Published in Bernhard Kittel, Wolfgang Luhan, and Rebecca Morton (eds). 2012. Experimental Political Science: Principles and Practices. Palgrave Macmillan

Laboratory Tests of Formal Theory and Behavioral Inference

Jonathan Woon*
University of Pittsburgh

May 25, 2011

> "One searches in vain for a detailed discussion of exactly how and when a model should be applied…If theorists blithely continue to prove more theorems, and applied scientists doggedly continue to gather more data, at some point data and theory might just miraculously conjoin. But we regard such a union as more likely to result from a determined effort than from a fortuitous accident." (Fiorina and Plott 1978, 576)

Political scientists seek to answer questions about political behavior, institutions, and outcomes. Why do (or don't) people cooperate to achieve common goals? To what extent do elections induce politicians to follow the wishes of the public? Why is government unable to enact new laws demanded by popular majorities? At the most general level, the method of advancing scientific knowledge of politics involves developing theories and testing their predictions. Theories are often expressed in terms of models that are purposeful, abstract simplifications of the real world, and *formal theory* involves a set of concepts and methods for systematically constructing and analyzing *mathematical models*. Robert Powell (1999) provides a succinct definition of a model as a "*constrained, best effort to capture what the modeler believes*

* Associate Professor, Department of Political Science, woon@pitt.edu

to be the essence of a complex empirical phenomenon or at least an important aspect of it" (24, emphasis original).

But in order to assess whether or not a model accurately captures the "essence" of real-world politics, the empirical predictions from a model must be tested against systematic evidence. The results of empirical tests should then guide the theorist in refining or extending the model. Theory testing is therefore an essential component of what Roger Myerson (1992) calls a "*modeling dialogue*…a process in which theorists and empiricists work together interactively on the difficult task of finding tractable models that capture and clarify the important aspects of real situations" (64, emphasis original). Without this dialogue, theoretical and empirical analyses might develop on separate, possibly divergent, paths and scientific progress would be significantly inhibited, as Morris Fiorina and Charles Plott so bitingly allude to in the epigraph.[1]

The fact that Fiorina and Plott's call for a "determined effort" to join theory and data is found in one of the earliest laboratory experiments testing the predictions of formal political theory indicates that laboratory experiments can contribute a great deal to the modeling dialogue. Because the defining feature of laboratory experiments is the control that researchers have over the data generating process, one obvious advantage of experimental control is that researchers can easily generate data to test a theory when naturally occurring data are otherwise unavailable.[2]

---

[1] The concern about a gap between theoretical and empirical research has also motivated the Empirical Implications of Theoretical Models (EITM) movement in political science to deliberately encourage greater interaction between theorists and empiricists (Aldrich, Alt and Lupia 2008). Note that I am neither arguing that all models must be tested—see Clarke and Primo (2007) on this point—nor that all theory must be formal, but rather that *when* models are intended to formalize an explanation for an empirical phenomenon, the modeling dialogue is an important component of the research process.

[2] The use of random assignment—which political scientists often associate with experiments—is not a necessary component of an experimental theory test, although it can often be useful to employ random assignment to identify a causal relationship in terms of a "treatment effect."

In tests of applied game theory models, this involves implementing a game in the lab and observing whether subjects' choices accord with the theoretical predictions.

In this chapter, I argue that laboratory experiments play an important role in testing behavioral models because of the precise control that they afford.[3] To make this argument, I present a simple framework that can be used to understand the kinds of inferences that can be drawn from empirical tests of formal theory predictions and then compare the inferences that can be made from laboratory experiments and observational studies. I then discuss important design principles that follow from the framework and provide an illustration from an experiment that I conducted on democratic accountability. The example also illustrates other kinds of design decisions that a researcher faces when implementing a laboratory theory test.

1 A Framework for Inferences from Theory Tests

A formal model can be thought of as comprising three distinct models of an empirical phenomenon. First, there is a *model of the social situation* that describes who the players are, the possible choices they can make, the information that they observe, and the outcomes that result from each possible combination of choices. In game theoretic terminology, this is called the game form. Second, there is a *model of preferences* that describes each player's goals, wants, or needs in terms of a well-ordered ranking of the possible outcomes of the social situation. These preferences are typically represented by utility or payoff functions. Third, there is a *model of behavior* that completes the model by specifying the properties of how people choose their

[3] For extended discussion of epistemological issues related to experiments, see Morton and Williams (2010) and Bardsley et al (2010). Morton and Williams also provide an extensive discussion of design issues, as do Kagel and Roth (1995) and Camerer (2003). For recent reviews of laboratory experiments in political economy see Palfrey (2006, 2009).

actions.[4] In non-cooperative game theory, behavioral assumptions are captured by the notion of Nash equilibrium and related equilibrium concepts (e.g., subgame perfect equilibrium or Perfect Bayesian equilibrium).[5]

Once each component is specified, the completed model is analyzed and solved, and the results of the analysis will include empirical implications—that is, testable predictions about outcomes and behavior. The following conditional statement summarizes how *empirical implications* (*EI*) follow logically from the joint assumptions of the model (the game form G, preferences P, and behavioral assumptions B).

$$G \wedge P \wedge B \longrightarrow EI \tag{1}$$

If the data are consistent with the empirical implications, then it is reasonable to conclude that the model appropriately captures the essence of the phenomenon being modeled. But if the evidence instead disconfirms the empirical implication, then what inference can be made? As a matter of logic, statement (1) must be true if the theoretical analysis is done properly. Taking the contrapositive of (1) produces the equivalent statement:

$$\neg EI \longrightarrow \neg G \vee \neg P \vee \neg B \tag{2}$$

Thus, falsifying the prediction implies that any one of the model's three components might be "false."[6] In other words, because the model is a set of joint assumptions, the theoretical

[4] See Smith (1994) for a general three-part framework for understanding experiments consisting of the environment, institution, and behavior.

[5] In cooperative game theory and social choice theory, there is no explicit model of behavior. Instead, the solution concepts involve axiomatic properties of group choice, such as the core. Although the framework can be generalized by replacing the behavioral model with a solution concept (such as in Fiorina and Plott), I use the notion of a behavioral model because the discussion will revolve around non-cooperative game theory.

[6] Although I use the notation of propositional logic, I do not mean to imply that the goal of theory testing is to establish whether or not a model is "true" or "false." Because models are

implications of an empirical test must necessarily involve some degree of *indeterminacy*: we can infer that one or more of the components must be a *poor approximation* to the phenomenon being studied, but it might be difficult to know for sure which the culprit is.

2 Observational Versus Laboratory Theory Tests

Applying the framework helps to highlight a crucial difference between observational studies and laboratory experiments, and this suggests an important role for laboratory experiments in the modeling dialogue. Two questions are of interest. First, given a falsified prediction, is it possible to resolve, or at least narrow down, the theoretical indeterminacy by establishing the empirical validity of some of the model's assumptions? For instance, establishing the validity of the game form would narrow down a model's failure to preferences or the behavioral model. Second, given that the theoretical indeterminacy is, or is not, resolved, what should the next step in the modeling dialogue be?

Although it may sometimes be possible to gather sufficient evidence, the lack of control a researcher has over the data generation process in an observational study more often than not makes it difficult, if not impossible, to make a compelling case about the empirical validity of a model's assumptions. As a result, the theoretical indeterminacy will remain unresolved. Falsification might imply that the game form is improperly specified, or that the assumptions about preferences are wrong, or that the behavioral assumption is inappropriate. It could also be due to the interaction of two components, or even the interaction of all three.

Typically, the theoretical response to observational falsification is to hold behavior B, and sometimes preferences P, constant while varying the model of the situation from G to G' in order

---

simplifications of reality—and false by definition—I merely use truth values as a shorthand to mean whether or not the model sufficiently captures the essence of the situation.

to produce a new model that better predicts and explains the data. For many research topics, such as the study of institutions, this appears to be a sensible and practical methodological move insofar as the focus is on identifying key features of the institution. For instance, congressional scholars debating the role of parties in the lawmaking process hold preferences (e.g., single-peaked preferences over a unidimenisonal policy space) and behavioral assumptions (e.g., sequential rationality characterized by subgame perfect equilibrium) constant while varying the key players and order of moves in the game form. Pivotal politics models (Krehbiel 1998) assume that the median legislator first proposes policies followed by the approval (or disapproval) from the veto and filibuster pivots (which represent supermajority voting rules). In contrast, party gatekeeping models (Cox and McCubbins 2005) assume that a party leader first decides whether or not to allow an issue onto the agenda followed by a policy choice by the median legislator.

Daniel Diermeier and Keith Krehbiel (2003) tacitly acknowledge the theoretical indeterminacy and defend the practice of holding behavioral assumptions constant on the grounds that "regular changes in behavioral postulates essentially guarantees that the field of study will fail to be cumulative" and if "the behavioral postulate were abandoned, then all of the prior institutional theories that contributed to the base of knowledge would have to be re-analyzed to gain comparability" (129). While frequent changes in behavioral postulates that prevent progress are certainly undesirable, holding behavioral assumptions constant in order to facilitate progress is defensible only up to a point. It can be defended to the extent that existing knowledge supports the maintained behavioral assumptions as good approximations of reality or if it is impossible to evaluate the assumptions empirically. But because laboratory experiments provide important tools for evaluating behavioral models, clinging to behavioral assumptions is ill advised for at

least two reasons.[7] First, if more and more elaborate theories are built upon a foundation that ultimately collapses, a lot of "progress" will be wasted when it comes time to rebuild them. Second, and more importantly, denying alternative behavioral assumptions severely limits the classes of theories and explanations that can be developed and therefore inhibits scientific progress. The strict rational choice approach explains institutions as optimal collective choices (given preferences and constraints), while alternative behavioral theories might explain institutions as myopic choices or the results of adaptive processes.

In contrast to the persistent indeterminacy of observational studies, clearer inferences can be drawn from laboratory theory tests. The hallmark of laboratory experiments is the precise control that experimenters have over key features of the data generating process. While experimental control has the obvious advantage that researchers can generate data that otherwise would not be available for theory testing, the more important consequence for the purposes of this discussion is that control ensures that the rules governing subjects' interaction in the laboratory exactly match the social situation described by the game form G. Doing so therefore immediately narrows (by disjunctive syllogism) the theoretical indeterminacy to preferences or behavior. Experimenters can also ensure that monetary payoffs from outcomes exactly match the preference structure of a model. By thus "inducing value" (Smith 1976), experimenters attempt to control preferences P, although this control is sometimes imperfect when subjects have intrinsic preferences (such as for fairness or other-regarding preferences). When control of preferences is successful, knowing that G and P correspond exactly to the formal model must then mean that a falsified prediction necessarily implies that the problem lies with the behavioral

[7] A better defense of the rational choice paradigm is that analyses making strong assumptions such as Nash equilibrium or perfect Bayesian equilibrium provide us with knowledge of an ideal benchmark and that theoretical analysis should also investigate the consequences of behavioral departures from idealized models.

assumptions *B*.[8] To be clear, the proper inference is not that *B* is "universally false," but that the model of behavior *B* does not accurately capture the essence of behavior in situations *G* with preferences *P*.

Although the asymmetric nature of falsification implies that unambiguous conclusions can be drawn from disconfirming a prediction in the lab, I do not mean to suggest that confirming or verifying the predictions of a theory are any less important. To the contrary, it is easy for many applied and empirically oriented researchers in political science and other social sciences to dismiss the sparseness of a behavioral model such as Nash equilibrium or its rationality requirement out of hand. Experiments therefore provide evidence that a model "works" and should be taken seriously and therefore can be applied to explain behavior outside the lab. For example, experiments on jury voting and information aggregation support key predictions of Nash equilibrium theories, as Thomas Palfrey summarizes: "many of the choice behaviors predicted by the equilibrium theories seem implausible, which makes their empirical validation in the laboratory quite surprising" (2009, 386).

Whereas the conventional response to observational falsification is to vary the game form or preferences, experimental falsification often leads to new models of behavior. Of course, these models need not involve wholesale rejection of the previous model. New models may retain key components of the old model that work well while altering the model where the assumptions are

[8] However, precise control over preferences is not always successful. Indeed, there is a vast literature on "social preferences" (e.g., Charness and Rabin 2002) that suggests how laboratory payoffs do not always correspond to subjects' actual preferences. Although this presents a potentially serious confound to making inferences about behavior, it does not mean that it is impossible to induce preferences. For example, control is likely to be successful in settings with common values or binary outcomes. The structure of preferences in the model and how closely it can be implemented in the lab is a design issue that must be considered by the experimenter. For an expanded discussion of this issue, see Bardsley et al (2010, chapter 3).

weakest. For instance, *quantal response equilibrium* (McKelvey and Palfrey 1995, 1998) retains the core principle of Nash equilibrium that players' strategies are mutual best responses but differs in that players' actions are subject to stochastic error (as in probabilistic choice models). Alternatively, *level-K* or *cognitive hierarchy models* (e.g., Nagel 1995, Camerer, Ho and Chong 2004) dispense with the equilibrium assumption of mutual consistency of beliefs and actions while retaining the assumption that players' strategies are best responses, but allow players to have mistaken beliefs about the behavior of others.

While control of the game form and preferences ensure that behavioral models can be tested, it is important to recognize that laboratory experiments have an important limitation: they cannot be used to definitively test a formal theory's assumptions about the game form and preferences. This issue is usually referred to by political scientists as "external validity," which is the extent to which the inferences from a particular sample (laboratory or otherwise) generalize to other samples.[9] More precisely, a limitation of laboratory experiments for theory testing is that they cannot establish the correspondence between a formal theory and the real-world. The reason is that the experimenter can implement *G* and *P* as stated in the model, but the empirical validity of *G* and *P* can only be established using non-experimental data. In other words, experimental researchers must share the same concern as formal theorists about whether *G* and *P* accurately represent the real-world.

Nevertheless, as I have argued, experiments play an important role in testing behavioral assumptions. Laboratory experiments therefore help to narrow down the set of behavioral models that might work outside the lab. That is, if a given behavioral model doesn't work in the lab when *G* and *P* are known, there is good reason to be skeptical that it won't work to explain behavior

[9] See Levitt and List (2007) for a discussion of various concerns about the generalizability of laboratory experiments.

outside the lab either. If *B* doesn't work in simple, controlled settings, why should it work in complicated, messy ones? Thus, laboratory experiments should be viewed as complements, rather than substitutes, to observational studies: when experiments are used to pin down behavioral assumptions, observational studies can draw stronger inferences about the whether the game form and preferences accurately capture the essence of the real-world situation.

3 Essential Design Principles

The framework suggests two simple, yet essential, design principles that should be followed in order to make meaningful, unambiguous inferences about the applicability of behavioral models:

1. *The experiment should implement the game form G and preferences P as closely as possible*.
2. *Subjects must understand the rules of the game but should not be told, or given suggestions, about how to behave*.

The first principle ensures that the results of the experiment can be interpreted clearly in terms of the behavior model while the second principle ensures that behavior arises endogenously from subjects' reasoning and thinking (otherwise, the experiment would be a simulation). Subjects must understand the rules clearly, but this does not mean they are required to understand or perceive the game the same way that a game theorist does—after all, whether or not they do so is an empirical question.

Of course, in practice, adhering to these principles may not always be simple, especially for complicated games, or when the literal implementation of a game may conflict with other design considerations. For some purposes, there may also be good reasons for departing from

these principles. Nevertheless, if the primary goal of the experiment is to make inferences with respect to behavioral models, then following these principles is crucial.

A simple experiment from the early history of game theory attests to the importance of the first basic principle. In 1950, Merrill Flood and Melvin Dresher devised a simple 2×2 (simultaneous move) matrix game (Flood 1958) to test John Nash's recently developed non-cooperative equilibrium theory (Nash 1951). The game payoffs, given in pennies, are described in Figure 1 (the first line in each cell). The game has a dominant strategy Nash equilibrium (Down, Left), but both players would be better off from the outcome (Up, Right) that results from playing dominated strategies (i.e., the Pareto superior outcome is not a Nash equilibrium). The structure of this game would later become widely known as the Prisoner's Dilemma.

Flood and Dresher had the same two subjects (AA and JW) play the game 100 times, and they regarded each play as an independent observation. The observed frequencies of play in Figure 1 (in parenthses) show that both subjects tended to play the dominated strategy and that the modal outcome is the Pareto superior (Up, Right) rather than the Nash equilibrium (Down, Left). Flood and Dresher regarded their finding as evidence against Nash equilibrium.

[Figure 1 about here.]

Although Flood and Dresher's procedures would not survive the peer review process today (and they admitted that their experiment was only preliminary), two aspects of the design and interpretation of the experiment are noteworthy. First, it is interesting to note that Flood and Dresher's purpose was explicitly behavioral:

> "To find whether or not subjects tended to behave as they should if the Nash theory were applicable, or if their behavior tended more toward the von Neumann-Morgenstern solution, the split-the-difference principle, or some other yet-to-be-discovered principle." (Flood 1958, 12)

Second, there is a poor correspondence between 100 plays of the experimental game between the same two subjects (who were also friends) and the single-shot nature of game they intended to implement. Today, their laboratory game would be quickly recognized as a (finitely) repeated game. Indeed, the lack of correspondence was noted by Nash himself (demonstrating a theoretical response in the modeling dialogue):

> "The flaw in this experiment as a test of equilibrium point theory is that the experiment really amounts to having the players play one large multimove game. One cannot just as well think of the thing as a sequence of independent games as one can in zero-sum cases. There is much too much interaction, which is obvious in the results of the experiment…If this experiment were conducted with various different players rotating the competition and with no information given to a player of what choices the others have been making until the end of all the trials…this modification of procedure would remove the interaction between the trials." (Nash's comments printed in Flood 1958, 16)

While closely following the game form and preference assumptions is important for behavioral inference, the point here is not that the implementation must always be literal. Some features of the design will also depend on the theoretical interpretation of the behavioral model

that go beyond their formal mathematical definitions. For example, in the case of testing the Nash equilibrium predictions of static games, some form of repetition is often necessary. This is the case if Nash equilibrium is interpreted as a stable outcome of the game played by experienced subjects. Thus, subjects must play the game more than once to gain experience, and so that the researcher can observe whether or not the outcome is stable. One solution, as Nash points out, would be to use random matching without feedback. Thus, the basic lesson from this example is that when it is not always possible or ideal to implement a game literally, experimenters must nevertheless have a clear understanding of how the behavioral assumptions apply to the game that is actually implemented. Usually, this is achieved by conducting additional theoretical analysis. When such analysis is not possible or if no proof of the empirical implication is otherwise possible (e.g., the truth of (1) is not known), the experiment should be thought of as a "stress test" rather than a direct test of the behavioral model.

4 Illustration: Democratic Accountability

To illustrate the kinds of behavioral inferences that can be made from an experiment that follows these principles, I will discuss an example from my own work on electoral accountability (Woon 2010). This illustration also demonstrates that when adhering to these two principles, a researcher still has wide latitude in designing the experiment and faces a number of important design choices. The additional aspects of the design that I will discuss are choices I made about the *complexity* of the game, the *parameters* to use, and the amount of *meaningful context*.

An important question in democratic theory concerns the extent to which elections provide incentives for politicians to act in the interests of voters. V.O. Key's (1966) traditional theory of retrospective voting suggests that elections should provide powerful incentives even

when voters do not fully understand the linkages between policies and outcomes. Voters only need to be able to judge the quality of outcomes, such as the health of the economy, and use a simple retrospective voting rule: reward (re-elect) politicians for good outcomes and punish them (kick them out of office) for bad ones. Traditional retrospective voting can be thought of as a *sanctioning* device.

Subsequent rational choice theorizing, however, identifies an important limitation of the traditional theory (e.g., Fiorina 1981). If voters are forward-looking, then elections should instead be viewed as *selection* devices: voters elect politicians based on expectations of future performance. Information about past outcomes is relevant only insofar as it helps to update beliefs about the incumbent. Several models formalizing the theory have identified conditions under which selection renders sanctioning an empty threat (e.g., Canes-Wrone, Herron and Shotts 2001, Fearon 1999). Moreover, the selection view implies that politicians will have incentives to *pander*. That is, they will choose policies that convince voters—and even mislead them—into thinking that the politician possesses desirable characteristics (thus raising voter expectations about future performance) even when the politician knows that such policy choices are not in the best interests of voters.

Because real elections involve many factors beyond those that the formal theories focus on, laboratory experiments are well-suited for creating a controlled environment for testing the behavioral predictions (and thus, the behavioral assumptions) of formal models of elections. In my experiment, I used the model of Fox and Shotts (2009) to investigate whether elections are better viewed as sanctioning or selection devices. I chose their model because it identifies conditions under which selection and sanctioning are observationally distinct as well as conditions under which they are observationally equivalent.

The model involves policy-making before and after an election, and the sequence of actions in the model is as follows:

1. Nature chooses a state of the world $\omega \in \{A,B\}$, the incumbent's type $t \in T$, the incumbent's signal $s \in \{A,B\}$, and the challenger's type $c \in T$.
2. The incumbent chooses a policy $p \in \{A,B\}$.
3. The voter observes both the policy p and the state ω, but not the type t, and chooses whether or not to re-elect the incumbent.
4. Nature chooses a new state ω' and a new signal s', and the elected politician (either the incumbent or challenger) chooses a new policy p'.

"Good" policies for voters are ones that match the state, while "bad" ones are ones that do not match. There are eight possible politician types in the model, which are characterized by three attributes: whether they are office or policy motivated, their level of expertise, whether they share the voter's policy preference or have an ideological bias for policy B.

The intuitive strategy for voters is to use a retrospective or *outcome-based rule*: vote for the incumbent if and only if the policy is "good" ($p = \omega$), in order to induce office-motivated politicians to use their expertise (follow their signals). But there exists a perfect Bayesian equilibrium—describing optimal strategic behavior—in which voters instead use a *policy-based rule*: re-elect politicians if and only if they choose $p = A$ (given that $\omega=A$ is more likely than $\omega = B$), which induces office-motivated politicians to ignore their signals and always choose $p = A$. This equilibrium can be thought of as a "delegate" or "pandering" equilibrium because voters only reward politicians for demonstrating that they do not have an ideological bias for policy B, and office-motivated politicians comply by always choosing A even if they know it is worse for

voters. Thus, the sequential rationality of perfect Bayesian equilibrium makes a distinctly different prediction about behavior than traditional retrospective voting.

4.1 Complexity

As someone who is knowledgeable about mathematical models, I recognize that I suffer from the "curse of knowledge" (Camerer, Loewenstein and Weber 1989). I might unconsciously perceive the game differently from subjects in the laboratory without the same knowledge. To mitigate the curse of knowledge, I try to imagine how I would think about the game if I were a subject in the lab. Although this is admittedly a subjective process, it led me to recognize that even though the game is an extreme simplification of real elections, it is nevertheless somewhat complicated—certainly more complicated than common experimental games such as a linear public goods game or the ultimatum game.

I therefore simplified several aspects of the game that would make the rules easier to convey, while still retaining the essence of the situation being modeled. One feature of the original model that was not absolutely necessary was that there is some probability that the policy issue in the second period is one of "common value"—that all types shared the same preferences. Removing this feature simplifies the game form while preserving the delegate equilibrium.

The second feature that I simplified was the nature of post-election policymaking. Rather than implementing Step 4 in the game sequence described above—having Nature choosing ω', s', and allowing the politician to choose p'—the voter receives payoffs based directly on the type of politician elected to office in the second period. These payoffs are identical to the expected value of the voter's payoffs based on the elected politician's type, so again the modification simplifies the game while preserving the delegate equilibrium. This simplification also has other benefits. It

reduces the amount of noise involved in subjects payoffs and ensures that subjects' clearly know the preference ordering over types. In other words, the modification involves greater control over subjects' preferences than if the realization of random variables and the politician's free choice in Step 4 were retained.

4.2 Parameters

Another important design consideration in testing formal theories is the selection of parameters. The model analyzed by Fox and Shotts has several exogenous parameters characterizing the distribution of states and politicians' types, and in a technical sense any feasible parameters can be used in the laboratory implementation of the game. However, some sets of parameters yield a greater number of useful observations for the purposes of discriminating between alternative theories (in this case, between selection/rationality and sanctioning/retrospection).

In the case of the accountability experiment, compare the outcome-based and the policy-based voting rules when they are described as strategies. In this case, a voter strategy specifies whether to elect the incumbent or challenger for each possible information set (each possible combination of policy and state). The two strategies are compared in Table 1. Notice that when the $\omega = A$, both strategies prescribe the same behavior. However, they make distinct predictions when $\omega = B$. Thus, in order to produce a sufficient number of observations where $\omega = B$, the probability of $\omega = B$ must be as high as possible subject to two constraints. The probability of state A must be higher than the probability of state *B* because this is a maintained assumption of the analysis. The distribution of states should also be easily represented as a simple fraction. While $\Pr(\omega = B) = 0.499999$ would technically satisfy the first constraint, it is so close to 0.5 that

subjects would not likely make the distinction, so I chose Pr($\omega = B$)=0.4. The probability can also be described as a "4 in 10 chance."

[Table 1 about here.]

The second parameter of interest is the probability that politicians are office-motivated. This is because only office-motivated politicians respond to differences in voter strategies. I chose to let this probability be $\frac{3}{4}$, described as a both a "75% probability" and a "three in four chance" which is large enough to ensure that many politicians would be office-motivated while also ensuring that a sufficient number of politicians would be policy-motivated and thus still contribute to voters' beliefs.

The third parameter of interest is the probability that politicians are ideological. The higher this probability is, greater the incentive that voters have to select politicians in order to avoid ideological types. I therefore let this probability be $\frac{3}{4}$, as it is sufficiently high to ensure a unique equilibrium while being simple to describe.

The final parameter of interest is the probability that politicians have perfect signals. If this probability is sufficiently low, then the delegate equilibrium (and the policy-based voting rule) is the unique pure strategy perfect Bayesian equilibrium. I let this probability be $\frac{1}{10}$, which ensures that voters will not much about politicians' quality and will therefore focus on avoiding ideological types. I also conducted sessions where this probability is $\frac{1}{2}$, which does not guarantee that the delegate equilibrium is unique, but enables me to test the comparative static prediction that delegate equilibrium voting behavior is less likely when the probability is higher.

4.3 Meaningful Context

The final design consideration is whether the game should be described to subjects in "context free" abstract terms or with real-world context. An abstract representation of the game might describe the players' roles as "Player 1" and "Player 2" while calling them a "Politician" and "Voter" would involve some meaningful context. Many experiments testing game theoretic predictions, such as the Flood and Dresher experiment, use abstract context.

One reason for using an abstract context is to try to ensure that subjects' decisions depend only on strategic factors (the game form and their incentives) and to minimize the potential influences of non-strategic or "psychological" factors that might be triggered by the presence of real-world context. For example, when the Prisoner's Dilemma is framed as a "Community Game" or "Wall Street Game" the level of cooperative behavior changes in a systematic and predictable way (Liberman, Samuels and Ross 2004). In this case, the experimenter has appeared to lose control over subjects' preferences, so using abstract context makes a lot of sense if the purpose of the experiment is to test abstract strategic thinking.

But removing real-world context also decreases the correspondence between the laboratory environment and the real-world situation that the experiment is intended to represent. The result may be a decision-making environment that is too artificial. Behavioral economist George Loewenstein argues that "the context-free experiment is…an elusive goal"—that is, an abstract setting is simply an unfamiliar context—and that "the goal of external validity is served by creating a context that is similar to the one in which…agents will actually operate" (Loewenstein 1999, F30) Thus, in testing applied models intended to represent a specific class of real-world situations, a balance must be struck between ensuring control over the game form and

preferences, and providing enough context so that the results may be interpretable in terms of real-world behavior.

An example from psychology of where some context increases the likelihood that subjects make objectively correct decisions is Wason's "four-card problem" (Wason and Shapiro 1971). In this task, subjects are given four cards, each with a letter on one side and a number on the other and are shown only one side of each card. Subjects are asked to turn over two cards in order to test the logical statement "a vowel on one side implies an even number on the other." When this task is presented in the abstract (with letters and numbers), only 10% of subjects' choices are logically correct. However, when a logically equivalent problem is presented using a familiar context, such as checking IDs to ensure that alcoholic beverages are consumed by persons over the minimum drinking age, the proportion of correct choices improves dramatically to 89% (Cosmides 1989).

Because the Fox and Shotts model is motivated by and intended to represent a specific context (elections) rather than a more general model intended to apply to a variety of contexts (e.g., collective action), some meaningful context is appropriate when implementing the game form in the laboratory. That is, the purpose of the experiment is to test voting behavior, but in a controlled setting where the game form and preferences match the formal model, so meaningful context is appropriate. I therefore chose to tell the subjects that they would be playing two different roles, "politicians" and "voters", and that politicians' attributes were described as "office-seeking" or "policy-seeking" "motivations", "pragmatic" or "ideological" "preferences", and "perfect" or "noisy" "quality of information."

In this case, the complexity of the game form is another reason for using meaningful context. Recall that there are eight politician types, characterized by three different kinds of

attributes. Using some context also helps to orient subjects by helping them to understand and remember the game form, which enhances experimental control by minimizing confusion.

4.4 Results: Unpacking Rationality

Table 2 presents the predicted versus observed choices made by subjects in the role of voters from the experiment. Voting behavior for each possible information set (combination of *p* and ω) is compared to the delegate equilibrium prediction. The data clearly disconfirm the game theoretic prediction, as voters appear to consistently employ an outcome-based retrospective rule rather than a policy-based rule. Given the close implementation of the game form and preferences assumed by Fox and Shotts, the data (when viewed in light of the framework) strongly suggest that the behavioral assumptions captured by the perfect Bayesian equilibrium concept are poor approximations of behavior in this setting.

[Table 2 about here]

Can we say anything more about behavior other than the fact that perfect Bayesian equilibrium is not applicable in this context? It turns out that we can. First note that the behavioral assumptions of perfect Bayesian equilibrium can be decomposed into three aspects of rationality: choices that are consistent with preferences (basic rational choice), choices that are best responses to other players' choices (strategic optimality), and beliefs that are consistent with Bayes' Rule (Bayesian inference). Subjects' behavior is consistent with the first two notions, which narrows down the failure of rationality to the third.

To see why this is the case we need to consider the behavior of politicians in the experiment. Their behavior is shown in Table 3, which presents the predicted versus observed choice frequencies. Policy-motivated politicians have non-strategic incentives—that is, their

payoffs do not depend on what voters do-and the data show that their actions are consistent with basic rational choice predictions. The data also show that office-motivated politicians do not use equilibrium strategies, but their actions in the experiment are best responses to the voter's outcome-based strategy. Subjects' behavior is therefore also consistent with strategic optimality.

[Table 3 about here]

To further understand why there was such a strong behavioral tendency for voters to use a retrospective or outcome-based rule, I conducted a set of additional treatments. I hypothesized that the failure of voters to use a strategy consistent with correct Bayesian inferences might have been for two reasons. One possibility was the complexity of the inference problem (updating beliefs about eight possible types). In the face of such complexity, people tend to rely on some sort of decision heuristic rather than trying to solve for the fully optimal strategy. A competing (but not mutually exclusive) hypothesis was that subjects used an outcome-based strategy because they wanted to induce politicians to use their information to choose the best possible policy.

Although I do not have space to provide the full details of the additional treatments, the main principle underlying the design of the additional treatments was to modify the game in such a way to remove the hypothesized source of non-equilibrium behavior while keeping the strategic incentives constant (i.e., the delegate equilibrium was still the unique perfect Bayesian equilibrium). In this way, game theory provides a basis for the null hypothesis of no difference between the baseline experiment and the behavioral treatment (since the strategic incentives are identical) while the behavioral hypotheses imply that voters are more likely to use the equilibrium policy-based rule in the modified games (treatments). Although the data from the

additional treatments remain inconsistent with the perfect Bayesian equilibrium point predictions, the treatment effects nevertheless provide support for both behavioral hypotheses.

5 Conclusion

In this chapter, I highlighted the advantages of laboratory experiments for testing formal theories. Precise control over the game form and preferences implies that laboratory methods are particularly well-suited for testing and investigating the behavioral components of formal models. My experiment on democratic accountability provides an illustration that not only falsifies the package of assumptions captured by perfect Bayesian equilibrium, but also that experimental methods can be used to pinpoint the source of an equilibrium concept's inapplicability. Experiments therefore play an especially important role in demarcating the bounds of rationality, identifying the circumstances when different rationality concepts do and do not apply.

While I have focused primarily on the benefits of laboratory experiments, it is also important to recognize their limitations. While controlled laboratory experiments permit strong inferences about *behavior given a social situation and preferences*, the strength of inferences about *real-world behavior* depends on how closely the laboratory environment corresponds to the relevant features of the real-world context the experiment is intended to investigate. The real-world validity of the artificial environment is therefore an important concern common to both laboratory experiments and formal theories. Laboratory experiments are best viewed as complements rather than substitutes for field and observational data.

References

Aldrich, John H., James Alt and Arthur Lupia. 2008. The EITM Approach: Origins and Interpretations. In *The Oxford Handbook of Political Methodology*, ed. Janet Box-Steffensmeier, Henry Brady and David Collier. Oxford University Press.

Bardsley, Nicholas, Robin Cubitt, Graham Loomes, Peter Moffatt, Chris Starmer and Robert Sugden. 2010. *Experimental Economics: Rethinking the Rules*. Princeton University Press.

Camerer, Colin. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Russell Sage Foundation and Princeton University Press.

Camerer, Colin F., Teck-Hua Ho and Juin-Kuan Chong. 2004. "A Cognitive Hierarchy Model of Games." *Quarterly Journal of Economics* 119(3):861–898.

Camerer, Colin, George Loewenstein and Martin Weber. 1989. "The Curse of Knowledge in Economic Settings: An Experimental Analysis." *Journal of Political Economy* 97(5):1232–54.

Canes-Wrone, Brandice, Michael C. Herron and Kenneth W. Shotts. 2001. "Leadership and Pandering: A Theory of Executive Policymaking." *American Journal of Political Science* 45(3):532–50.

Charness, Gary and Matthew Rabin. 2002. "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics* 117(3):817–869.

Clarke, Kevin A. and David M. Primo. 2007. "Modernizing Political Science: A Model-Based Approach." *Perspectives on Politics* 5(04):741–753.

Cosmides, Leda. 1989. "The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies With the Wason Selection Task." *Cognition* 31(3):187–276.

Cox, Gary W. and Mathew D. McCubbins. 2005. *Setting the Agenda: Responsible Party Government in the House of Representatives*. New York: Cambridge University Press.

Diermeier, Daniel and Keith Krehbiel. 2003. "Institutionalism as a Methodology." *Journal of Theoretical Politics* 15(2):123.

Fearon, James D. 1999. Electoral Accountability and the Control of Politicians: Selecting Good Types versus Sanctioning Poor Performance. In *Democracy, Accountability, and Representation*, ed. A. Przeworski, S.C. Stokes and B. Manin. Cambridge University Press.

Fiorina, Morris A. 1981. *Retrospective Voting in American National Elections*. New Haven, Conn.: Yale University Press.

Fiorina, Morris P. and Charles R. Plott. 1978. "Committee Decisions Under Majority Rule: An Experimental Study." *The American Political Science Review* 72(2):575–598.

Flood, Merrill M. 1958. "Some Experimental Games." *Management Science* 5(1):5–26.

Fox, Justin and Kenneth W. Shotts. 2009. "Delegates or Trustees? A Theory of Political Accountability." *Journal of Politics* 71(4):1225–37.

Kagel, John H. and Alvin E. Roth, eds. 1995. *The Handbook of Experimental Economics*. Princeton University Press.

Key, V.O., Jr. 1966. *The Responsible Electorate*. Belknap Press.

Krehbiel, Keith. 1998. *Pivotal Politics: A Theory of US Lawmaking*. Chicago: University of Chicago Press.

Levitt, Steven D. and John A. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? " *The Journal of Economic Perspectives* 21(2):153–174.

Liberman, Varda, Steven M. Samuels and Lee Ross. 2004. "The Name of the Game: Predictive Power of Reputations versus Situational Labels in Determining Prisoner's Dilemma Game Moves." *Personality and Social Psychology Bulletin* 30(9):1175–85.

Loewenstein, George. 1999. "Experimental Economics from the Vantage Point of Behavioral Economics." *The Economic Journal* 109:F25–F34.

McKelvey, Richard D. and Thomas R. Palfrey. 1995. "Quantal Response Equilibria in Normal Form Games." *Games and Economic Behavior* 10:6–37.

McKelvey, Richard D. and Thomas R. Palfrey. 1998. "Quantal Response Equilibria in Extensive Form Games." *Experimental Economics* 1:9–41.

Morton, Rebecca B. and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality*. Cambridge University Press.

Myerson, Roger B. 1992. "On the Value of Game Theory in Social Science." *Rationality and Society* 4(1):62.

Nagel, Rosemarie. 1995. "Unraveling in Guessing Games: An Experimental Study." *The American Economic Review* 85(5):1313–1326.

Nash, John F. 1951. "Non-Cooperative Games." *Annals of Mathematics* 54(2):286–295.

Palfrey, Thomas R. 2006. Laboratory Experiments. In *The Oxford Handbook of Political Economy*, ed. Barry R. Weingast and Donald A. Wittman. Oxford University Press.

Palfrey, Thomas R. 2009. "Laboratory Experiments in Political Economy." *Annual Review of Political Science* 12:379–388.

Powell, Robert. 1999. *In the Shadow of Power: States and Strategies in International Politics*. Princeton University Press.

Smith, Vernon L. 1994. "Economics in the Laboratory." *Journal of Economics Perspectives* 8(1):113–131.

Smith, V.L. 1976. "Experimental Economics: Induced Value Theory." *The American Economic Review* 66(2):274–279.

Wason, P.C. and Diana Shapiro. 1971. "Natural and Contrived Experience in a Reasoning Problem." *Quarterly Journal of Experimental Psychology* 23:63–71.

Woon, Jonathan. 2010. "Democratic Accountability and Retrospective Voting: A Laboratory Experiment." Working paper: University of Pittsburgh.

Figure 1. Flood and Dresher's experimental test of Nash equilibrium

		Column (JW)		
		Left	Right	
Row (AA)	Up	-1, 2 (8%)	0.5, 1 (60%)	(68%)
	Down	0, 0.5 (14%)	1, -1 (18%)	(32%)
		(22%)	(78%)	

Source: Flood (1958), Notes: Payoffs (in pennies) are given by the pair of numbers in the top line of each cell. The observed frequencies (out of 100 plays) are shown in parentheses. The original action labels used in Flood (1958) are "Row 1," "Row 2," "Column 1," and "Column 2."

Table 1. Comparison of alternative voting rules

State (ω)	Policy choice (p)	Retrospective rule (Outcome-based)	Delegate equilibrium rule (Policy-based)
A	A	Incumbent	Incumbent
A	B	Challenger	Challenger
B	*A*	*Challenger*	*Incumbent*
B	*B*	*Incumbent*	*Challenger*

Table 2. Voter behavior

State (ω)	Policy choice (p)	Votes for incumbents		N
		Predicted	Observed	
A	A	100%	95%	603
A	B	100%	35%	211
B	A	0%	16%	352
B	B	0%	87%	418

Table 3. Politician behavior

Type	Signal	Policy *A* chosen		N
		Predicted	Observed	
Pragmatic policy-motivated	A	100%	95%	17
	B	0%	17%	36
Ideological policy-motivated	A	0%	21%	140
	B	0%	9%	140
Office-motivated (all)	A	100%	91%	648
	B	100%	20%	564