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

Abstract

What effect does repeated play have on reputation building? The international relations literature remains divided on whether, when, and how reputation matters in both inter-state and intra-state 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 build reputations in many ways consistent with the model predictions, but we also identify and attempt to explain several interesting deviations. 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 last 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 (2005) found that governments that invested in reputations for honesty were able to negotiate more effectively with foreign leaders. Tomz (2007) found that governments that invested in reputations for honesty were able to obtain better loans. And Walter (2009) found that governments that built reputations for toughness could deter future separatists. But Snyder and Diesing (1978), Mercer (1996), and Press (2005) all found that reputations did not appear to have any influence on how state leaders behaved even if leaders believed they would.

These disparate findings are puzzling. If reputations work sometimes 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 will choose to invest in reputation, when reputations are likely to affect behavior, and why investments might matter sometimes more than others.

What follows is broken into three parts. In part one 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 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 argued, "the value of reputation

and the extent of reputation building (will) increase with the frequency and opportunities for its use" (1982, pg. 304).

Part two tests these predictions in the controlled, incentivized environment of the laboratory (Bolton and Ockenfels, 2007; Camerer and Weigelt, 1988; Jung et al., 1994). We believe experiments offer three advantages in studying this puzzle (Kinder and Palfrey, 1993; McDermott, 2002). 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, but that these investments do not seem to change behavior in any way. Laboratory experiments allow us to see if individuals react to incentives in ways existing models predict, and if not, what other factors may be influencing their behavior. Second, lab experiments allow the researcher to manipulate the decision environment in specific ways. Instead of relying on observational data from historical cases (which can be highly unreliable), lab experiments allow us to alter the values of specific variables while carefully controlling other parts of the decision context.[1] Third, experiments also reveal how individuals may deviate from purely rational behavior, opening a window for additional theorizing about when and why reputation might matter (McKelvey and Palfrey, 1995)

Our tests reveal three surprising findings. First, in the simple entry-deterrence game we use, repeated play did not have as great an effect on reputation and deterrence as the model predicted. Defenders who played against fewer entrants were just as likely to invest in reputation building as defenders who played against twice as many entrants. Moreover, entrants in the

[1] 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 (i.e., China) that faced four potential outside aggressors over territory, and compare it to a country (i.e., 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.

shorter game were similarly likely to be deterred as those in the longer game. The repeated nature of play, therefore, did not appear to be the critical variable determining whether reputation building would be important. Second, 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 under-invested in reputation building against early entrants, and over-invested 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.

In the third and final section, we discuss why subjects behaved this way, and why the model failed to predict these patterns of behavior. The laboratory experiment suggests two possible explanations. First, our subjects appeared to be motivated by a wider range of 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 a small subset of subjects did not appear to be motivated by monetary goals. This meant that reputation building was less important to some individuals 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, some subjects continued to play poorly. The results of the experiment, therefore, help explain why reputation building may not have been observed in some situations despite the fact that all the purported conditions conducive to reputation building existed.

These are interesting findings and reveal the conditions under which individuals may deviate from otherwise expected behavior. Still, readers should be aware that the strength of

laboratory experiments lies in their ability to test the internal logic of the reputation theory and not in their applicability to real-world settings. Showing when undergraduate students build reputations in a lab (or choose not to) 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 in some contexts but not others.[2] Laboratory experiments are simply the first step in beginning to identify when and why a particular set of behaviors 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 will continue to behave in the future as it has behaved in the past, 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 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.

[2] 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.

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 et. al. (1994) in their laboratory experiment on reputation building. What is different about our experiment is that we vary the number of repetitions, which has not been done. Changing the number of entrants a defender faces allows us to see if repeated play really is a key factor affecting reputation building dynamics.

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 one 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. [3]

<Insert Figure 1 Here>

The game begins with nature randomly choosing whether the defender is committed or uncommitted to fighting a challenge with probability p.[4] 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.[5] Once nature has chosen the defender's type, the entrant must decide whether 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).

[3] Payoff parameters are those used by Jung et al. (1994). We use these parameters to retain consistency in the literature and because they produce clear equilibrium differences depending on the frequency of play.
[4] In reality, defender types may change over time. A defender that was once tough may lose some of their resolve if repeated "fights" against entrants have weakened it. Designing an experiment with endogenous types is beyond the objectives of this particular study.
[5] In this case, the payoffs are 160 for not fighting a challenger and 70 for fighting a challenger.

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.

Theoretical Predictions

In order to determine whether repeated play had an 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. The standard sequential equilibrium model makes three predictions about how defenders and entrants should play with eight entrants.[6] 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 in order to acquire a reputation for toughness and then acquiesce with increasing probability as the number of remaining challengers

[6] We use the sequential equilibrium concept 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. Our appendix reviews a proof by Jung et.al. (1994) and then applies the model to when there are 4 or 8 entrants. In principle, it is possible to utilize other equilibrium concepts, but this has largely not been done by theorists for games of the complexity we consider.

decreases. Third, entrants should base their strategy on information they glean about the type of defender they face and the incentives this defender 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. They 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.

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?

According to the model, 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.

<Insert Figure 2 Here>

Two key differences between the two versions of the game are evident. 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:

> H1: Defenders should be less likely to build reputations (i.e., 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.

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 8 entrant or 4 entrant treatment and subjects could not participate in both treatments.[7] In each session, subjects were assigned randomly to two separate groups, entrants and defendants, referred to simply as first movers and second movers. These neutral terms were used in order 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 1/3rd chance of being committed and 2/3^{rds} chance of being uncommitted. This uncertainty, together with the repeated nature of play, made reputation building a valuable strategy to pursue.

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

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).[8] Each entrant 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.[9] 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 payoffs. Subjects knew that these payoffs would be translated into U.S. dollars at the end of the experiment. Subjects then repeated the

[8]We did this to observe the decision of a defender even when their opponent did not choose to enter. Our design is not equivalent to using the normal form version of the entry-deterrence game, as the defender was told to make a decision conditional on entry. The behavior we observe is very similar to that observed by (Bolton and Ockenfels, 2007) whom elicited strategies sequentially. Furthermore, we compare two treatments that used the same protocol. While in principle the mechanism of strategy solicitation can influence choices, there is considerable debate on this (Bosig et al., 2003; Brandts and Charness, 2000; McLeish and Oxoby, 2004).

[9] This design allowed us to keep all subjects engaged throughout the experiment, as well as maximize the amount of data we could collect within an experimental session.

experiment four times in order to take into account the effects of learning and to generate sufficient data for the analysis.[10]

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.[11] 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)? And, more generally, would subjects deviate in any way from the expectations of our model? The results, as mentioned above, were striking.

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,

[10] 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 re-assigned 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.

[11] 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.

defenders in the eight person design were actually significantly *more* likely to acquiesce than those in the game with fewer repetitions.[12] The model predicted just the opposite.

<Insert Figure 3 Here>

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*. 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 4 and 8 person games did emerge in the 2nd and 3rd periods but it was much smaller than the model predicted.[13] This suggests that repeated play – at least a change from eight to four periods - was not the key factor determining when reputation building would matter and when it would not.

<Insert Figure 4 Here>

Result #2: In the early phases of each game defenders under-invested 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

[12] Difference in mean rate of fighting 24% ($p<.05$, standard errors clustered at individual level).
[13] 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.

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.

As the right side of Figure 3 shows, this was not the case, especially in period one. 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 periods two and three in the eight person game, that probability never approached the 100% 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 entered at a much higher rate than expected. According to the model, approximately 36% of entrants in the four-entrant design should have entered in each period where there was no observed accommodation. Yet, as Figure 4 shows, *over eighty percent* of subjects in the four-entrant game entered in the very first period. This percentage dropped in the second and third period, 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% 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.

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

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

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, therefore, did not have the large effect on behavior that most scholars have assumed. The second is that uncommitted defenders under-invested 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 did so longer than they should have. Uncommitted defenders invested in reputation building and challengers were deterred, but not exactly as our model expected.

Discussion

Why did defenders and entrants behave this way? One explanation for these unexpected results is that subjects simply 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 sub-optimally as a result.

To check if this was true, we examined whether subjects learned how to play over time. Recall that we repeated the experiment four times (each defender faced a sequence of 4 or 8 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 not all mistakes were corrected.

Figure 5 reveals that in both the four and eight entrant design, weak defenders increased their investment in reputation building in early periods as they gained more experience with the game. For the 8 entrant treatment, this is exactly what the theory predicted. Weak 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 continued to back down about half the time against the very first entrant.

Figure 5 also reveals that little 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 latter periods because mistakes were less costly to the defenders who made them and were, therefore, easier to overlook.

<Insert Figure 5 Here>

Still, the fact that defenders continued to under and over-invest in reputations even as they gained experience 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 rest. In an attempt to identify any heterogeneity in our subjects, 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. A subset of individuals acquiesced in every period and never changed their behavior, and it was these uncommitted defenders who appeared to be driving the anomalous results in the early periods. Furthermore, 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 offers further evidence that a certain type of individual – one who always acquiesces when in the uncommitted role and, therefore, always plays according to type – is driving the unusual results.

<Insert Figure 6 Here>

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

If one focused only on the immediate strategic interaction, this logic would be correct. Uncommitted defenders who chose to acquiesce in any given period received 160 points versus 70 points if they chose to fight. This logic, however, does not hold up if one has a longer-term view of payoffs. Fighting in one period delivers fewer points in the short-term, but *could* have the effect of deterring the next challenger, delivering far greater points over time. 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. A number of subjects revealed this myopic logic in 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 explains why some uncommitted defenders would consistently over-invest in reputation building late in each game? We believe there are three potential explanations. It is possible that these individuals 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 8 entrant design who backed down early paid a high price for this mistake because they then faced a higher rate of entry from the remaining challengers. Over-investment late in the game, however, was less costly and, therefore, harder to catch. Figure 5 above revealed that behavior changed very little in the last stages of the game, suggesting that less learning was taking place.

A second related explanation, however, has to do with differences in preferences. It is possible, for example, that some weak 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 be unwilling to reveal their weakness to any player, and will only acquiesce 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."

A third explanation again has to do with cognitive abilities. Some subjects may never figure out the game and will fight longer than necessary thinking this is the optimal strategy to pursue. Numerous subjects, for example, offered the following advice to future players in our exit questionnaire:

> "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 final explanation is that individuals discount the future differently. Some may value immediate gains more highly than future ones, or feel the pain of immediate losses more than future losses. Both of these preferences would affect the type of strategy a subject would pursue.

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 this strategy because it

delivered better payoffs. 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, or fight too late. If this was the case, then entrants (or at least particularly strategic ones) could anticipate that less fighting would take place early on, more fighting would take place later, and they could adjust their strategy to match this. Second, they may have also anticipated that some individuals would not want to bluff, or conversely would enjoy bluffing, and that this would change the strategic game at least in early and late periods.[14] Either way, it appears as if subjects were aware that not all people would behave rationally all the time and they adjusted 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 was a critical variable affecting behavior. In order to do this, we introduced a methodology relatively unknown in the IR literature: incentivized laboratory experiments.[15] This allowed us to isolate variables that game theoretic models suggested should affect behavior, and to test whether they had the predicted effect.

[14] 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).

[15] We are, however, not the first to use laboratory experiments in IR. The first experimental work in IR was conducted by Mort Deutsch and Marc Pilisuk in the early 1960s. Currently, important work in this vein is being done by Alex Mintz, Nehemia, Geva, Francis Beer, Jonathan Wilkenfeld, Mark Shafer, Rose McDermott, Daniel Druckman and Phil Tetlock.

The sequential equilibrium model has always assumed that reputation building would become more valuable as the number of repetitions increased. It predicted that weak 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 too much. 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 served to deter entry – it also revealed that a certain subset 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 tells us that investments in reputation building are likely to be made in games that are repeated relatively few times, in fact, fewer times than our model had predicted. Second, it tells us that some 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. This may help explain the mixed results in empirical studies of reputation building in international relations. Some leaders may never bluff, others always will, and this affects the strategies of rival leaders with whom they are interacting.

The experimental results suggest 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 (2004) argue that the role of emotion plays a more central role than standard models permit. 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 strategically more sophisticated than others. Cognitive hierarchy equilibrium (CHE) models (Camerer et al., 2004) assume a distribution of strategic sophistication, and best responses are made in light of the fact that not all opponents will be equally talented. 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.

What does this tell us about reputation building in international relations? 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. It may be that some leaders will take longer to process how a particular strategic game is played, and will, therefore, behave in ways that are less-than-rational. Others may be more principled, honest or aggressive.

What the experiment offers, at the very least, is a starting-point of where to look if we hope to explain certain anomalies in the real world. More work needs to be done, but if we can apply these insights to studies of actual interactions in IR, we may be able to better understand behavior, better predict outcomes, and better identify optimal counter-strategies than what we are currently able to do.

Figure 1

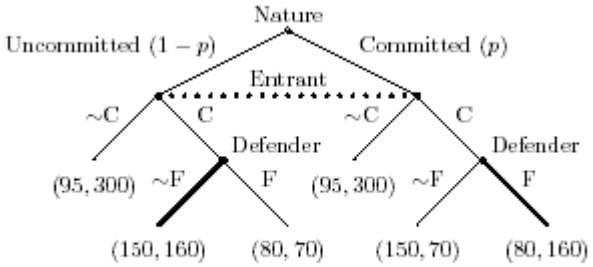
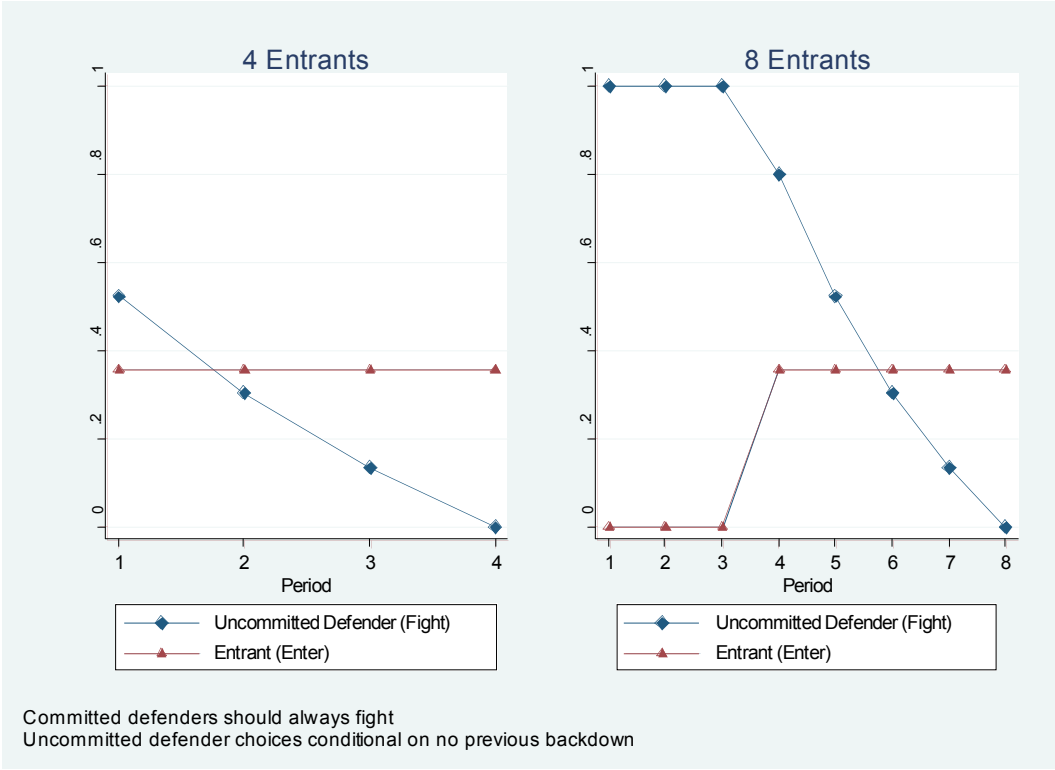


Figure 1: The Structure of a Single-Shot Play

Figure 2



Committed defenders should always fight
Uncommitted defender choices conditional on no previous backdown

Figure 3

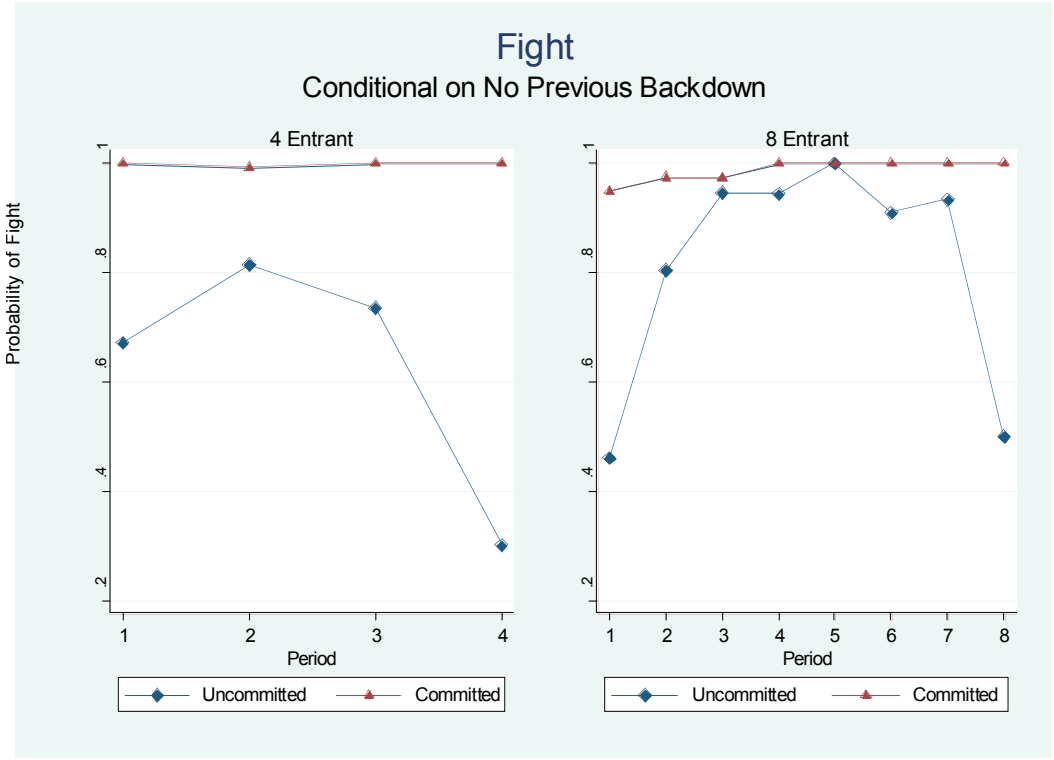


Figure 4

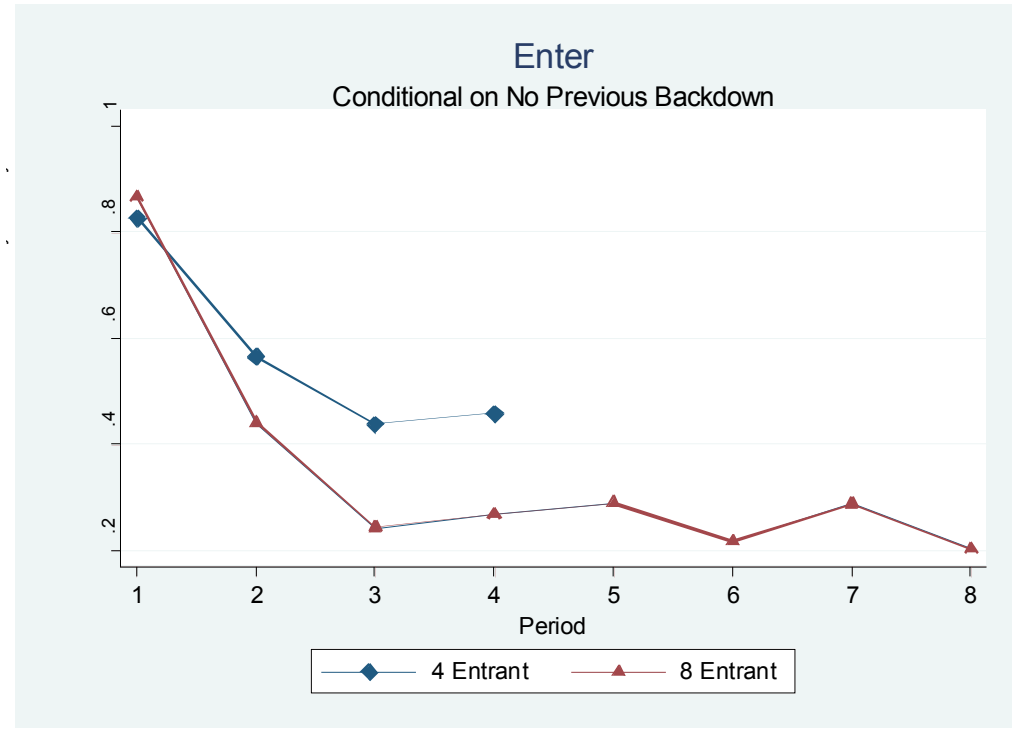


Figure 5

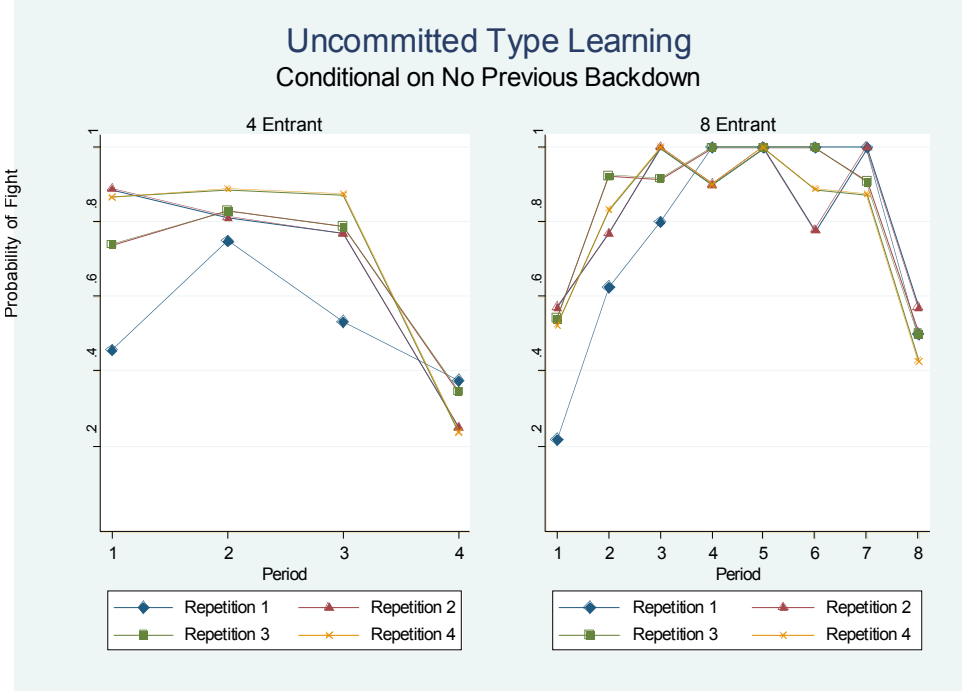
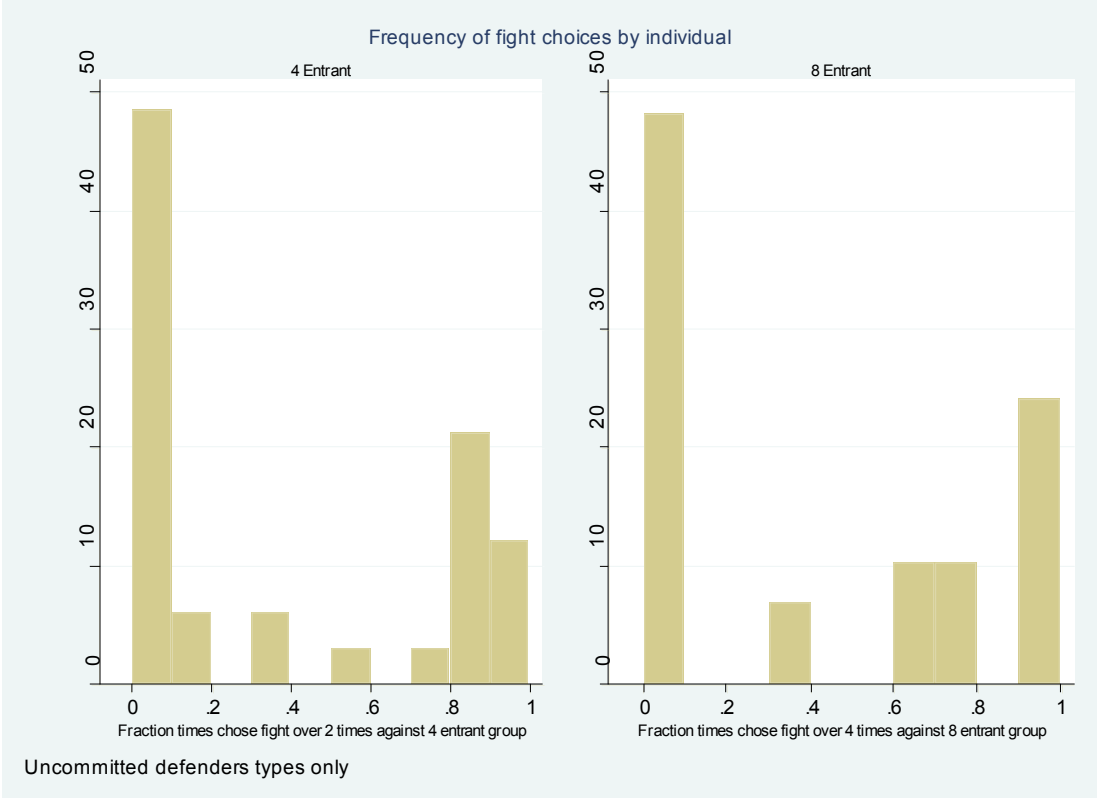


Figure 6



Bibliography

Bolton, G. E. and A. Ockenfels, 2007, Information Externalities, Matching and Reputation Building - A Comment on Theory and an Experiment, (University of Cologne Working Papers Series in Economics: 17, http://ockenfels.uni-koeln.de/download/papers/bolton_ockenfels_information_externalities.pdf).

Bosig, J., J. Weimann and C. L. Yang, 2003, The Hot versus Cold Effect in a Simple Bargaining Experiment. *Experimental Economics* 6, 75-90.

Brandts, J. and G. Charness, 2000, Hot vs. Cold: Sequential Responses and Preference Stability in Experimental Games. *Experimental Economics* 2, 227-238.

Camerer, C. and K. Weigelt, 1988, Experimental Tests of a Sequential Equilibrium Model. *Econometrica* 56, 1-36.

Camerer, C. F., T.-H. Ho and J.-K. Chong, 2004, A Cognitive Hierarchy Model of Games. *Quarterly Journal of Economics* 119.

Fehr, E. and K. Schmidt, 1999, A Theory of fairness, competition and cooperation. *The Quarterly Journal of Economics* 114, 817-868.

Jung, Y. J., J. H. Kagel and D. Levin, 1994, On the existence of predatory pricing: an experimental study of reputation and entry deterrence in the chain-store game. *The Rand Journal of Economics* 25, 72-93.

Kinder, D. R. and T. R. Palfrey, 1993, *Experimental foundations of political science* (University of Michigan Press, Ann Arbor) Pages.

McDermott, R., 2002, Experimental Methodology in Political Science. *Political Analysis* 10.

McKelvey, R. and T. Palfrey, 1995, Quantal Response For Normal Form Games. *Games and Economic Behavior* 10, 6-38.

McLeish, K. and R. Oxoby, 2004, Specific Decision and Strategy Vector Methods in Ulimatum Bargaining: Evidence on the Strength of Other-Regarding Behavior. *Economics Letters* 84, 399-405.

Mercer, J., 1996, *Reputation and International Politics* (Cornell University Press, Ithaca) Pages.

Mesquita, B. B. d. and R. McDermott, 2004, Crossing No Man's Land: Cooperation from the Trenches. *Political Psychology* 25, 271-287.

Milgrom, P. and J. Roberts, 1982, Predation, Reputation, and Entry Deterrence. *JOURNAL OF ECONOMIC THEORY* 27, 280-312.

Press, D. G., 2005, *Calculating Credibility: How Leaders Assess Military Threats* (Cornell University Press Ithaca) Pages.

Sartori, A., 2005, *Deterrence By Diplomacy* (Princeton University Press, Princeton, NJ) Pages.

Snyder, G. H. and P. Diesing, 1978, *Conflict Among Nations* (Princeton University Press, Princeton) Pages.

Tomz, M., 2007, *Reputation and International Cooperation: Sovereign Debt Across Three Centuries* (Princeton UP, Princeton, NJ) Pages.

Walter, B., 2009, *Reputation and Civil War: Why Separatist Conflicts Are So Violent* (Cambridge University Press, Cambridge) Pages.