The Effect of Repeated Play on Reputation Building: An Experimental Approach

Dustin H. Tingley and Barbara F. Walter

Abstract What effect does repeated play have on reputation building? The literature on international relations remains divided on whether, when, and how reputation matters in both interstate and intrastate conflict. We examine reputation building through a series of incentivized laboratory experiments. Using comparative statics from a repeated entry-deterrence game, we isolate how incentives for reputation building should change as the number of entrants changes. We find that subjects in our experiments generally build reputations and that those investments pay off, but we also find that some subjects did not react to incentives to build reputation in ways our model had predicted. In order to explain this, we focus on the heterogeneity of preferences and cognitive abilities that may exist in any population. Our research suggests that rational-choice scholars of international relations and those using more psychologically based explanations have more common ground than previously articulated.

Reputation building is one of the most talked-about, yet least understood strategic phenomenon in international relations. Over the past sixty years, scholars and policymakers have debated the degree to which governments invest in reputations for toughness, and if they do, whether these reputations deter aggressive behavior. Anecdotal evidence suggests that state leaders care deeply about reputations. But empirical studies have found, at best, only mixed support that such reputations matter. Sartori found that governments that invested in reputations for honesty were able to negotiate more effectively with foreign leaders.[1] Tomz found that governments that invested in reputations for honesty were able to obtain better loans.[2] Walter found that governments that built reputations for toughness could deter future separatists.[3] But Snyder and Diesing, Mercer, and Press all found that

The authors wish to thank Zoltan Hajnal, Alice Hsiaw, Bob Keohane, Tom Palfrey, and Kris Ramsay for their comments and encouragement, as well as Ernesto Reuben for his assistance with coding in Ztree.

1. Sartori 2005.
2. Tomz 2007.
3. Walter 2009.

International Organization 65, Spring 2011, pp. 343–65

doi:10.1017/S0020818311000026

reputations did not appear to have any influence on how state leaders behaved even if leaders believed they would.[4]

These disparate findings are puzzling. If reputations work in some instances but not others, what explains this variation? We believe laboratory experiments can help answer this question. By carefully controlling the environment in which decisions are made, laboratory experiments can reveal when individuals choose to invest in reputation, when reputations are likely to affect behavior, and why investments might matter sometimes more than others.

The article that follows is broken into three parts. In the first part, we walk the reader through the standard sequential equilibrium model commonly used in economics to explain when and why reputations should matter. According to the model, under conditions of incomplete information, individuals should invest more heavily in reputation building (and draw more inferences from past behavior), if they believe a game will be repeated many rather than few times. Conversely, they should care less about reputation in situations that are repeated less often. As Milgrom and Roberts have argued, "the value of reputation and the extent of reputation building (will) increase with the frequency and opportunities for its use."[5]

The second part of this article tests these predictions in the controlled, incentivized environment of the laboratory.[6] We believe experiments offer three advantages in studying reputation building.[7] First, experiments allow the researcher to obtain unambiguous evidence about causation. Currently, the main criticism of reputation arguments is not that leaders do not invest in reputation building—most scholars and policymakers agree that they do—but that these investments do not appear to work. Laboratory experiments allow us to determine whether individuals react to incentives to build reputations in ways predicted by the model, and whether reputations, once built, affect behavior at all. Second, lab experiments allow the researcher to manipulate the decision environment in specific ways to isolate the importance of specific factors. Instead of relying on observational data from historical cases (which can be highly unreliable), lab experiments allow us to alter the values of certain variables while carefully controlling other parts of the decision context.[8] Third, experiments also reveal how individuals may deviate

4. See Snyder and Diesing 1977; Mercer 1996; and Press 2005.

5. Milgrom and Roberts 1982, 304.

6. See Bolton and Ockenfels 2007; Brandts and Figueras 2003; Camerer and Weigelt 1988; Grosskopf and Sarin forthcoming; Jung, Kagel, and Levin 1994; and Neral and Ochs 1992.

7. See Kinder and Palfrey 1993; and McDermott 2002.

8. Testing the effects of repeated play using observational data is difficult. This is because it is impossible to find a situation where all the elements of a game are exactly the same except for the number of iterations. One could try to study a country (for example, China) that faced four potential outside aggressors over territory and compare it to a country (for example, India) that faced eight potential outside aggressors over territory, but there would be so many potentially confounding factors that it would be impossible to determine whether the repeated nature of play was driving differences in behavior and not some other factor.

from purely rational behavior, opening a window for additional theorizing about when and why reputation might matter.[9]

Our tests reveal four surprising findings. First, individuals invested in reputation building even when a game was repeated relatively few times. In the simple entry-deterrence game we use, defenders who played against four entrants were just as likely to invest in reputation building as defenders who played against twice as many entrants, at least early on. Second, building a reputation for toughness helped defenders deter entry. Subjects who had observed a defender backing down in the past were significantly more likely to challenge than those who had never observed such backing down. This provides clear evidence that building a reputation for toughness can be valuable at least in the controlled environment of the laboratory. Third, there were clearly times when individuals invested more and less heavily in reputation building, and when reputation building was more or less likely to work. Defenders underinvested in reputation building against early entrants, and overinvested in reputation building against later entrants. This suggests that if one looked only at early iterations of a particular game, the incidence and effects of reputation building would be much weaker than if one looked only at later iterations. This may explain some of the discrepancy that has been observed in existing empirical studies. Finally, individuals invested more heavily in reputation building as they gained more experience with the game. However, even after gaining experience with the game, subjects continued to deviate somewhat from the expectations of the model. As will be discussed in more detail below, this off-the-equilibrium path behavior appears to be driven by individuals with different preferences and information processing abilities from the larger subject pool. The bottom line, however, is that most subjects did invest in reputation building and if they did, these investments paid off.

The third part attempts to explain those subjects who did not make these investments, and why the model failed to predict their behavior. The laboratory experiment suggests two possible explanations. First, some subjects appear to be motivated by different preferences than the model took into account. The model assumed that all subjects would be driven by a desire to maximize profits, but our experiment revealed that some individuals cared less about maximizing profits and this caused them to ignore opportunities to engage in deterrence. Second, the experiment also revealed real limits in cognitive ability amongst some of our subjects. Even after gaining experience with the game, a subset of subjects continued to play poorly, something that the model does not take into account.[10] The results of

9. McKelvey and Palfrey 1995.

10. Experimental tests of models of reputation in the prisoner's dilemma (PD) game (Kreps et al. 1982) by Cooper et al. (1996) suggest similar cognitive limitations. There, reputations for being willing to play tit-for-tat in a repeated PD game appeared to explain relatively little of the subjects' decisions in the game. Cooper et al. suggest that a model combining reputational concerns, altruism, and cognitive limitations could explain more of their data. Andreoni and Miller (1993) manipulate beliefs about reputations for altruism in a finite PD experiment and find slightly more support for the Kreps

the experiment, therefore, may explain why some leaders fail to invest in reputation building even if all the purported conditions conducive to reputation building exist.

These are interesting findings and reveal the conditions under which individuals may or may not focus on reputation when making decisions. Still, readers should be aware that the strength of laboratory experiments lies in their ability to test predictions of the reputation theory and not in their applicability to real-world settings. Showing when undergraduate students build reputations in a lab (or react to someone else's attempt to build a reputation) does not mean that state leaders will do the same in the more complex world of international politics. It does, however, mean that researchers will have a better idea about the underlying conditions that must exist before reputation becomes a rational strategy to pursue, and why reputation building might emerge and be influential in some contexts but not others.[11] Laboratory experiments are simply the first step in beginning to identify when and why a particular type of strategic behavior is likely to be observed.

The Model

The model we employ focuses on one type of reputation—a reputation for resolve—because it has dominated much of the literature on reputation building in international relations. We define a reputation for resolve as a belief by others that a player who fought a challenger in the past will continue to fight challengers in the future, given a sufficiently similar situation. In international relations, reputations for resolve are a way for state leaders to deter future challengers by credibly signaling that they will continue to be tough given similar circumstances.

A standard model for this type of reputation is the market entry-deterrence game in economics. In this game, a monopolist can dissuade smaller firms from entering the marketplace by engaging in predatory pricing. Lowering prices and starting a price war allows the monopolist to signal to other firms that entry will be costly and thus allows the monopolist to gain a reputation for being tough. This type of behavior is deemed rational because the short-term cost the monopolist pays to fight early entrants is offset by the long-term profits it obtains by deterring later ones.

Entry-Deterrence Stage Game

This article uses a repeated version of the entry-deterrence stage game displayed in Figure 1. This is the same game used by Jung, Kagel, and Levin in their labo-

et al. sequential equilibrium predictions. More recent literature has begun to explore experimentally the role of "bad" reputations (Grosskopf and Sarin forthcoming).

11. Determining the theory's relevance to actual disputes requires a different set of tests using different data. These data might be historical, or they may come from natural experiments or survey experiments.

ratory experiment on reputation building.[12] What is different about our experiment is that we vary the number of repetitions.[13] Changing the number of entrants a defender faces allows us to see if repeated play really is a key factor affecting reputation-building dynamics and explains why reputation building emerges in some contexts and not others.

In the game, a defender faces a series of potential entrants who must decide whether to challenge the defender or stay quiet. The defender in turn must decide whether to fight entry or allow the challenger to enter. Figure 1 reveals the structure of a single-shot play of the game as well as the payoffs each of the players knows it will receive for the different outcomes.[14]

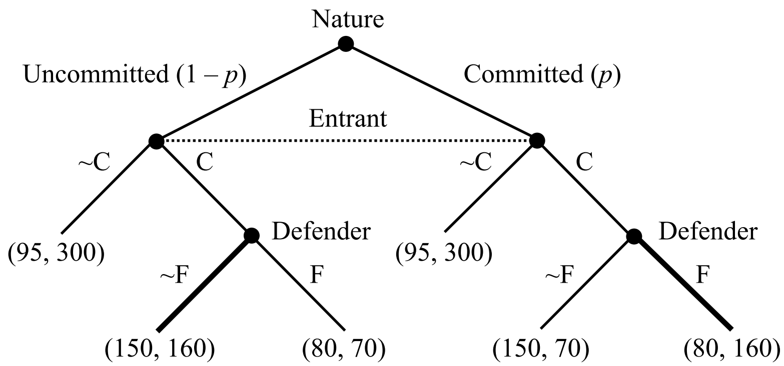


FIGURE 1. *The structure of a single-shot play*

The game begins with nature randomly choosing whether the defender is committed or uncommitted to fighting a challenge with probability p.[15] This introduces the element of uncertainty necessary for reputation building to be a rational strategy to pursue. If the defender is committed, it will always prefer to fight entry rather than acquiesce since this will always deliver better payoffs (see Figure 1). If it is uncommitted, it would prefer to acquiesce rather than pay the costs of fighting.[16] Once nature has chosen the defender's type, the entrant must decide whether

12. Jung, Kagel, and Levin 1994. They found that subjects generally engaged in reputation building but departed from several specific predictions.

13. Similarly see Brandts and Figueras 2003.

14. Payoff parameters are those used by Jung, Kagel, and Levin (1994). We use these parameters to retain consistency in the literature and because they produce clear equilibrium differences depending on the frequency of play.

15. In reality, defender types may change over time. A defender that was once tough may lose some of his resolve if repeated "fights" against entrants have weakened it. Designing an experiment with endogenous types is beyond the objectives of this particular study.

16. In this case, the payoffs are 160 for not fighting a challenger and 70 for fighting a challenger.

to challenge (C) or not (~C), knowing that there is some probability that it is facing a committed defender, and some probability $(1 - p)$ that it is uncommitted. If the entrant decides to challenge, the defender then chooses whether to fight (F) this challenger or not (~F).

In the repeated play version of this game, once the defender makes his or her choice, a second entrant then chooses whether to challenge, after which the defender again decides whether to fight or accommodate. As each entrant plays, they obtain information about how previous entrants played against the defender they are currently matched with, and how the defender played if the previous entrant decided to challenge. Thus, they are able to update their beliefs about what type of defender they are likely to face. How the defender behaves toward an early entrant, therefore, can be interpreted as important information about how the defender is likely to behave toward later entrants. The game continues until the defender has been pitted against a commonly known number of entrants.[17]

Theoretical Predictions

In order to determine whether repeated play had a strong effect on behavior, we had our subjects play two versions of the same game. In one version, there were eight entrants, and in the second version there were four. Assuming risk-neutral utility functions, the standard sequential equilibrium model makes three predictions about how defenders and entrants should play with eight entrants.[18] First, committed defenders should always fight no matter what period they are in. Second, uncommitted defenders should pursue a strategy that depends on how many of the eight entrants still need to choose whether to enter or not. The more entrants that remain, the more valuable deterrence becomes and the more likely uncommitted defenders should be to fight. Uncommitted defenders, therefore, have the incentive to bluff (fight) in early periods to acquire a reputation for toughness and then acquiesce with increasing probability as the number of remaining challengers decreases. Third, entrants should base their decision about entry on information they glean about the type of defender they face and the incentives this defender

17. Future work might consider models where there is no fixed number of entrants but instead a fixed probability of future bargaining. For experimental work along these lines, see Tingley forthcoming.

18. We use the sequential equilibrium concept with the standard belief restriction because it produces a unique equilibrium characterization, has been used by others studying the type of game we employ, and explicitly incorporates the dynamics of updating across periods of play—a feature that is central to the formation or dissolution of reputation that we are interested in. Our predictions use the proof by Jung, Kagel, and Levin (1994), and then applies the model to when there are four or eight entrants. In principle, it is possible to utilize other equilibrium concepts, but this has not been done by theorists for games of the complexity we consider. We attempted to use computational game theoretic techniques for solving for the quantal response equilibrium (using *Gambit*; see McKelvey, McLennan, and Turocy 2007) but were unsuccessful. We note conceptual issues that some have raised generally about sequential equilibrium/subgame perfect concepts in applications to extensive form games (Reny 1992).

has to invest in reputation building. If the defender never backed down, entrants should never enter in the early periods, since in equilibrium both committed and uncommitted defenders will fight in early periods. Entrants should then enter with greater likelihood during the middle and latter periods (knowing that uncommitted defenders will be increasingly likely to back down at these times). If a defender backed down in an earlier period, entrants know they are facing an uncommitted opponent, and they should always enter.

Readers should note, however, that our assumption about risk neutrality could affect both defender and entrant behavior. Camerer and Weigelt discuss this issue in a similar model for either the entrant or defender.[19] They claim that if entrants have risk-averse utility functions they will enter less often and (weak) defenders will back down less often. If defenders have risk-averse utility functions then entrants will enter less often. We include neutral utility functions because doing so substantially simplifies the analysis. We also include them because there is no reason to believe that risk tastes will differ across our experimental treatments and that the relatively small stakes mean that any concavities in the utility function will be moderate.[20] Still, a model allowing for nonrisk neutral utility functions for all players is needed to make more definitive equilibrium predictions, something that is beyond the scope of this research note.

The question we are interested in answering, however, is whether defenders and entrants change their behavior if they face only four entrants. Would the same reputational dynamics emerge if a game were repeated far fewer times?

The model predicts that defender and entrant strategies should change dramatically when the number of repeated plays is reduced. Figure 2 shows the different predicted probabilities of entry and fighting depending on the number of entrants. Note that these are all cases where the entrant has never observed the defender accommodating in the past. These graphs illustrate the fundamental difference in how individuals should behave depending on whether a game is repeated few or many times.

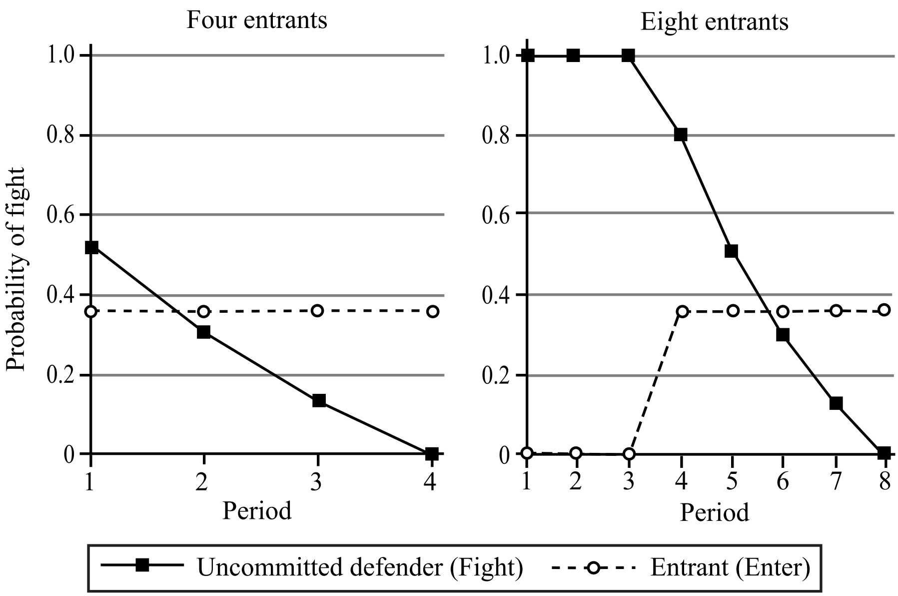
Two key differences are evident between the two versions of the game. First, uncommitted defenders facing only four entrants should be far more likely to acquiesce against the first three entrants than uncommitted defenders facing eight entrants. This is because defenders facing a total of four entrants have fewer incentives to invest in reputation building given the smaller number of entrants they hope to deter.

Second, in response to these differing incentives, entrants should also alter their strategy depending on the total number of entrants. Entrants in the four-person game should be more likely to enter early on. In the four-entrant case, the entrants know they are more likely to gain concessions in the very first period and have, therefore, greater incentives to challenge. This gives us two basic hypotheses for testing:

19. Camerer and Weigelt 1988, 11.
20. This said, there are difficulties with both of these reasonings (Rabin 2000) as well as with procedures designed to control for risk tastes (see Cox and Oaxaca 1995; and Davis and Holt 1993, 476) and our results should be interpreted with the appropriate caveats.

H1: Defenders should be less likely to build reputations (that is, fight entry) in the early stages of a four-person game than in an eight-person game.

H2: Entrants should be more likely to enter in the early stages of the four-person game than in the eight-person game.



Notes: Committed defenders should always fight. Uncommitted defender choices conditional on no previous backdown and risk-neutral utility functions.

FIGURE 2. *Equilibrium predictions*

The Experiment

In this section we describe our experimental design for the entry-deterrence game. A more detailed explanation of both the eight- and four-entrant design is available in a supplemental appendix online. Subjects were recruited through Princeton University's Laboratory for Experimental Social Science (PLESS). A single experimental session used only the eight-entrant or four-entrant treatment and subjects could not participate in both treatments.[21] In each session, subjects were assigned randomly to two separate groups, entrants and defendants, referred to simply as

21. All subjects were told prior to commencing whether they were in the four- or the eight-entrant game.

first movers and second movers. These neutral terms were used to avoid leading the subjects in any way. Defenders were also assigned a "type," either uncommitted or committed, which we called "type 1" or "type 2." Defenders knew their type, but entrants did not. Entrants were told only that each defender had a one-third chance of being committed and two-thirds chance of being uncommitted. This uncertainty, together with the repeated nature of play, made reputation building a valuable strategy to pursue.

During the experiment, entrants were given information on how the defender played against all other previous entrants. If a previous entrant had challenged the defender, all subsequent entrants would see whether the defender had backed down or had fought. This allowed entrants to update their beliefs about the type of defender they faced. If an entrant chose not to challenge, no information about the defender's choice would be recorded and no information about the defender's type would have been revealed.

The experiment proceeded as follows. Defenders faced the entrants sequentially. Within each pairing, entrants were asked to choose between entering the game (and thus challenging the defender), or not entering. We elicited defender choices using the strategy method: defenders were asked to select a strategy based on what an entrant might do: "if the first mover enters I will choose B1 or B2" (not fight or fight). The big advantage of the strategy method is that it allows the researcher to observe the decision of a defender even when their opponent did not choose to enter. The drawback is that subjects now play a normal form game that is subtly different from the sequential form entry deterrence game, and this could bias their behavior.[22]

Each entrant then made one decision with no available history (in the first period), one decision with a previous period's history against a different defender (in the second period), and so on.[23] At the end of each repetition (after each entrant had played each defender once), subjects saw a screen with their decision history, the decisions of the subject they were paired with in each period, and their own pay-

22. We suspect this design feature would likely have the most influence on entrant choices, as they might think that defenders would make different choices depending on whether the entrant choice was known or hypothetical. Experimentalists with strong reasons to use such a strategy solicitation procedure frequently face this problem and there is considerable debate on the exact consequences of this (see Brosig, Weimann, and Yang 2003; Brandts and Charness 2000; and Oxoby and McLeish 2004). We tried to minimize any bias by designing an instructional procedure that made clear the sequential nature of the game being played. Despite using the strategy method, the behavior we observed in our experiment is very similar to that observed by Bolton and Ockenfels (2007) who elicited strategies sequentially. In an additional follow-up experiment with a one-shot version of the game, we had subjects play the game sequentially or using the strategy method. We found virtually no differences in behavior. Nevertheless, future experiments on reputations that compare completely sequential play and the strategy solicitation procedure are desirable, especially as the normal form version of the single-shot game has multiple equilibria.

23. This design allowed us to keep all subjects engaged throughout the experiment, as well as to maximize the amount of data we could collect within an experimental session. The engagement consideration is key in light of recent critiques of previous experiments studying reputation (Grosskopf and Sarin forthcoming).

offs. Subjects knew that these payoffs would be translated into U.S. dollars at the end of the experiment. Subjects then repeated the experiment four times in order to take into account the effects of learning and to generate sufficient data for the analysis.[24]

Results and Interpretation

The Effects of Repeated Play

Our goal in running the experiment was to collect data on how subjects reacted to a situation where varying incentives existed to invest in reputation.[25] We did this to answer two questions. First, would defenders be significantly less likely to invest in reputation building and entrants more likely to challenge when the number of repetitions was low (H1 and H2)? Second, more generally, would subjects deviate in any way from the expectations of our model? The results, as mentioned above, were striking.[26]

Result 1: Defenders were not less likely to invest in reputation and entrants were not more likely to challenge in games with fewer repetitions.

Contrary to the predictions of the model, defenders in the four- and eight-entrant games played surprisingly similarly. Figure 3 pools across all repetitions of the experiment and shows that uncommitted defenders in both treatments invested in reputation building half-heartedly against the very first entrant, then invested more heavily against the middle entrants, and then acquiesced against the very last entrant. Perhaps the biggest surprise was that in the first period, defenders in the eight-

24. The precise number of repetitions was unknown to subjects; they were simply told that the experiment "may or may not be repeated." Across repetitions of the experiment all positions (entrant/defender) stayed the same, entrants were randomly assigned when they would move against each defender, and defender types (committed/uncommitted) were randomly reassigned according to the commonly known type distribution. Points were converted to dollars at the rate of $2.00 per 4000 points and subjects were paid on their total points from the experiment. Total subject earnings ranged from $18–$30 for slightly under an hour of time in the laboratory.

25. Our empirical strategy is to break defenders out by those who had already backed down and those that had not. We also break out entrants into those that face a defender who had not yet backed down, and those that faced a defender who had. We do this because the equilibrium model makes this important distinction.

26. We ran a series of follow-up experiments that had no repeat play and hence were single-shot games like that depicted in Figure 1. The experiments were otherwise identical to the experiments below and used different subjects. In this setting a single-shot equilibrium prediction is that uncommitted defenders should never fight. Only 8 percent of uncommitted defender choices were to fight entry, which is in stark contrast to the results below. This suggests that it is unlikely that subjects simply had raw preferences for fighting and instead that repetition is an important structural parameter. Our results for the eight-entrant case were largely similar to what Jung, Kagel, and Levin (1994) found in their experiment. Below we examine individual level differences in behavior that they did not do.

person design were actually significantly more likely to acquiesce than those in the game with fewer repetitions.[27] The model predicted just the opposite.

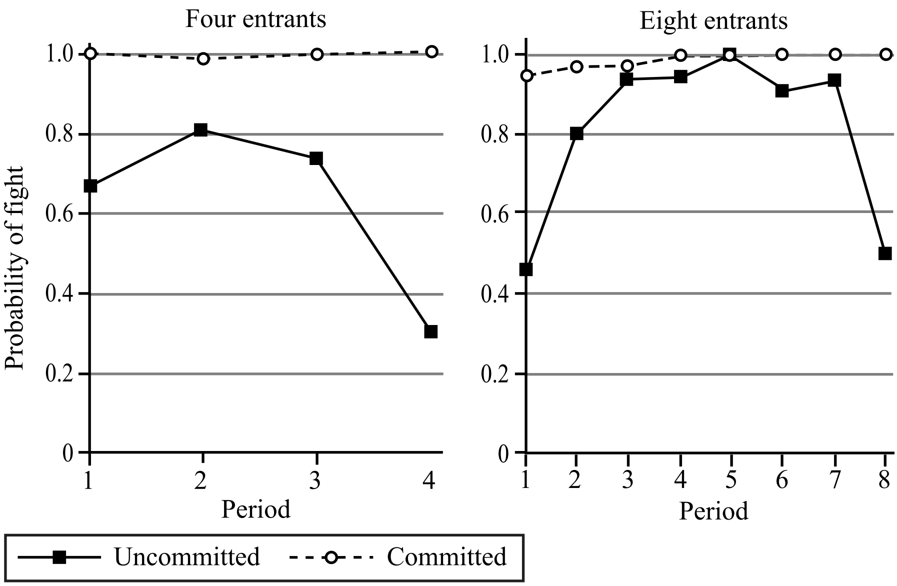


FIGURE 3. *Fight: Conditional on no previous breakdown*

We now shift our attention to entrants. Figure 4 reveals that entrants also did not change their behavior dramatically if they were playing shorter or longer games. The very first entrants in both the four- and eight-person game entered at very high rates, after which the rate of entry declined until it stabilized with the third entrant. This is striking. According to the model, the first three entrants in the eight-entrant design simply should not enter, which is conditional on the defender having never backed down. If early entrants enter and defenders fight them (as they are supposed to), entrants will receive fewer points than if they had chosen not to enter at all. Yet the vast majority of entrants in the eight-person design did what most subjects in the four-person design did: they chose to enter in the very first period. As Figure 4 shows, a small difference between the four- and eight-person games did emerge in the second and third periods but it was much smaller

27. The difference in mean rate of fighting was 24 percent ($p < .05$, standard errors clustered at individual level). This result in the eight-entrant case is similar to that found by Jung, Kagel, and Levin (1994, 80) for inexperienced players.

than the model predicted.[28] This suggests that a change from eight to four periods was not the key factor determining when reputation building would matter and when it would not.

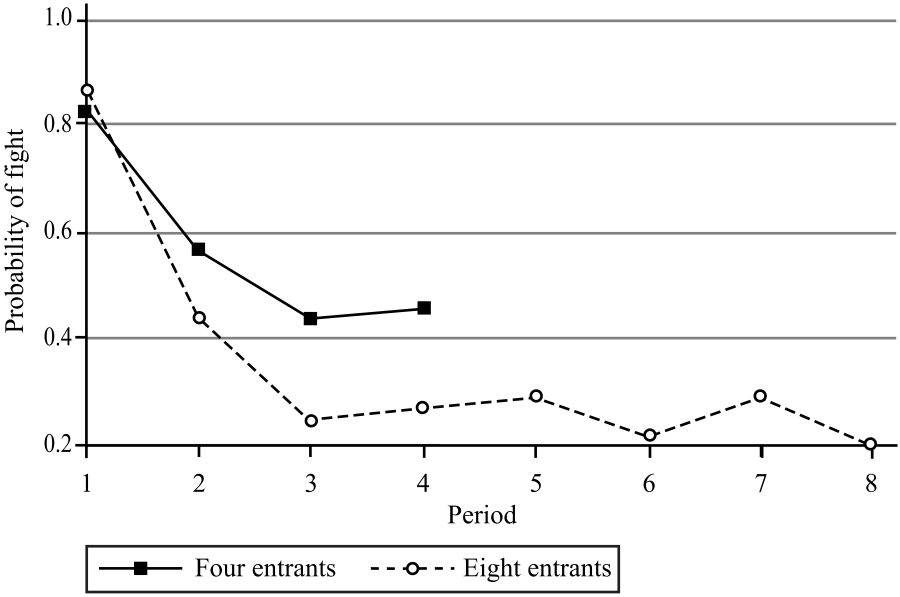


FIGURE 4. *Enter: Conditional on no previous breakdown*

Result 2: In the early phases of each game, defenders underinvested in reputation, while entrants challenged more regularly than they should have.

What appeared to matter more was the sequencing of play. Uncommitted defenders were much less likely to invest in reputation building against early entrants, and early entrants were much more likely to challenge than was expected. This was true in both the four- and eight-person game. The equilibrium model predicted that uncommitted defenders in the four-person design would fight with a mixed strategy in all periods, while uncommitted types in the eight-entrant design would fight against the first three entrants and then play a mixed strategy against the remaining entrants. This did not occur.

28. Entry frequencies in the first period between the two treatments are not statistically distinguishable, entry was higher in the four-entrant case in the second ($p < .1$) and third ($p < .05$) periods.

As the right side of Figure 3 shows, this was not the case, especially in the first period. The majority of uncommitted defenders in both the four- and eight-person design chose to acquiesce against the first entrant rather than fight. Moreover, while the probability of fighting did increase in the second and third periods in the eight-person game, that probability never approached the 100 percent fight rate predicted by the model. For reasons to be discussed below, uncommitted defenders invested less in reputation building in the very first round than was predicted.

Entrants also deviated from equilibrium predictions in these early periods. As Figure 4 shows, early entrants in games with four and eight entrants entered at a much higher rate than expected. According to the model, approximately 36 percent of entrants in the four-entrant design should have entered in each period where there was no observed accommodation. Yet, as Figure 4 shows, more than 80 percent of subjects in the four-entrant game entered in the very first period. This percentage dropped in the second and third periods but never reached the low rate predicted by the model.

The same is true in the eight-entrant design. According to the theory, entrants in the eight-entrant design should never enter in the first, second, and third periods. After that, approximately 36 percent of them should enter if no previous accommodation is observed. But Figure 4 shows that entrants again entered at a high rate in the first period, reducing their entry in the second and third periods, and eventually entering close to equilibrium predictions.

The fact that early entrants in both the four- and eight-period designs behaved this way does not mean that reputation building did not occur. One of the striking findings revealed by Figure 4 is that even though subjects were not deterred in the first two periods of play, they did tend to be deterred in the third, fourth, and remaining periods when paired with a defender who had not backed down. This is especially surprising in the four-period design where entrants were not expected to be deterred. Reputation building and deterrence did emerge, it just emerged later than expected.

Part of the reason reputation building emerged later was due to subjects' inexperience with the game. As will be discussed below, uncommitted defenders increased the rate at which they fought early entrants as they learned how to play the game. In later rounds, defenders fought more, though interestingly, this did not have a substantial effect on early entrant behavior which on average showed little learning.[29]

Result 3: In the latter phases of each game, weak defenders over-invested in reputation.

Finally, uncommitted defenders deviated from equilibrium predictions in the final stages of both the four- and eight-person designs. In both cases, uncommitted

29. One potential explanation for this is that entrants felt defenders had an unfair advantage in the game, and hence by entering earlier this advantage would be lessened for some set of defenders. We discuss this further below.

defenders should have increasingly backed down as the number of remaining entrants declined. With fewer remaining entrants, these defenders had fewer incentives to build a reputation for toughness. Yet Figure 3 reveals that defenders continued to invest in reputation building until the second to last period of play—far longer than was rational given the payoff structure. Jung, Kagel, and Levin found some evidence for this overinvestment in their eight-entrant design where 50 percent of the inexperienced weak defenders who had not yet backed down fought entry in the penultimate period.[30] Our results are even more extreme.

The laboratory experiment, therefore, left us with three surprising findings. The first was that uncommitted defenders were willing to invest in reputation building whether they faced few or many entrants. Repeated play—at least a change between four and eight iterations—did not have the large effect on behavior that most scholars have assumed. The second is that uncommitted defenders underinvested in reputation building against the very earliest challengers, while early entrants challenged at a higher rate. The third is that uncommitted defenders who invested in reputation building, invested beyond the point where it was optimal.

Discussion

Why did defenders and entrants behave this way? One explanation for these unexpected results is that subjects did not understand how to play the game. Although we gave careful instructions and allowed subjects to ask questions before the experiment began, it is possible that some subjects still did not grasp the strategic logic underlying the game and played poorly as a result.

To check if this was true, we examined whether subjects learned how to play better over time. Recall that we repeated the experiment four times (each defender faced a sequence of four or eight entrants a total of four times) in order to check whether they would play closer to equilibrium predictions as they gained more experience with the game. We found that some learning did take place, but that subjects never played exactly the way the model predicted.

Figure 5 reveals that in both the four- and eight-entrant design, uncommitted defenders increased their investment in reputation building in early periods as they gained more experience with the game.[31] For the eight-entrant treatment, this is exactly what the theory predicted. Uncommitted defenders in the eight-player games learned that fighting in early periods paid off. As one subject wrote in the post-experiment questionnaire: “If you are weak, choose to fight in the first period. Subsequent first movers will think you’re strong and will choose not to challenge, so you will earn 300 points instead of 160. It took a while to figure out.” Still, even after four repetitions, uncommitted defenders who faced eight entrants con-

30. Jung, Kagel, and Levin 1994, 56.
31. Analogous analysis of entry rates showed little change through repetitions of the experiment.

tinued to back down about half the time against the very first entrant, far more than the model predicted.

Figure 5 also reveals that less learning occurred in the final periods of play. Regardless of how many times the game was played, defenders in both the four- and eight-person game continued to invest in reputation building long after the model predicted they should. As will be discussed in greater detail below, we believe that learning was less likely to take place in these later periods because mistakes were less costly at this point in the game and harder to catch as a result.

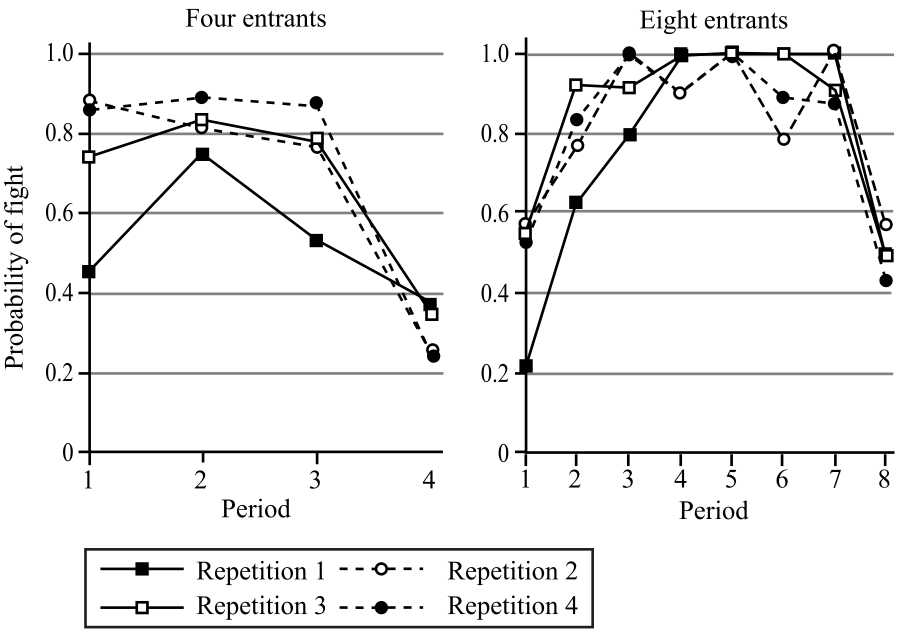
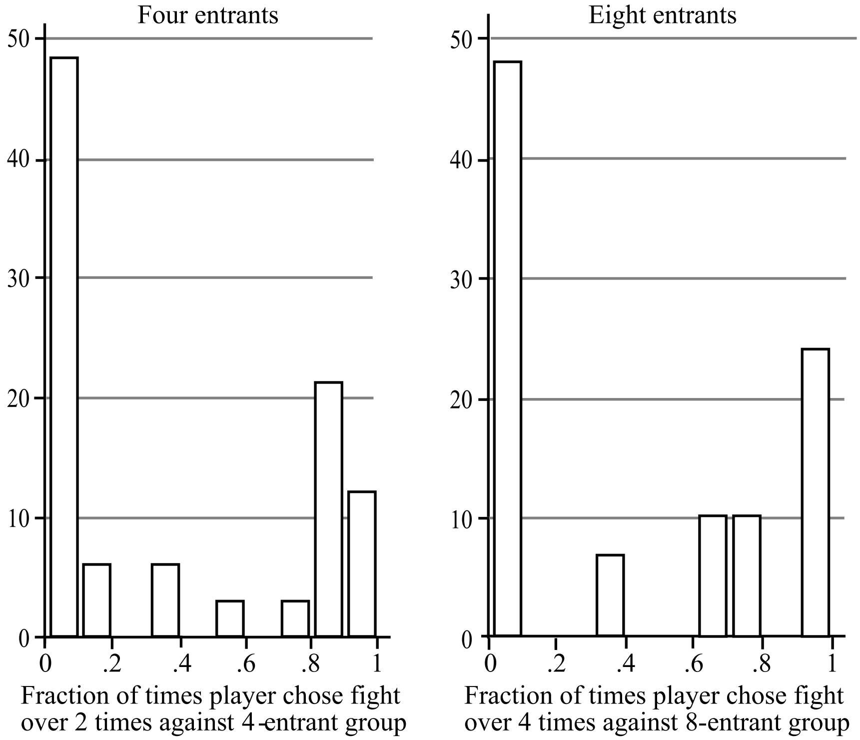


FIGURE 5. *Uncommited type learning: Conditional on no previous breakdown*

Still, the fact that defenders continued to under- and overinvest in reputations even as they learned how to play the game suggests that something else is going on. A second possible explanation is that this off-the-equilibrium path behavior is being driven by a subset of subjects with different preferences or information-processing abilities than the model takes into account. In an attempt to determine if this is true, we looked more closely at two different types of data. First, we examined how different subjects played the role of uncommitted defender. Since defenders repeated the same sequence of decisions four or eight times, we looked at the percentage of times each subject chose to fight in more than half the periods. Figure 6 shows that many subjects in the uncommitted defender's role fell into one of two groups: they either regularly backed down or they regularly fought. Thus, a subset of individuals acquiesced in every period and never changed its behavior, and it was these uncommitted defenders who appeared to be driving the anomalous results in the early periods. Additional analysis revealed that uncommitted defenders who chose to back down early in the game were also more likely to continue to back down. This suggests that a certain type of individual—one who always acquiesces when in the uncommitted role—helps explain the underinvestment in reputation in the early rounds of play.



Note: Uncommitted defenders types only.

FIGURE 6. *Frequency of fight choices by individual*

As a second check, we reviewed what subjects wrote in a postexperiment questionnaire that asked them to describe the logic behind their decisions. Their responses were illuminating. Uncommitted defenders who backed down early in the eight-person design offered a fairly consistent explanation for their behavior: they backed down because they believed it would deliver the highest reward.

Acquiescing early did not deliver the highest reward, although it is easy to see why some defenders might have thought this. If a defender focused only on the immediate round of play and not on future rounds, acquiescing offered the greatest number of immediate points. Uncommitted defenders who backed down in any given period received 160 points versus 70 points if they chose to fight. If, however, a defender took a longer-term view of payoffs, then backing down could be more costly if it encouraged more entrants to challenge in the future. It appeared that some defenders were choosing to acquiesce because they focused only on immediate gains and not on the cumulative value of deterring future entrants. This logic was revealed by a number of subjects in their responses to our exit survey:

- "Pick whatever option gives you 160 every time. It will give you the most points."
- "If you are a weak type, always acquiesce. If you are strong, always fight. It guarantees the most points regardless of what the first mover does. Play the game the same way throughout."
- "Choose the option that gives you more points. If you are a weak type, acquiesce every move. Even if the first mover chooses not to challenge you can't lose. It will get you more money."

What then explains why some uncommitted defenders consistently overinvested in reputation building late in each game? We believe there are three potential explanations. It is possible that subjects invested in reputation building longer than was rational because they received little feedback that this was a bad strategy to pursue. Weak defenders in the eight-entrant design who backed down early paid a high price for this mistake since they faced a higher rate of entry from the remaining challengers. Uncommitted defenders who overinvested late in the game, however, paid a less hefty price, making the mistake harder to detect. The fact that behavior changed very little in the last stages of the game, as shown in Figure 5, suggests that less learning did occur, perhaps for this reason.

A second related explanation, however, has to do with cognitive abilities. It is possible that some subjects never completely understood the game and fought longer than necessary, thinking this was the optimal strategy to pursue. This lack of strategic sophistication was revealed in the exit questionnaire by a number of subjects who offered misguided advice to future players:

- "Always choose to fight until the last move, then pick whatever gives you the best return on the last play."
- "Always choose to fight. If you can convince every first mover that you'll fight no matter what they do, then if they're rational they should choose not to challenge."
- "Always fight. You get the most money considering the fact that the first mover will most likely choose to challenge."

A third explanation has to do with differences in preferences. It is possible, for example, that some uncommitted defenders continued to invest in reputation building simply because they received personal satisfaction from being known as strong, or at least "not weak." Some uncommitted types may resist revealing weakness to any player, only acquiescing when their behavior cannot be observed by any additional challengers. According to one subject in the four-entrant treatment: "If you're a weak type, pick fight for the first, second, and third periods. Acquiesce on the last period. At the fourth period no one's going to see the results of your action."

We now have some idea why defenders behaved the way they did, but what about entrants? Recall that entrants were more likely to fight in early rounds and more likely to be deterred in later rounds than was predicted by the model. On the surface the answer seems obvious. Entrants entered at a higher rate in early periods because defenders were more likely to back down in these periods, and they were deterred at a higher rate in late periods because defenders were more likely to fight. In other words, entrants pursued a strategy designed to deliver the best possible payoffs, given how defenders were behaving. But how did entrants know that defenders were going to behave this way, especially since it was not in the defender's interest to do this?

There are at least two reasons entrants may have believed defenders would play this way. For one, they may have already understood that a subset of individuals existed who would fail to play correctly; some would acquiesce too early, some would fight too late. If this were the case, then entrants (or at least particularly strategic entrants) could adjust their strategy to match this. Second, entrants may have also anticipated that some individuals would consciously choose not to bluff, while others would want to appear strong, and that these individuals would change the strategic game in early and late periods.[32] The results from the experiment reveal that most subjects appeared to be aware that not all defenders would behave rationally all the time and they altered their play accordingly.

Conclusion

This article attempted to determine when individuals would invest in reputation building, when it would have an effect, and whether the number of repeated plays

32. The high rates of entry in early periods might also reflect social comparison issues. If an entrant does not challenge, they get 95 points whereas the defender gets 300 points. If the entrant enters, the point "spread" is much smaller whether or not the defender fights, or whether it is committed. Thus, entry may serve to equalize earnings for both groups. There was some evidence of this in our post experiment polls, as some entrants expressed frustration that the experiment favored defenders. Fairness norms may explain part of what is going on here (Fehr and Schmidt 1999). Similarly, if entrants were altruistic toward other entrants and believed that entry in period 1 could push a weak defender to reveal their type, then early entrants may want to enter at a higher rate in order to provide this information to later entrants. See Bolton and Ockenfels 2007, for an experiment motivated by information externalities and reputation building.

was a critical variable affecting behavior. In order to do this, we introduced a methodology relatively unknown in the international relations literature: incentivized laboratory experiments.[33] This allowed us to isolate variables that game theoretic models suggest should affect behavior, and to test whether they had the predicted effect.

The sequential equilibrium model has always assumed that reputation building would become more valuable as the number of repetitions increased. It predicted that uncommitted defenders would invest most heavily in reputations for toughness against early challengers, and then taper off as the number of remaining challengers declined. Our experiment, however, revealed that subjects invested in reputation building even when they expected only a small number of interactions. It also revealed that a certain minority of subjects invested in reputation much less than expected, and that some subjects invested much more. Thus, while our experiment revealed that most subjects did react to incentives in ways consistent with our model—they invested in a reputation for toughness and these reputations deterred entry—it also revealed that a certain group of individuals did not.

These findings offer three insights into when reputation building is likely to emerge and why it may matter sometimes but not others. First, it demonstrates that investments in reputation are likely to be made even in games that are repeated relatively few times. Uncommitted defenders will invest in reputation, and entrants will be influenced by reputation, even in games with only four iterations. Second, it reveals that a certain subset of individuals will choose never to invest in reputation building even if it is to their advantage, while others will almost always invest even if it has decreasing value to them. Third, most subjects (at least most subjects in our laboratory) are quite strategic and appear to pursue strategies that take into account these subsets of individuals when determining how to play.

This study offers insights into at least two issues related to international relations. First, it offers strong empirical evidence that reputation building works. Studies by Mercer, Press, and by Snyder and Diesing have questioned the value of investing in a reputation for resolve; leaders who backed down in one crisis did not appear to suffer reputational costs for doing so.[34] Our experiment suggests that backing down can have very negative effects if a game is expected to be repeated, and past behavior can be clearly observed. Indeed, backing down against the first entrant led to substantially lower earnings in subsequent periods. Figure 7 plots the average earnings for weak defenders in each period, broken out by whether entry was fought or not fought in the first period. Subjects who built reputations for toughness early in a game earned less in the first period but more in subsequent periods in both the four- and eight-entrant designs.

33. We are not the first to use laboratory experiments in international relations. The first experimental work in international relations was conducted by Mort Deutsch and Marc Pilisuk in the early 1960s. Currently, important work in this vein is being done by a number of scholars including Alex Mintz, Nehemia Geva, Francis Beer, Jonathan Wilkenfeld, Mark Shafer, Rose McDermott, Daniel Druckman, and Phil Tetlock.

34. Mercer 1996; Press 2005; Snyder and Diesing 1977.

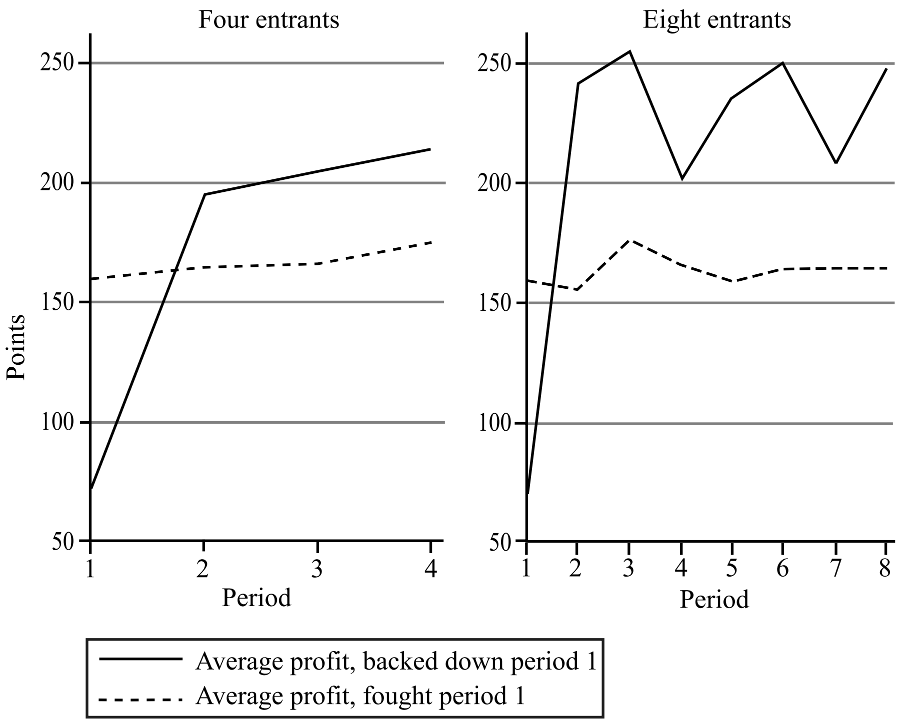


FIGURE 7. *Differences in profit by fighting or backing down in first period*

Figure 7 reveals that subjects clearly took into account how a defender behaved in the past and factored this into their decision about whether to challenge. If a defender had backed down, most entrants interpreted this to be a sign of weakness and most entrants chose to challenge more as a result. In this experiment, past behavior did matter, and investments in reputation building appeared to work.

This study also offers insight into why state leaders might deviate from purely rational reputation-building behavior. As noted, it is a long way from undergraduate subjects and experimental manipulations to state leaders and the battles they wage around the globe. Thus, considerable caution needs to be taken when trying to make predictions about real world behavior from this kind of analysis. Nevertheless, it is possible that the differences in preferences and cognitive capabilities suggested by our experiment also exist among state leaders. Some leaders may take longer to process how a particular strategic game is played, and will, therefore, fail to invest in reputation building at least against early opponents. Indeed, as shown in Figure 6, nearly 50 percent of our defender subjects never tried to build a reputation. Others may be more principled or honest and will refuse to try to build a reputation for toughness even if this would serve to deter more challengers.

Skeptics might argue that the results of the experiment are unlikely to apply to state leaders. Had these individuals been cognitively challenged or highly principled they would never have ascended the political ladder and gained top leadership positions. This may be true, although two arguments are worth considering. Princeton undergraduates must also navigate a highly competitive path to gain admittance into an elite Ivy league school. A strong selection effect, therefore, also occurred with our subjects. Second, anyone who has interacted with state leaders knows that real differences exist in terms of their cognitive ability and personal preferences. We suspect that some state leaders, like some Princeton undergraduates, will not play the game as well or as efficiently as others, and that patterns of reputation building will be affected as a result.

All of this suggests that new models need to be developed that better reflect the different preferences and abilities that exist in the human population and the way in which this affects optimal strategies to pursue. Bueno de Mesquita and McDermott argue that the role of emotion plays a more central role than standard models permit.[35] Our experiment revealed heterogeneity amongst defenders in terms of how likely they were to defend against entry, especially in terms of our subjects' ability to understand the strategic incentives of the game. Individual heterogeneity might be modeled formally by allowing some actors to be more strategically sophisticated than others. Cognitive hierarchy equilibrium (CHE) models assume a distribution of strategic sophistication and assume that best responses are based on the fact that not all opponents are equally talented.[36] Such models have only been worked out for much simpler games than the one we consider here, but they do suggest a way to link behavioral irregularities to rational choice models in ways that could be productive. We believe that an active interplay between experimental, theoretical, and observational work in international relations, therefore, would help scholars explain many of the puzzling patterns of behavior we see in the world today.[37]

References

Andreoni, James, and John H. Miller. 1993. Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence. *Economic Journal* 103 (418):570–85.

Bolton, Gary E., and Axel Ockenfels. 2007. Information Externalities, Matching and Reputation Building—A Comment on Theory and an Experiment. Working Paper Series in Economics, 17. Cologne, Germany: University of Cologne. Available at ⟨http://ockenfels.uni-koeln.de/download/papers/bolton_ockenfels_information_externalities.pdf⟩. Accessed 4 January 2011.

Brandts, Jordi, and Gary Charness. 2000. Hot vs. Cold: Sequential Responses and Preference Stability in Experimental Games. *Experimental Economics* 2 (3):227–38.

35. Bueno de Mesquita and McDermott 2004.
36. Camerer, Ho, and Chong 2004.
37. See Mercer 2005; and Schotter 2006.

Brandts, Jordi, and Neus Figueras. 2003. An Exploration of Reputation Formation in Experimental Games. *Journal of Economic Behavior and Organization* 50 (1):89–115.

Brosig, Jeannette, Joachim Weimann, and Chun-Lei Yang. 2003. The Hot Versus Cold Effect in a Simple Bargaining Experiment. *Experimental Economics* 6 (1):75–90.

Bueno de Mesquita, Bruce, and Rose McDermott. 2004. Crossing No Man's Land: Cooperation from the Trenches. *Political Psychology* 25 (2):271–87.

Camerer, Colin, Teck-Hua Ho, and Juin-Kuan Chong. 2004. A Cognitive Hierarchy Model of Games. *Quarterly Journal of Economics* 119 (3):861–98.

Camerer, Colin, and Keith Weigelt. 1988. Experimental Tests of a Sequential Equilibrium Reputation Model. *Econometrica* 56 (1):1–36.

Cooper, Russell, Douglas V. DeJong, Robert Forsythe, and Thomas W. Ross. 1996. Cooperation Without Reputation: Experimental Evidence from Prisoner's Dilemma Games. *Games and Economic Behavior* 12 (2):187–218.

Cox, James C., and Ronald L. Oaxaca. 1995. Inducing Risk-Neutral Preferences: Further Analysis of the Data. *Journal of Risk and Uncertainty* 11 (1):65–79.

Davis, Douglas D., and Charles A. Holt. 1993. *Experimental Economics*. Princeton, N.J.: Princeton University Press.

Fehr, Ernst, and Klaus M. Schmidt. 1999. A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* 114 (3):817–68.

Grosskopf, Brit, and Rajiv Sarin. Forthcoming. Is Reputation Good or Bad? An Experiment. *American Economic Review*.

Jung, Yun Joo, John H. Kagel, and Dan Levin. 1994. On the Existence of Predatory Pricing: An Experimental Study of Reputation and Entry Deterrence in the Chain-Store Game. *RAND Journal of Economics* 25 (1):72–93.

Kinder, Donald R., and Thomas R. Palfrey. 1993. *Experimental Foundations of Political Science*. Ann Arbor: University of Michigan Press.

Kreps, David M., Paul Milgrom, John Roberts, and Robert Wilson. 1982. Rational Cooperation in the Finitely Repeated Prisoners' Dilemma. *Journal of Economic Theory* 27 (2):245–52.

McDermott, Rose. 2002. Experimental Methodology in Political Science. *Political Analysis* 10 (4):325–42.

McKelvey, Richard D., Andrew M. McLennan, and Theodore L. Turocy. 2007. Gambit: Software Tools for Game Theory, Version 0.2007.01.30. Available at ⟨http://www.gambit-project.org⟩. Accessed 4 January 2011.

McKelvey, Richard D., and Thomas R. Palfrey. 1995. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior* 10 (1):6–38.

Mercer, Jonathan. 1996. *Reputation and International Politics*. Ithaca, N.Y.: Cornell University Press.

———. 2005. Rationality and Psychology in International Politics. *International Organization* 59 (1):77–106.

Milgrom, Paul, and John Roberts. 1982. Predation, Reputation, and Entry Deterrence. *Journal of Economic Theory* 27 (2):280–312.

Neral, John, and Jack Ochs. 1992. The Sequential Equilibrium Theory of Reputation Building: A Further Test. *Econometrica* 60 (5):151–69.

Oxoby, Robert J., and Kendra N. McLeish. 2004. Sequential Decision and Strategy Vector Methods in Ultimatum Bargaining: Evidence on the Strength of Other-Regarding Behavior. *Economics Letters* 84 (3):399–405.

Press, Daryl G. 2005. *Calculating Credibility: How Leaders Assess Military Threats*. Ithaca, N.Y.: Cornell University Press.

Rabin, Matthew. 2000. Risk Aversion and Expected-Utility Theory: A Calibration Theorem. *Econometrica* 68 (5):1281–92.

Reny, Philip J. 1992. Rationality in Extensive-Form Games. *Journal of Economic Perspectives* 6 (4):103–18.

Sartori, Anne E. 2005. *Deterrence by Diplomacy*. Princeton, N.J.: Princeton University Press.

Schotter, Andrew. 2006. Strong and Wrong: The Use of Rational Choice Theory in Experimental Economics. *Journal of Theoretical Politics* 18 (4):498–511.

Snyder, Glenn. H., and Paul Diesing. 1977. *Conflict Among Nations: Bargaining, Decision Making, and System Structure in International Crises*. Princeton, N.J.: Princeton University Press.

Tingley, Dustin. Forthcoming. The Dark Side of the Future: An Experimental Test of Commitment Problems in Bargaining. *International Studies Quarterly*.

Tomz, Michael. 2007. *Reputation and International Cooperation: Sovereign Debt Across Three Centuries*. Princeton, N.J.: Princeton University Press.

Walter, Barbara F. 2009. *Reputation and Civil War: Why Separatist Conflicts Are So Violent*. Cambridge: Cambridge University Press.