In this case, something akin to a "sophisticated" choice could be made in which early choices, yielding potentially greater rewards are bypassed

translates into a simple point that, regardless of the theory's motivational assumptions, actors do not tire of the reward medium (whatever it may be).

Smith's second point is that the rewards should be salient to the subjects (1982, p. 931). Saliency implies that the structure of the institution is sufficiently well defined such that subjects know the correspondence between the institution and the reward. This means that the reward is solely a function of the experiment and not of things unrelated to the experiment. In this sense the reward earned by subjects should be embedded in the experiment. Motivation, too, is fixed in the institution through the reward structure.

The third point is one of dominance (Smith, 1982, p. 934). Dominance requires that the reward structure embedded in the experiment be sufficiently desirable to override any costs or values that lie external to the experiment. Since volunteers in laboratory experiments can always do something else with their time (work, socialize, perhaps even study), the reward must exceed these opportunity costs. Otherwise, what motivates subjects is unclear -- whether the experiment or the subject's own external concerns.

The final point Smith makes is that the reward structure should be private (1982, p. 935). In other words, actors should know only their own reward schedule, and not that of others. This seeks to prevent actors from valuing something beside the reward medium. Two problems emerge if subjects know the rewards of others. First, instead of being motivated by their own rewards, they may be motivated by the rewards of others. In this sense a subject may choose strategies purely contingent on the rewards of others. While interesting questions (and indeed models) of actors minimizing or maximizing the payoffs of others might be raised, it is impossible to behaviorally sort out what motivates subjects if some seek to minimize the rewards of others, some subjects focus on maximizing the reward to a single individual, and still others ignore the rewards gained by other subjects. A second problem with making the rewards publicly known is that subject's may ignore the reward medium, converting the experiment into

---

in favor of choices that lead to more of the reward in the long run. See for example the model in an agenda setting context discussed by McKelvey and Niemi (1978).

a different "game" in which they now compete with one another to gain the most (or the least) of the reward medium. In part this has the potential to restructure the salience of the reward medium. Results from these experiments may not be anything close to that predicted under a theoretical model. In effect assuring privacy as to a subject's reward schedule provides another control over the motives of subjects so that the experimenter can be certain about what is driving subjects.

In the absence of non-satiation, salience, dominance or privacy of rewards, researchers are left with numerous threats to the interpreting experimental results. Since theoretical models generate predictions for experiments based on assumed motivations for subjects, if the motivations of subjects can be questioned, because of lapses in experimental design, then the results of the experiment are unclear.

A Two-Person Non-cooperative Bargaining Experiment.

An example of an experiment with mixed motivations is based on a series of two-person non-cooperative bargaining games (see Morgan and Wilson, 1989). The question under investigation is: why do countries in conflict break off negotiations and resort to war? At the heart of this question was why do negotiations sometimes fail? The model driving this research stems from formal theories of n-person spatial games. In its most simplified sense, the model has five components. First, a pair of autonomous actors are placed in a conflict setting. Second, these actors are assumed to be rational utility maximizers and as such are given single peaked utility functions over a single dimension of conflict. Third, actors are allowed a message space in which they can propose points on the single dimension. Unanimity rule, the fourth component, is imposed in order to reach a settlement on a point by the actors. Finally, actors are given a measure of power by which they can unilaterally end negotiation.

With these elements of the model in place the model works in the following manner. Actors have only two choices: to negotiate a settlement or go to war. Negotiation is characterized a one party making a proposal and seeking agreement on it. If the other party

agrees to the proposal, negotiation is completed and actors gain the utility for the point on which they settled. Going to war is characterized as unilaterally ending negotiations and having a point on the policy space imposed based on the power of the actors. Going to war is part of a stochastic process in which power is defined as a probability distribution over outcomes, with war a draw from that distribution. Since actors are assumed to be utility maximizers, then a set of predictions for an arbitrary policy space, a set of utility functions, and a power distribution is possible. This prediction is known as the "bargaining range" (see Morgan, 1984). The bargaining range is that set of points where the value of reaching a negotiated settlement exceeds the expected value of going to war for both parties. Since the probability distribution defining power is known, each actor can calculate his expected payoff for breaking off negotations and compare this with proposed points of settlement. Proposed points with utility exceeding the expected value of going to war for both parties, then, are in the bargaining range. Rational utility maximizers, are predicted to settle on points in this range rather than go to war when the bargaining range is non-empty.

The experiment involved groups of two individuals who participated on computer terminals which were physically separated from one another. Participants were recruited from various undergraduate classes and the "commons" in student housing and were free to select the time in which they desired to participate. All subjects were promised that they would be paid in cash at the conclusion of the experiment and no information was provided about the expected earnings from the experiment. Participants were told that they would be in two distinct experiments, one in which they would participate with the computer, and the second in which they would participate with one another. The experiments were run with 4 or 6 individuals at a time. Participants, upon arriving for the experiment were allowed to select their own computer terminal and were kept physically separated. Each participant was randomly assigned to a two-person group and an experimental condition. Before beginning the experiment, participants worked through a short series of instructions at their individual terminal. These instructions

incorporated several tests, both on how to use the computer equipment and emphasizing key elements of the experiment.

The first experiment was in fact a "pre-experiment" designed to allow participants to earn an endowment for subsequent play. In the "second" experiment, individual payoffs could become negative, so the initial earnings were a means of ensuring that participants had an endowment for the experiment. Rather than "staking" each individual to a fixed amount prior to beginning the experiment, we chose to have member's "earn" their endowment. As Hoffman and Spitzer (1985) show in a different experimental context, participants respond quite differently to endowments they earn than to endowments that are assigned. Once the pre-experiment was completed, participants were given instructions on their individual terminals concerning the bargaining experiment. These instructions included several tasks in which they were required to give a proper response before moving on. The instructions informed members that they would be randomly assigned with another partner, that they would interact with that individual via their computer terminal, and that at the conclusion of the experiment they would be paid, in private, their earnings from both experiments. Before beginning the experiment, individuals were quizzed as to their randomly assigned identity and other components of the experiment.[5]

The task for participants in the experiment was to either reach agreement on a single point taken from a line or to allow the computer to randomly choose, from a normal distribution, a point on that line. The line, with integers ranging from 1 to 100, was displayed on the computer terminal. A participant was motivated by being assigned a specific point on that line, which constituted her "ideal point" for the proposal space. At that point an individual received her maximum payoff. Points further from a member's ideal point decreased linearly in value. Table 1 gives the ideal points, valuation, and utility functions for the participants. The only notable point is that, although participants' utility functions were linear, the utility functions for the players were asymmetric. Since individuals participated in two distinct

5 These instructions are available from the author upon request.

periods, in the first individuals were randomly assigned an ideal point. At the second period, the ideal points and utility functions of the players were swapped, though players were not informed in advance that this was to be the case. Using a mouse to point to an alternative on the line, an individual was always provided information concerning the location and value of that point to her.

<Table 1 About Hero

To motivate subjects to behave as utility maximizers the reward medium in these experiments was dollars. The first presumption was that the payoff structure satisfied the principle of non-satiation, since subjects were assumed to value more money rather than less. Secondly, the reward structure was thought to be salient, since the rules of the experiment were well defined. Subjects could quickly and easily find the value of any point by simply using their mouse and letting the computer calculate that value. In addition, there were given the likelihood for each point of being randomly selected if negotiations were ended. Subjects, then, were given explicit linkages between their earnings and how to participate in the experiment. In addition the payoffs were considered dominant, since the expected earnings for an experiment lasting less than 40 minutes was $8.85 — a rate well above minimum wage. Finally, all participants only were given information concerning their own payoffs, were explicitly told other participants would not have the same valuation for proposals, and were told they would be privately paid at the conclusion of the experiment. All in all, this design was thought to motivate subjects to maximize their dollar earnings, in accordance with the theoretical model.

During the experiment participants were given two distinct menus from which to choose. The first menu allowed individuals to propose an alternative on the proposal space to another player. Such a proposal was an indication that the member was willing to end the period, earning the value of that proposal. A choice was simply made by clicking a point on the proposal space, then affirming that this proposal was what the individual wished to send to her partner. Once a proposal was sent, the partner was notified as to which proposal was made, its location on the proposal space, and its value. That individual then chose to accept or reject the

proposal. The initial proposer was then notified of her partner's choice. If the proposal was rejected, both participants continued the experiment. If the partner accepted the proposal, the initial proposer was then given the choice to affirm or reject the proposal. If rejected, play resumed. If accepted, the period was finished and both players were credited with the amount earned from accepting this proposal. Thus, the experiment required unanimity for accepting a proposal.

A participant could also choose to unilaterally end the period. The second option on the menu enabled a participant to stop the experiment. If a player confirmed that this was what she wished to do, the computer randomly selected a point from a normal distribution. That point became the final proposal and players were paid their value for it. Players were fully informed as to the shape of the normal distribution, since the likelihood that a particular point would be selected by the computer was displayed and individuals could freely examine those likelihoods by moving the cursor across the proposal space.[6] In addition, ending the period was not costless. The individual choosing to end the period was assessed a fee of $1.00, an amount which was subtracted from his/her earnings. This was intended to replicate a feature of the model which holds that war is not costless.

A number of experimental conditions were investigated in the original experiments (see Morgan and Wilson, 1989), including differential power distributions and differential costs for bargaining. However, for this example we focus on the symmetric power case. Also, since bargaining costs and the sequence of the decision had no independent effect on outcomes, data for these experiments are pooled. In the symmetric power case, the probability of going to war was given by a normal distribution with a mean of 50 and a standard deviation of 17. Given the ideal points of the players, both had roughly equal chances that, if negotiations were broken off, the computer would choose an outcome closer to their ideal point. Given the expected

6 The values displayed were not the true probabilities. Instead, these values were displayed as the probability multiplied by 1000. In pre-tests, participants indicated that the extremely small values reported as probabilities made little sense to them, and they were not able to adequately discriminate across changes in those values. In looking at the larger magnitudes, however, they could see how fast the distribution dropped off.

value for "going to war," the cost of unilaterally ending negotiations, the positions of the various players, and their respective utility functions, the predicted bargaining range is the interval [46,58], The predictions, then, is that for experiments in which a negotiated settlement is reached, those outcomes will appear in the interval [46,58].

A total of 40 outcomes were obtained under the symmetric power condition.[7] Under the symmetric power condition, 27 of 40 outcomes were negotiated (67.5%). These negotiated settlements are plotted on Figure 6. From this Figure, it is clear that the bargaining range does not predict very well. Of the negotiated outcomes, only 5 fall in the predicted bargaining range (18.5%). Although there is considerable variance in choices, there is a substantial clustering of outcomes shifted toward player 2, whose ideal point is at the point 92.

<Figure 6 About Hero

Upon closer examination of these outcomes, it is apparent that there is an ill fitting match between the motivational assumption of utility maximization in the model and the motivations of subjects in the experiment. After all, the theoretical model assumes that actors use an expected value calculus to decide whether or not to settle on an outcome. In the experimental setting a plausible rival hypothesis is that actors use a simple behavioral rule-of-thumb: they accept proposed points that yield positive value in the reward medium. In these experiments, payoffs are private to each player and there are no opportunities for players to communicate their earnings to one another. The only communication allowed in the experiment is the location of a proposal, who proposed it, and whether or not the proposal is agreed to by the other party. Nonetheless, given the payoff functions used in this experiment, there is a narrow range across which both players gain positive (non-zero) payoffs. This range extends across the interval [60,70], and for the sake of clarity, we call it the "behavioral bargaining range."

7 These results were obtained from 31 different pairs of actors. A number of conditions were crossed in these experiments, including asymmetric power distributions. In addition, each pair made two independent decisions, although not always under the same experimental conditions.

If we replot the outcomes and now include the behavioral bargaining range, we find a substantial improvement in fit among the negotiated outcomes (see Figure 7). Now just under half the outcomes appear in this behavioral bargaining range. Even though power is approximately symmetrically distributed for the two players, because of the shape of the utility functions outcomes are skewed toward player 2. Note that player l's expected value for going to war exceeds that of settling on any point in the behavioral bargaining range. Yet neither of the players appear to use this expected value calculation.

<Figure 7 About Hero

Is this behavioral bargaining range is spurious? After all the fit for the negotiated outcomes is not perfect. To test this conjecture, we ran a second, limited series, of experiments in which the behavioral bargaining range was eliminated. To eliminate the behavioral bargaining range, the same parameters in the previous experiment were retained. However, the valuation of individual ideal points (and thus every other point) were decreased by $1.50. While this was more than sufficient to eliminate any segment on the proposal space where both players jointly gained a positive payoff the expected value bargaining range does not change. Our experimental results make it clear that eliminating the behavioral bargaining range has a profound effect on individual choices. Quite simply, players in this second experimental series negotiated significantly fewer outcomes than did players in the first series. Only two of ten outcomes were negotiated. Of these two, one was in the bargaining range predicted using an expected value calculation. What is striking about these results is that as the behavioral bargaining range was eliminated, the likelihood that players resort to war increases dramatically. Consequently, it appears that there is a fundamental mismatch between the motivations driving the theoretical model and those driving subjects in these experiments. The former assumes that actors have the ability to make complex calculations, while the latter relies on a simple rule of thumb.

Conclusion

Laboratory experiments provide a powerful tool for political scientists. Yet, as with an tool, the researcher must exercise caution. Laboratory experiments present many pitfalls for the wary and unwary alike. Not only are problems of external validity endemic to experimentation, but problems of internal validity remain omnipresent, taking on many guises.

Part of the argument in this paper is that threats to external validity are not serious problems for researchers conducting laboratory experiments. By their very nature, such experiments will be simplifications and abstractions of a more complex environment. Generalizing from the experimental setting to a natural setting, then, will be problematic. This, however, does not excuse the experimenter from designing experiments that address problems in natural settings., It only says that an experimenter must not use the laboratory to build something isomorphic with the world, but instead look elsewhere for the design of an experiment. Experiments should aim at testing theoretical constructs. To this end, theoretical models should dictate the design of experiments.

If laboratory experiments are designed to test theories, then problems of external validity fade and threats to internal validity increase. At this point experiments must be aligned with what the researcher intends to test. Certainly there is no dearth of theories in political science. Not all are amenable to empirical test — ordinarily due to problems in instrumentation and measurement. Laboratory experimental methods allow the researcher to impose controls and design crucial tests of a theory. However, this can only be accomplished if a match is achieved between theory and experiment. To do so, fundamental threats to internal validity must be resolved. Most of these threats can be taken care of through careful experimental design.

Of particular concern to political scientists, especially as we deal with collective choice problems, are problems of internal validity concerning subject misinterpretation and subject motivation. That these can be important problems was illustrated by two experimental designs. In both instances the experimental data appeared to have patterns, but not as predicted by theory. Both designs, however, had problems with internal validity, making it difficult to

interpret what was happening in the experiment. In future revisions, these designs will be altered to overcome these flaws.

The general point to this paper is quite simple. It involves cautioning political scientists to take care in the design of laboratory experiments. The laboratory is the one place in which we can impose strict controls. Consequently our experimental designs should strive to meet this potential. It will be attained only by following the lead of theory and avoiding threats to internal validity.

References

Campbell, Donald T. and Julian C. Stanley (1963) Experimental and Quasi-Experimental Designs for Research. Rand McNally College Publishing Co.: Chicago.

Festinger, Leon (1953) "Laboratory Experiments" In Research Methods in the Behavioral Sciences, edited by Leon Festinger and Daniel Katz. New York: Dryden Press, pp. 136-172.

Fiorina, Morris (1975) "Formal Models in Political Science" American Journal of Political Science V. 19: 133-159.

Fiorina, Morris and Charles R. Plott (1978) "Committee Decisions Under Majority Rule: An Experimental Study" American Political Science Review V. 72.

Herzberg, Roberta Q. and Rick K. Wilson (1989) "Effects of Agenda Access Costs in a Spatial Committee Setting" Unpublished Manuscript, Rice University.

Hoffman, Elizabeth and Matthew Spitzer, "Entitlements, Rights, and Fairness: An Experimental Examination of Subjects' Concepts of Distributive Justice" Journal of Legal Studies V. 14 (June, 1985), 259-297.

McKelvey, Richard (1976) "Intransitivities in Multidimensional Voting Models and Some Implications for Agenda Control" Journal of Economic Theory, V. 16: 472-482.

McKelvey, Richard (1986) "Covering, Dominance, and Institution Free Properties of Social Choice" American Journal of Political Science V. 30: 283-314.

McKelvey, Richard and Richard Niemi (1978) "A Multistage Game Representation of Sophisticated Voting for Binary Procedures" Journal of Economic Theory, V. 18: 1-22.

McKelvey, Richard, Peter C. Ordeshook and Mark Winer (1978) "The Competitive Solution for N-Person Games Without Sidepayments" American Political Science Review V. 72.

Morgan, T. Clifton (1984) "A Spatial Model of Crisis Bargaining" International Studies Quarterly 28, 407-426.)

Morgan, T. Clifton and Rick K. Wilson (1989) "The Spatial Model of Crisis Bargaining: An Experimental Test" Paper presented at the 1989 Annual Meeting of the International Studies Association, London, March 28 - April 1, 1989.

Plott, Charles R. (1979) "The Application of Laboratory Experimental Methods to Public Choice." In Collective Decision Making: Applications from Public Choice Theory,

edited by Clifford Russell. Washington, D.C.: Resources for the Future.

Schofield, Norman (1978) "Instability of Simple Dynamic Games" Review of Economic Studies V. 45: 575-594.

Shepsle, Kenneth and Barry Weingast (1984) "Uncovered Sets and Sophisticated Voting with Implications for Agenda Institutions" American Journal of Political Science V. 28: 49-75.

Smith, Vernon (1982) "Microeconomic Systems as an Experimental Science" American Economic Review 72: 5 (December): 923-955.

Wilson, Rick K. (1986) "Results on the Condorcet Winner: A Committee Experiment on time Constraints" Simulations and Games V. 17: 217-243.

Table 1

Participant's Ideal Points and Utility Functions for the Experiment

Series 1

Player Function	Player Ideal Point	Player Maximum Value	Player Utility
1	4	$8.00	$8.00 - ($.1212 * |4 - X|)
2	92	$8.00	$8.00 - ($.25 * |92 - X|)

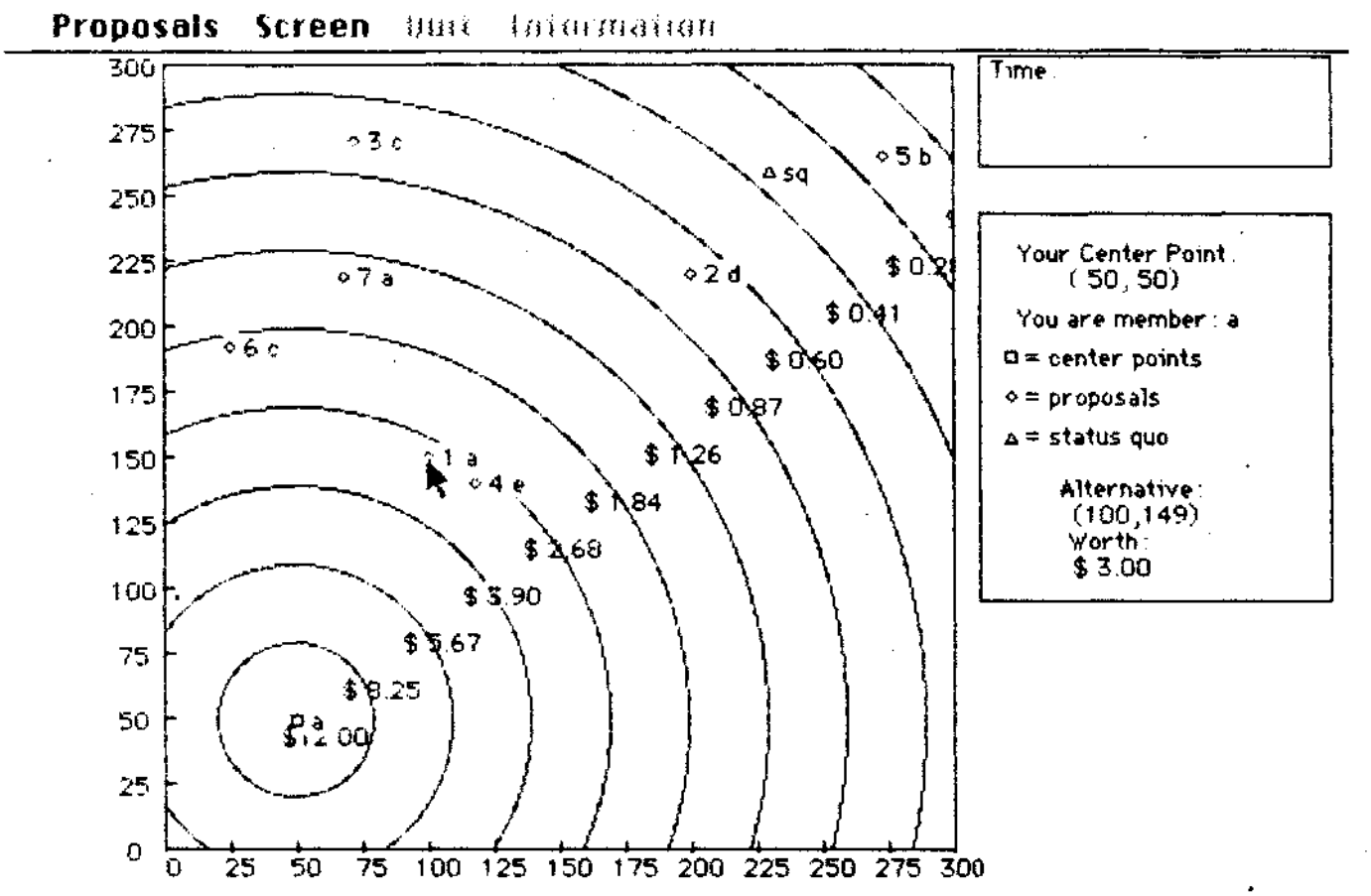
Series 2

Player Function	Player Ideal Point	Player Maximum Value	Player Utility
1	4	$6.50	$6.50 - ($.1212 * |4 - X|)
2	92	$6.50	$6.50 - ($.25 * |92 - X|)

where: X is an integer from the line (1,100).

Figure 1

Screen for Committee Experiments And Utility



Ideal Points and Utility Functions Used in Experiments.

Member	Ideal Point (ID)	Utility Function (X = arbitrary point)
1	(22,214)	$18.00 - ||X - ID|| * .07
2	(171,290)	$18.00 - ||X - ID|| * .07
3	(279,180)	$18.00 - ||X - ID|| * .07
4	(225, 43)	$18.00 - ||X - ID|| * .07
5	(43, 75)	$18.00 - ||X - ID|| * .07

Figure 2

Star Preferences with $.75 Costs

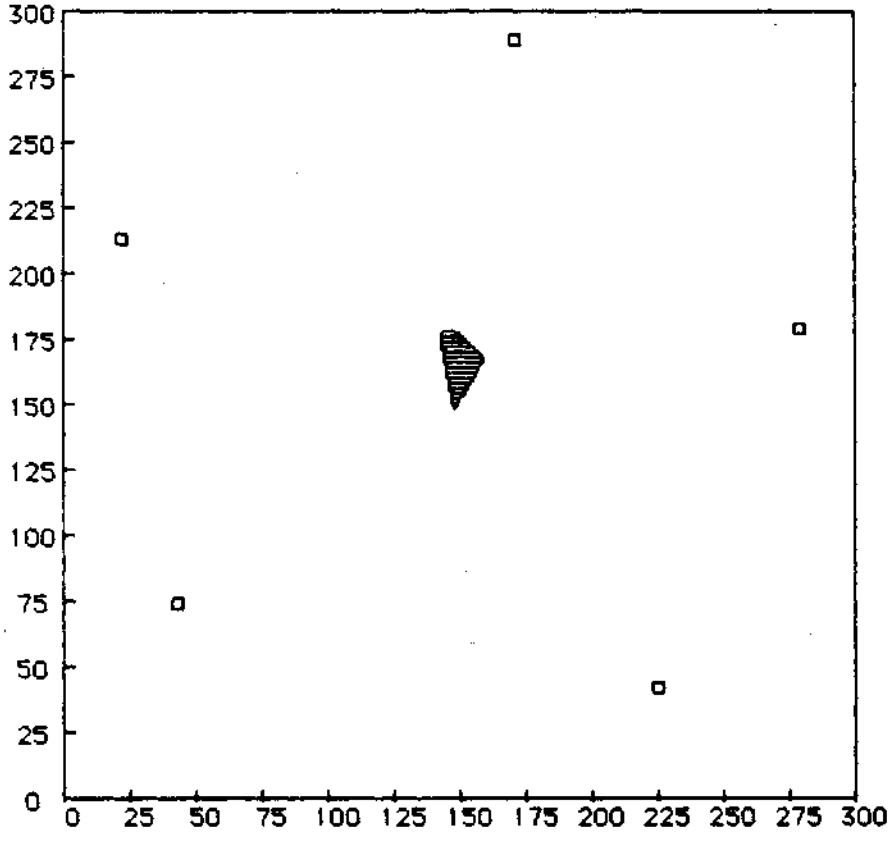
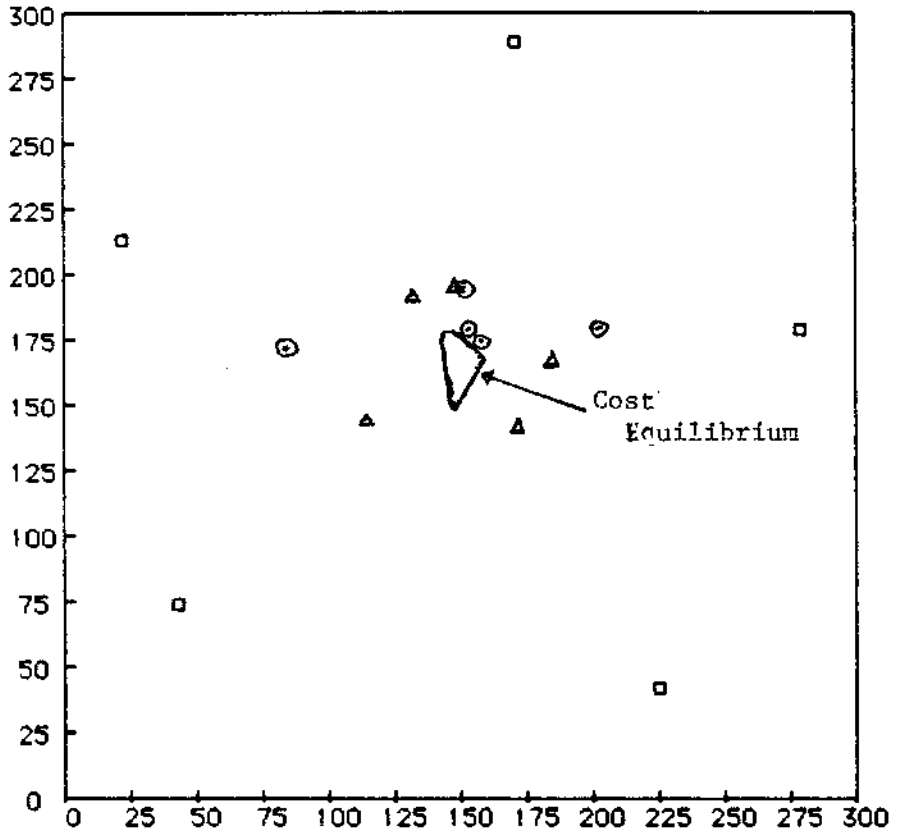


Figure 3

Plot of Committee Outcomes

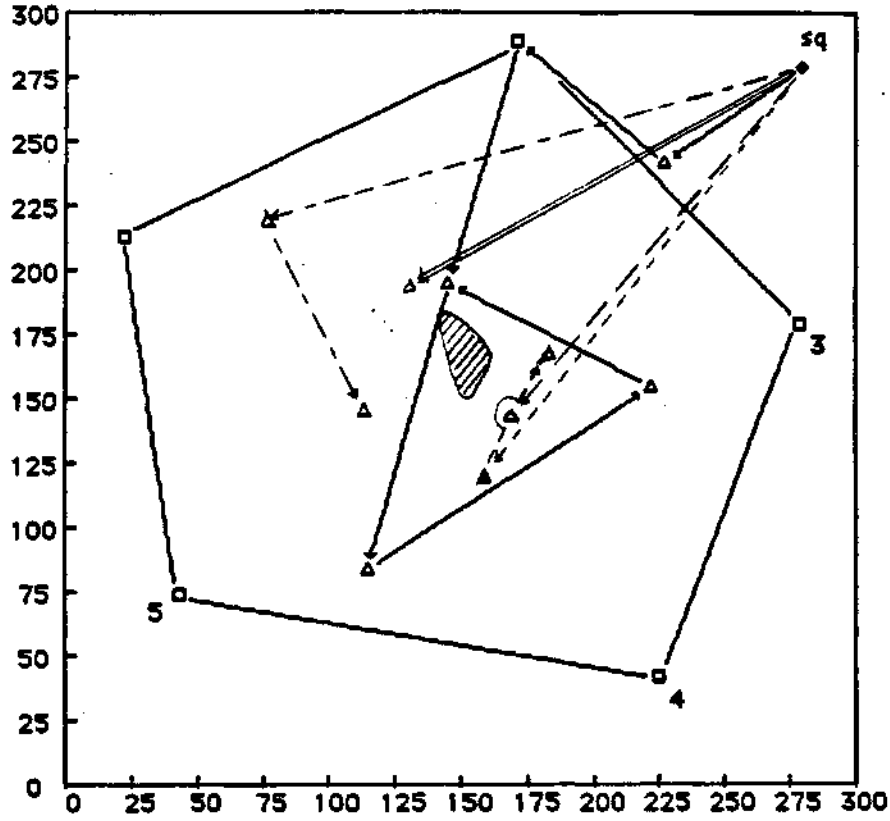


⊙ = No Cost Outcomes

△ = Cost Outcomes

Figure 4

Cost Treatment Committee Agendas



□ = Member Ideal Points

⟹ = Agenda Trajectory for Experiment: Cost1

- - - → = Agenda Trajectory for Experiment: Cost2

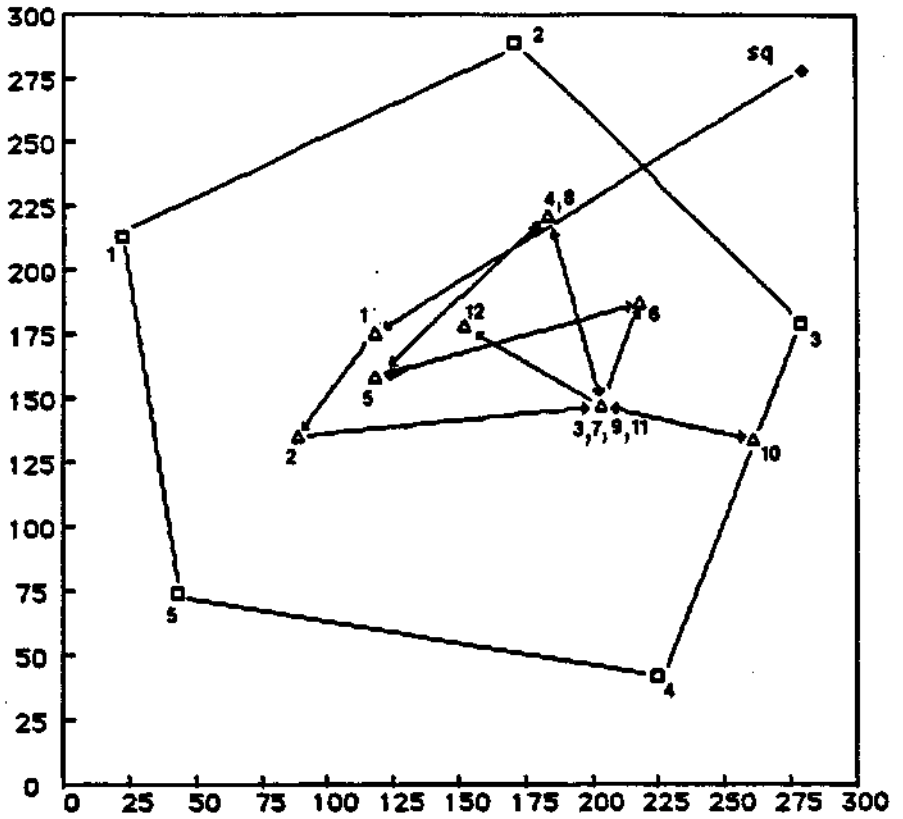
— — → = Agenda Trajectory for Experiment: Cost3

⟶ = Agenda Trajectory for Experiment: Cost4

— → = Agenda Trajectory for Experiment: Cost5

Figure 5

Agenda Trajectory for No Cost Experiment #4



□ = Member Ideal Points

⟶ = Agenda Trajectory for Experiment

Figure 6

Negotiated Outcomes for the Symmetric Power Condition

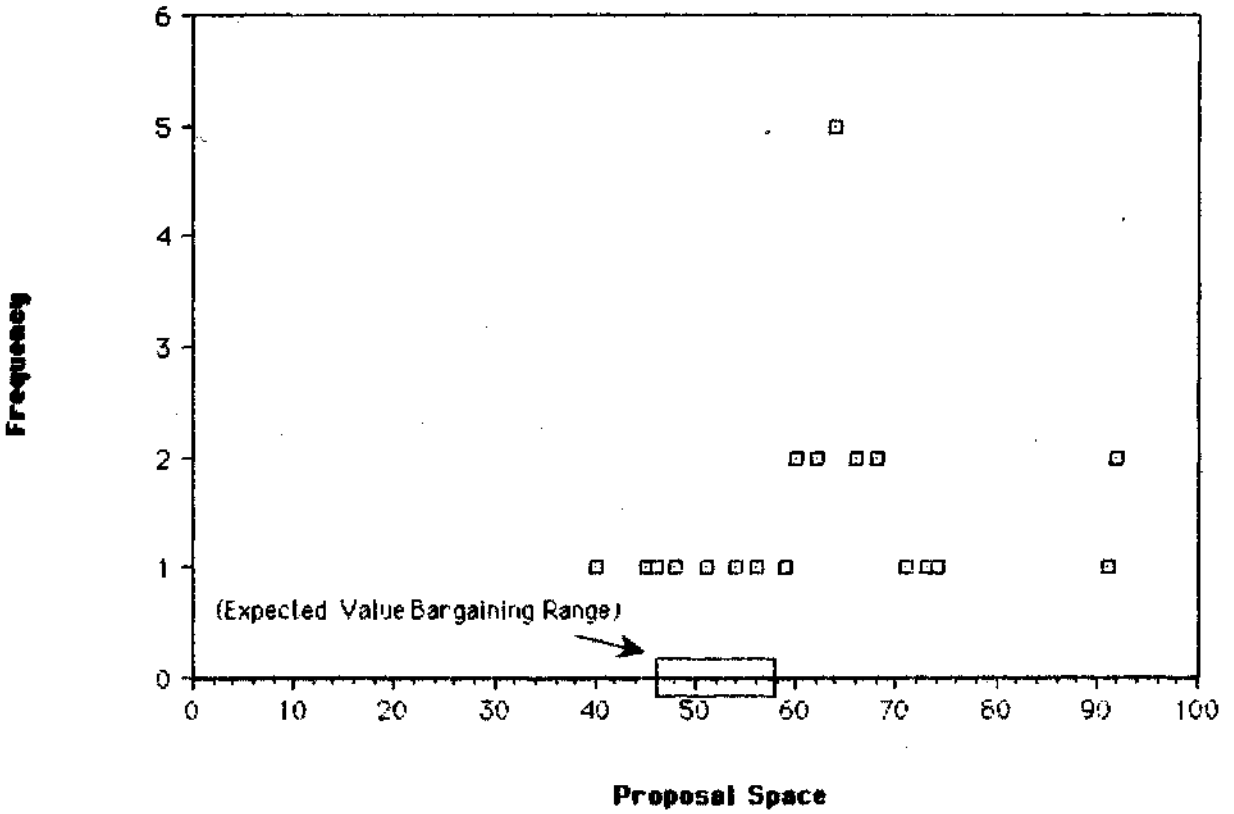


Figure 7

Negotiated Outcomes for the Symmetric Power Condition