MIT
POLITICAL SCIENCE

Massachusetts Institute of Technology
Political Science Department

Working Paper No. 2013-16

Are Costly Signals More Credible?
Evidence from Three Experiments

Kai Quek, MIT

Are Costly Signals More Credible?
Evidence of Sender-Receiver Gaps

Kai Quek*

Abstract

The idea that costly signals are more credible is a long-standing hypothesis in international politics. However, little is known on how costly signaling actually works. Causal evidence is elusive because the effect of a costly signal is almost always confounded with the effects of other previous or simultaneous information. I design three controlled experiments to study how the logic of sinking costs operates. I find that signalers randomly assigned with high resolve are more likely to sink costs, but receivers do not acquiesce in line with signaler expectations, despite the sunk costs suffered. The logic of sunk-cost signaling is strong at the signaler's end but not at the receiver's end. There is a sender-receiver gap in how the same deterrence interaction is perceived at the two ends of the signaling mechanism, contrary to what the theory of costly signaling automatically assumes.

Forthcoming in *Journal of Politics* (2016)

* Assistant Professor, Department of Politics and Public Administration, University of Hong Kong, Pokfulam Road, Hong Kong. Email: quek@hku.hk.

For comments and criticisms, I am grateful to Dan Altman, Eric Grynaviski, Mike Sances, Dustin Tingley, Joshua Wu, and particularly Dave Clark, Jim Snyder, B.K. Song, and my anonymous reviewers. I also thank the Behavioral Research Laboratory at the Massachusetts Institute of Technology (MIT) for its facilitation. The experiments were approved by the Committee on the Use of Humans as Experimental Subjects at MIT. The MIT Department of Political Science provided seed funding. Supplementary materials for this article are available in the appendices in the online edition. Replication files are available in the JOP Data Archive on Dataverse (http://thedata.harvard.edu/dvn/dv/jop).

Signaling is ubiquitous in international politics.[1] States send signals to one another as they communicate threats or promises with the intent to deter, compel, or persuade. But what kinds of signals are credible? The theory of costly signaling suggests that a credible signal should differentiate between resolved and unresolved states by carrying some costs that would discourage unresolved states from sending that signal.[2] Two general mechanisms of costly signaling exist: One is sunk-cost signaling, which creates direct costs that cannot be recovered, and which does not affect the relative value of escalation or compromise (Fearon 1997). To date, however, no study has provided a clean causal examination of how the logic of sinking costs works in crisis deterrence, despite its importance.

Theoretically, the concept of costly signaling is central to our understanding of how international exchanges are conducted under asymmetric information. While there are two mechanisms of costly signaling – sinking costs and tying hands – much attention has centered on the latter after Fearon's (1994) seminal article on audience costs.[3] Despite its relative neglect in the literature, the logic of sinking costs features in a wide range of international actions, from military mobilizations and peaceful reassurances to international market distortions and economic sanctions (e.g. Morrow 2000; Slantchev 2005; Glaser 1994; Kydd 2005; Gartzke, Li and Boehmer 2001; Lektzian and Sprecher 2007). States have also frequently used sunk costs to signal their grand-strategy interests (Fearon 1997, 71). In the context of international security, these costs include that of stationing troops and weapons abroad (Fearon 1997) as well as coordinating military alliances and foreign policies (Morrow 1994).

The mechanism also connects to a crucial puzzle in international security: Why do states

[1] A signal is a piece of information intentionally communicated by one party ("signaler") and observed by another ("receiver"). Credible signaling is achieved when it is convincingly shown that the signaler is resolved to fulfill its threat or promise. Non-credible signaling fosters miscalculations that can pave the way to bargaining failure and war (e.g. Blainey 1988; Fearon 1995; Filson and Werner 2002).

[2] See, esp., Spence (1973) and Fearon (1994; 1997). Crawford and Sobel (1982) and Farrell (1987) show that costless signaling is occasionally informative.

[3] Tied-hands signaling binds the signaler to a higher cost of backing down if the opponent does not back down, but is costless if the opponent does back down (Fearon 1997).

pay high costs to deploy and maintain military assets that are in fact militarily superfluous? Consider, for instance, why the United States deployed and maintained nuclear missiles in Europe from the Cold War to the present day.[4] The missiles are unlikely to change the outcome of a nuclear war, given long-range delivery technology and mutually assured destruction. Nonetheless, by sinking costs, these missiles may help to signal U.S. resolve in defending its European allies (O'Neill 1990; Fearon 1997; Slantchev 2011). Since proving a state's commitment to fight for another state is hard, states sink costs to sharpen their alliance credibility (Morrow 1994; Morrow 2000; Fuhrmann and Sechser 2014). The costs are high: deployments of nuclear assets in Europe increase the risks of nuclear mishaps and misuse (Sagan 2003; Kristensen 2005), and incur large financial costs (House Committee on Armed Services 2013). These costs and risks justify themselves on the assumption that sinking costs sharpens the credibility of one's resolve. But does sunk-cost signaling really work?

This paper presents controlled experimental evidence on the logic of sinking costs. Despite its practical significance to international politics, there is surprisingly little evidence on how the mechanism of sunk-cost signaling works in practice. A rigorous empirical test involves an examination of *both* the signaler and receiver ends in the mechanism: First, do resolved signalers use sunk costs to signal their resolve? Second, does a sunk-cost signal change the receiver's behavior?

Due to the noisy nature of the real world, it is challenging to construct a clean causal test of the signaling mechanism with observational data. To cut through the challenge, I design a series of controlled experiments to explore whether and how resolve shapes the signaler's willingness to send a sunk-cost signal, and whether and how that costly signal shapes the receiver's behavior. Specifically, I randomly divide the subjects into signalers and receivers, and incentivize their interactions using a signaling game with a separating equilibrium. The

[4] NATO had hosted as many as 7,300 tactical nuclear weapons across Europe during the Cold War (Knops 2010).

experiments test whether signalers randomly assigned with high resolve are indeed more likely to use a signal with sunk cost, and whether receivers that see a signal with sunk cost are indeed less likely to challenge the signaler.

I find that resolved signalers are much more likely to sink costs, but sunk-cost signals have an insignificant effect on the receiver's acquiescence rate. The findings expose a surprising gap between signalers and receivers: Signalers believe sunk costs make their threats more credible, and they choose to suffer sunk costs willingly. Yet receivers do not acquiesce in line with the signalers' expectation, despite the sunk costs suffered. The key point is *not* that sunk-cost signaling has no credibility effect, but that those randomly assigned to the role of the signaler behaved in a way consistent with the logic of sunk-cost signaling whereas those in the role of the receiver did not. Signalers and receivers do not respond to the logic of sinking costs in the same way. There is a systematic asymmetry in how the same deterrence interaction is perceived at the two ends of the signaling mechanism, contrary to what the theory of costly signaling assumes.

I used a variety of design strategies to test the robustness of the results. I replicated the experiment with two different subject samples and in two different environments, over the Internet (Experiment 1) and in the laboratory (Experiment 2). I also implemented a survey experiment (Experiment 3) that contextualized sunk-cost signaling in a concrete crisis scenario and facilitated a straightforward elicitation of credibility estimates free from a game-theoretic structure. While there is variation in the point estimates, the main conclusions are consistent across the different designs.

While surprising, the experimental results resonate with a recent observational study of sunk-cost deterrence in the context of stationing nuclear weapons on foreign territories. Historically, great powers have sunk massive costs in stationing nuclear arsenals on their protégé's territory, yet these sunk costs appear to have no effect on whether their protégé would be targeted in a militarized dispute (Fuhrmann and Sechser 2014). Great powers sink costs thinking that it would bolster deterrence, but the receivers do not respond in line with the signalers' expectation. The experimental evidence at the individual level clicks with the observational evidence at the interstate level, suggesting that the sender-receiver gap may be a deeper phenomenon in human behavior.

The paper is organized as follows. First, I describe the signaling game used in the experiments and extract a set of testable theoretical expectations. Second, I explain how each experiment is designed and implemented. Third, I report the results. Finally, I conclude by connecting the implications of the results with the puzzle of why sunk-cost signaling is frequently used despite its costs and inefficiencies. For the rest of the paper, the generic term "costly signal" refers specifically to a signal with sunk cost.

STUDYING COSTLY SIGNALING

Do resolved signalers sink costs to signal their resolve, and does sinking costs change the receiver's behavior? Two methodological challenges prevent us from answering these questions.

Two Fundamental Challenges

The first is the problem of observability. It is hard to identify both the signaler's true resolve at the point of signaling and the receiver's credibility belief at the point of signal reception, given that these are private information. The second is the problem of confounding. This problem can be analytically broken down into *between-signals* confounding and *within-signal* confounding.

In the former, the credibility effect of a sunk-cost signal is confounded with the effects of other current or previous signals or information – at least some of which are opaque to the researcher. Between-signals confounding almost always occur in real-world cases of international signaling, insofar as each case comes with a history and cannot reasonably be viewed as *tabula rasa.* Yet even if a particular case is assumed to be history-free, it may still be subject to within-signal confounding in one (or both) of two forms. The first is when the case involves more than one signaling mechanism. For example, the effects of sunk-cost and tied-hands signals are often mixed together in real-world cases (Fearon 1997, 70). The second is when the signaling effect of the signal is contaminated by the substantive effect of the action. When a signal involves an action, the target audience has to consider not only what is the *information* transmitted in the signal, but also how to react to the *substance* of the action. For instance, while costly-signaling actions such as the mobilization of troops may help to communicate the signaler's resolve, they may also induce strategic fears in the target arising from the substantive nature of the action as an escalatory move or offensive maneuver. The information effect of the signal and the substantive effect of the action can pull the target in opposite directions: the information effect directs the target to negotiate; the substantive effect directs the target to escalate.[5]

The twin problems of confounding and observability make it challenging to infer causation in a signaling mechanism using observational data. To overcome the two problems, I use experiments to construct a causal test in a controlled setting. Observability is achieved with the random assignment of resolve and the direct measurement of credibility responses at the point of signal reception. Between-signals confounding is eliminated with signaling games conducted in a *tabula-rasa* setting. Within-signal confounding is avoided through the experimental isolation of the signaling effect while holding the substantive effect of the action constant.

Signaling experiments are rare in international relations, but relevant experiments can be found in the economics literature that tests different equilibrium selection devices in signaling games.[6] The closest examples are by Miller and Plott (1985) with signaling games set in

[5] I thank Dave Clark for highlighting this point. This point connects to the literature on whether deterrent threats and military moves can inadvertently incite the opponent to escalate, especially in the context of a security dilemma (Lebow 1987, 211; Huth 1999, 29; Slantchev 2005, 546).

[6] See Brandts and Holt (1992); Banks, Camerer and Porter (1994); Cooper, Garvin and Kagel (1997);

experimental markets with buyers and sellers, and by Potters and Van Winden (1996) with a signaling game that relates to lobbying and advertising.[7] The experiments reported here differ in several ways. First, the signaling game is structured and framed to relate directly to the international relations literature on crisis interaction. The game captures a deterrence crisis with two parties contending for a valuable prize; the signaler issuing a threat to the receiver; and the receiver deciding whether to confront the signaler. Unlike signaling experiments in which the receiver either receives a message or doesn't, the receiver here *always* receives a threat from the signaler with the same content: the only difference lies in whether the threat carries a sunk cost. Second, the experiments are designed to provide a sharp test of the threat-based mechanism of sunk-cost signaling. The payoffs of the two signaler types are configured to make type-separation possible and the type-separation logic as transparent as possible. I also black-box the signaler's reaction to the receiver's choice: By definition – and as reflected in the payoffs – the resolved (unresolved) type will always (never) fulfill its threat to fight if the receiver does not acquiesce. This design ensures that the receiver's focus in the game falls sharply on the signaler type and the signal sent, with absolute certainty on how a given type of signaler would react. Third, I measure signal credibility not only based on the receiver's action, but also based on a measurement of signal credibility as perceived by the receiver. I also use a survey experiment to elicit signal credibility evaluations from a diverse national sample free from a game-theoretic structure. To my knowledge, this is the first attempt to experimentally isolate the effect of sunk-cost signals on credibility estimates in the literature.

and Cooper and Kagel (2005). In international relations, Tingley and Walter's (2011) cheap-talk experiment showed that when signalers and receivers are given a private channel of costless communication, it deters the receivers in early periods of play and makes the signalers more eager to fight.

[7] Miller and Plott (1985) had experimental markets where sellers choose product enhancements that differed in cost based on the quality of their product. The study found that in general, high-quality sellers are more likely to opt for more costly signals. Markets with relatively low marginal costs of signaling a high-quality product separated at least in the final periods, but not markets with relatively high signaling costs. The general finding is replicated in Potters and Van Winden (1996) (based on data from the last ten periods), where the signaler chooses between a costly message and no message over twenty periods with the signalers and receivers reversing roles in the last ten periods. The game has no separating equilibrium.

Finally, the experiments control for learning effects and reputational concerns. Signaling experiments in economics are often conducted with exactly repeated rounds to allow for learning over time. This is relatively realistic insofar as many market transactions are of a repetitive nature. However, this is less realistic in the case of an international crisis, where a failed deterrent signal can end the crisis in a war, with no further replay of the same crisis. In addition, understanding the behavior of first-time players is important in itself, given that inexperience is all too common in international politics.[8] It is also necessary to control for learning and focus on the one-shot responses generated in the interaction, in order to evaluate cognitive responses in their purest form (Costa-Gomes and Crawford 2006). In particular, it is important to control for reputation and the shadow of the future if we are to have a clean test of the causal effects of sunk-cost signaling. The experiments are set up to explicitly control for these confounders. Two of the experiments (Experiments 1 and 3) are designed in a one-shot setting. In Experiment 2 with multiple rounds, subjects play the same game repeatedly but are not told the outcomes until the end of the experiment. Thus, subjects become familiarized to the game without any leakage of information about the receiver's choice and the signaler's type.

Theoretical Expectations

I construct a signaling game to capture the mechanism of sunk-cost signaling in its purest form. The goal is to strip the mechanism down to its simplest and most basic level to allow for a clean experimental test. The game sets up a deterrence crisis between two players A and B interested in a valuable prize. A sends a threat to B that it will fight B if B does not stay out. There are two types of Player A: true type and fake type. By definition, a true type will *always* fulfill its threat while a fake type will *never* fulfill the threat. Thus, the true type is resolved but the fake type is not.

The signaler gets a higher payoff if the receiver stays out, while the receiver who challenges

[8] I thank an anonymous reviewer for highlighting this point.

the signaler gets a payoff conditional on the signaler's resolve: a low payoff if the signaler is resolved, and a high payoff if the signaler is not. The signalers are randomly assigned with high and low private valuations of the prize that generate high and low levels of resolve.

Formally, let Nature randomly assign A as a high or low-resolve type unknown to B. A chooses a threat with a sunk cost (c = 2) or a threat without a sunk cost (c = 0). B observes the threat and decides whether to confront A. If B stays out, the payoff is (10 - c, 6) if A is high-resolve and (3 - c, 6) if low-resolve. If B confronts, the payoff is (4 - c, 2) if A is high-resolve (i.e. A will, by definition, fulfill its threat and fight), and (2 - c, 10) if A is low-resolve (i.e. A will never fight). Figure 1 diagrams the game.

[Figure 1 here]

The utility and cost parameters are chosen to maximize the clarity and simplicity of the game, with the following considerations. First, the payoffs are constrained within the integer range in [0, 10] to keep the calculations as simple as possible for the subjects. Second, B's payoffs are constrained such that an equal expected value between the two choices (confront or stay out) is assured, given a 0.5 ex-ante probability that A is a high-resolve type. B's maximum payoff is the same as the maximum payoff for a high-resolve-type A; B's minimum payoff is the same as the minimum payoff for a low-resolve-type A, disregarding sunk cost which is a voluntary choice made by A. Finally, A's payoffs are defined to ensure that a threat with sunk cost is type separating, given the first and second constraints. This allows the threat with sunk cost to be strictly dominated by the threat without sunk cost for the low-resolve type: the low-resolve type gets *at most* 1 if it sends the costly threat, but *at least* 2 if it sends the costless threat. Even without solving the game, the intuition is clear: all players should know that low-resolve types would never send the costly threat, which allows for type separation and identification based on whether a costly signal is sent.

As a consequence, the game has a separating perfect Bayesian equilibrium as follows: A will send the threat with sunk cost if it is a high-resolve type and the threat without sunk cost if it is a low-resolve type. B will stay out if it receives the costly threat and will confront if it receives the costless threat. In B's belief, given that the costless threat is observed, the probability that A is high-resolve is 0 while the probability that A is low-resolve is 1. In other words, signalers are incentivized to send a costly threat if they are a high-resolve type and a costless threat if they are a low-resolve type, whereas receivers are incentivized to confront if they see a costless threat and to back down if they see a costly threat.[9]

Two expectations are extracted from the game. First, we should expect the signaler to sink costs if it is resolved. Second, we should expect the receiver to stay out when it sees a sunk-cost signal. These theoretical expectations follow the logic of costly signaling laid by Schelling (1960, 1966) and developed by others in the context of international politics (e.g. Jervis 1970; Powell 1987; Morrow 1989; Fearon 1997; Slantchev 2005). Resolved actors reveal their resolve by incurring costs and risks that unresolved actors would not be willing to take. Thus, sinking costs enhances credibility by separating resolved actors from unresolved ones (Fearon 1997).

EXPERIMENTAL DESIGN

I design three experiments. Experiment 1 implements the signaling game over the Internet. Experiment 2 replicates the same game in the laboratory. Experiment 3 deploys an online national survey. Table 1 summarizes the differences between the three. The following subsections specify the design of each experiment. For ease of presentation, Experiment 3 will be discussed after this section.

[Table 1 here]

[9] Appendix 3 provides the proofs.

Experiment 1

253 U.S. adult residents were recruited on 21-22 February 2013 through Amazon.com's Mechanical Turk (AMT). The validity of AMT as an experimental tool has been tested across different fields in social science, including economics (see Horton, Rand, and Zeckhauser 2011) and political science (see Berinsky, Huber, and Lenz 2012). AMT is particularly useful for recruiting a diverse sample of motivated subjects for experimental purposes (Berinsky, Huber and Lenz 2012, 361).[10] Participants were linked from AMT to the website where the experiment was hosted. Each participant received $0.51 as participation fee and a bonus payment between $0.00 and $1.00, which are up to about three times the normal payment rates in the AMT world at the time of the experiment. Participants played the sunk-cost signaling game at the very start of the session. After completion, they transited into a bargaining game followed by a risk-aversion game that measured their risk preferences.[11]

Participants were randomly divided into two groups: signalers (Player A) and receivers (Player B). Those assigned as Player A were randomly divided into true types (high-resolve) and fake types (low-resolve). All groups received the same instructions. The instructions highlighted that Player A knows whether it is a true type or a fake type, but A's type is unknown to Player B. B gets a high payoff if it confronts a fake type but a low payoff if it confronts a true type. The game was explained to participants in detail, with questions at the end to test their understanding. Participants also saw a summary of the game and payoffs twice – on the screen just before the decision screen, and on the decision screen

[10] Berinsky, Huber and Lenz (2012) found that "the MTurk [AMT] sample does not perfectly match the demographic and attitudinal characteristics of the U.S. population but does not present a wildly distorted view of the U.S. population either. Statistically significant differences exist between the MTurk sample and the benchmark surveys, but these differences are substantively small. MTurk samples will often be more diverse than convenience samples and will always be more diverse than student samples. Thus, if we treat the MTurk as a means for conducting internally valid experiments, instead of a representative sample, the MTurk respondent pool is very attractive." (361)

[11] Participants were told that they would play three different games. The computer would randomly choose one out of the three games they played, and count their point earnings in that game as bonus payment. Each game had a total possible value of 10 points, with each point equivalent to $0.10 in bonus. The average total earning was $1.02.

itself.

Thereafter, the participants made their decisions. Here is where A could choose whether or not to sink costs. Specifically, A could send either a threat without sunk cost ("Threat X with cost = 0 points") or a threat with sunk cost ("Threat Y with cost = 2 points"). Then, B would observe the signal (with or without sunk cost), and decide whether or not to back down. Specifically, B would observe the threat sent by A, which is displayed on B's computer screen, and decide whether "to stay out or to confront." Appendix 1 shows the experimental instructions in full. The rest of the paper will refer to a true type as "resolved" (high-resolve) and a fake type as "unresolved" (low-resolve).

To generate the game outcome, each participant's decision was randomly matched with the decision of another participant in the opponent role. Opponents were randomly drawn from participants who had played the game in the opponent role. I collected the pool of opponent decisions by implementing the same game in a pre-experiment three weeks earlier on AMT. These decisions were programmed into the actual experiment to generate the game outcomes for payment purposes, and the different frequencies of costly signals conditional on the opponent's signaler type. Participants in the pre-experiment were excluded from the actual experiment.[12] Additional details are documented at the end of Appendix 1.

To ensure that respondents paid attention and understood the signaling game, respondents were tested with four questions on how the game worked. Two questions were straightforward. The other two questions were designed to be hard tests: they were more complicated and could not be easily answered without careful thought. Aside from the test questions, which also served a training purpose, there were three additional safeguards to ensure that respondents understood how the game worked. First, in the screen immediately after the test questions, respondents saw the answers to the questions, which showed what a respon-

[12] The recruitment notice for the actual experiment prohibited repeat participants. Each AMT subject has a unique "Worker ID," which allows me to trace repeat participants and exclude them from the dataset.

dent got right or wrong. Next, respondents saw a screen that summarized the game and the payoffs.[13] Finally, the information also appeared on the decision screen itself.

There is a potential concern that Internet respondents may not pay sufficient attention to the experimental setting, which can affect the results in Experiment 1. I dealt with the concern in five ways. First, I recruited subjects from AMT, who are known to be more attentive than the average Internet respondent (Berinsky, Huber and Lenz 2012). Second, the monetary bonuses were performance-contingent, providing real incentive for respondents to play and "win" the game. Third, recruitment was restricted to those with a minimum 95% approval rate for prior AMT tasks, preventing respondents without a good record from participating in the game. Fourth, the data analysis excludes all subjects who answered more than one test question incorrectly. The final sample has a total of 222 subjects after excluding 31 subjects. Finally, there were three additional safeguards for respondents to revise and confirm their understanding of the game, as described earlier.

Experiment 2

The laboratory experiment was programmed and implemented on the z-Tree platform with the participants interacting with each other anonymously through computers (Fischbacher 2007). The experiment was conducted in four sessions at the MIT Behavioral Research Laboratory from 7 February to 8 March 2013. A total of 64 students were recruited from MIT through the laboratory. As the school has one of the most competitive and stringent selection standards in the nation, we can expect these subjects to be sharper and quicker with problem-solving than the average U.S. adult. The MIT sample is recruited to allow for a "stress-test design," so that we can see if the same result is obtained despite the different samples and settings (Morton and Williams 2010, 242-3). Participants were paid solely

[13] The summary screen was excluded in Session 1 of Experiment 2 ($n = 14$). The conclusions from Experiment 2 remain unchanged when Session 1 data is excluded from the statistical analysis.

based on their performance. [14] Each subject participated in only one session.[15]

The rules and instructions in the signaling game were largely similar in Experiments 1 and 2 (see Appendix 1), with two basic differences. Experiment 2 used a one-stage elicitation of retrospective credibility estimates, as two-stage elicitation might be more cumbersome and less effective with repeated rounds.[16] In addition, subjects in Experiment 2 played three rounds of the signaling game instead of a single round. This design feature serves three objectives: first, to collect more observations without deviating too far from a one-shot game setting; second, to replicate the one-shot result in Experiment 1 with the first-round result in Experiment 2; and third, to test the robustness of the result with repeated rounds that make subjects engage in repeated thinking about the game and their roles in the game. Subjects played the first round without knowing if the subsequent round would be similar. In each round, subjects were randomly assigned as either Player A or Player B, and randomly matched with one another. Anonymous random matching prevents reciprocity and reputation effects from confounding the causal estimates. Subjects were *not* told the game outcomes – which would leak information about the receiver's choice and the signaler's type – until the end of the entire experiment.

RESULTS

In Experiments 1 and 2, signalers are randomly assigned as either a resolved type or an un-

[14] The computer randomly chose nine out of the fifteen rounds they played, and counted their point earnings in those rounds as payment. Each round had a total possible value of 10 points, with each point equivalent to $0.50. Subjects earned $19.42 on average for the one-hour session.

[15] Each session is time-shared between the sunk-cost signaling game and a bargaining game. At the start of each session, participants were told that they would be playing a total of fifteen rounds over multiple scenarios. A scenario would have one or more rounds, and each round would be separate and independent from one another. Participants played eleven rounds of a bargaining game before transiting into the sunk-cost signaling game. After playing three rounds of the signaling game, participants played a risk-aversion game in the final round that measured their risk preferences. The potential for spillover effects is limited, given that the bargaining game is clearly different from the signaling game, and that subjects know that each round is independent from previous rounds. Subjects also know that they will be randomly matched with different opponents in the signaling game.

[16] One-stage elicitation presents Player B with the full menu of options when they were asked if they thought that their opponent was likely to be a true type. Appendix 1 describes the two-stage elicitation process in Experiment 1.

resolved type. Due to randomization, the experimental groups are identical in expectation across all observed and unobserved characteristics including their traits and demographics. As random assignment rules out by design the possibility of confounding by omitted variables, we can obtain unbiased causal estimates with straightforward significance tests. Do resolved signalers use sunk costs to signal their resolve? Figure 2 compares the percentages of resolved and unresolved signalers who sent a costly threat.

[**Figure 2 here**]

In Experiment 1, 30% of signalers randomly assigned as resolved signalers chose the threat with sunk cost compared to only 11% of unresolved signalers (two-tailed test of proportion, $p = 0.013$, $n = 112$).[17] In total, costly threats made up 21% of all threats sent regardless of type. 74% of costly threats were sent by resolved signalers.

In Experiment 2, 49% of resolved signalers chose the costly threat compared to just 10% of unresolved signalers ($p < 0.0001$, $n = 96$). 29% of all threats sent (regardless of type) were costly threats, and resolved signalers sent 82% of all costly threats. In the first round of the signaling game, 43% of resolved signalers sent a costly threat compared to 0% of unresolved signalers ($p = 0.0021$, $n = 32$). Costly threats accounted for 19% of all threats sent in the first round (regardless of type), all of which were sent by resolved signalers.

The evidence shows that resolved signalers are much more likely to sink costs – or "burn money" in common parlance – than unresolved signalers. Table 2 presents a robustness check with a logit analysis of the relationship between the level of resolve and the choice of costly threat. The binary variable, *High-Resolve*, is coded 1 if the signaler was randomly assigned as a resolved signaler, and 0 if otherwise. The model specifications include a baseline bivariate model and an alternative model with controls for risk preference (for Experiments 1 and

[17] Unless stated otherwise, the p-values in parentheses in the rest of this paper are based on a two-tailed test of proportion.

2) and round and session fixed-effects (for Experiment 2). Risk preference is measured on a summed score based on the risk-aversion game at the end of the experiment: the higher the score, the greater the willingness to take risk (see Appendix 1 for details). The control variables have no causal interpretation, however, and the only causal interpretation drawn is from the *High-Resolve* variable. The analysis shows that the *High-Resolve* variable is positive and significant across the different model specifications in Experiment 1 ($p < 0.02$) and Experiment 2 ($p \leq 0.001$), with or without controls.

[Table 2 here]

Next we turn to the receiver end of the mechanism: Does sinking costs change the receiver's behavior? Here we examine whether receivers choose to stay out or confront, conditional on whether they receive a threat with or without sunk cost. Given the payoffs, we should expect receivers to stay out if they believe they have encountered a resolved opponent. By definition, a resolved opponent will always fulfill its threat to fight – and depress the receiver's payoff to the minimum – if the receiver does not stay out. Hence, the receiver's decision to stay out provides a clear behavioral measure of the credibility of the threat.

Figure 3 compares the percentage of receivers who chose to stay out based on whether they received a threat with sunk cost. In Experiment 1, the percentages are similar across the two groups: 55% of receivers who saw a costly threat decided to stay out compared to 51% of receivers who saw a costless threat, with no significant difference (p = 0.74, n = 110). On the whole, 53% of all receivers decided to stay out in Experiment 1.

[Figure 3 here]

60% of receivers stayed out in Experiment 2. In total, 71% of receivers who received a costly threat stayed out in Experiment 2, compared to 56% of receivers who received a costless threat. The difference is larger compared to Experiment 1, but insignificant (p = 0.16, n

= 96). Based on independent observations in the first round of the signaling game, 50% of receivers who received a costly threat stayed out and exactly 50% of receivers who received a costless threat also stayed out (p = 1.00, n = 32).

As robustness check, I use logit models with controls for risk preference (for Experiments 1 and 2) and round and session fixed-effects (for Experiment 2) to estimate the relationship between receiving a costly threat and staying out. The costly-threat variable is coded 1 if the receiver received a threat with sunk cost, and 0 if not. Table 3 shows the logit estimates compared to a baseline model without controls. Across the models, there is no significant relationship between the costly-threat dummy and the decision to stay out in Experiment 1 ($0.55 \leq p \leq 0.75$) and Experiment 2 ($0.15 \leq p \leq 0.20$). The risk-preference variable is significant in Experiment 1 (p = 0.04) but insignificant in Experiment 2 (p = 0.27).

[**Table 3 here**]

It is important to be careful with our interpretations. We cannot reject the hypothesis that a threat with sunk cost has a similar effect on receiver acquiescence as a threat without. Statistical insignificance does not imply practical insignificance, however, and these results should not be taken as a generic claim that a sunk-cost signal has no credibility effect whatsoever. First, this would be a very strong claim that is hard to substantiate unless the result is successfully replicated with several internally-valid studies conducted across different samples and settings. Moreover, as our experiments explicitly control for learning, it remains an open question as to whether experiments that allow for learning would lead to better play. Finally, although no effect is found on the behavioral measure, the MIT sample (Experiment 2) seems to be more aware of what is the "correct" choice than the national sample (Experiment 1) on the *non-behavioral* retrospective measure (see below). Nevertheless, with these caveats, we can conclude as follows: There is evidence that the sunk-cost signaling mechanism operates well at the signaler's end, but no clear evidence that the same mechanism works well at the receiver's end. The two ends of the signaling mechanism do not operate at equal fidelity to the logic of sinking costs. Signalers and receivers respond asymmetrically to the same logic, exposing a gap in expectations between signalers and receivers.

Retrospective Evaluation of Credibility

Our research question focuses on whether and how a sunk-cost signal shapes the receiver's behavior. As a stress test, however, a non-behavioral measure is also injected into the experiments, based on the receiver's retrospective evaluation after the crisis interaction. After the crisis, all receivers were asked: "Do you think that Player A is a TRUE type?" The receiver responded on a seven-point scale that ranged from "Very unlikely" (0) to "Very likely" (6). At the midpoint was "Neither likely nor unlikely" (3).

In Experiment 1, receivers gave the costly threat an average credibility score of 3.05 compared to 3.01 for the costless threat, with no significant difference between the two scores (two-tailed t-test, $p = 0.94$, $n = 110$). In Experiment 2, however, the costly threat received an average credibility score of 3.82 and the costless threat an average score of 2.71 (two-tailed t-test, $p = 0.0010$, $n = 96$). Restricting the comparison to the first round of Experiment 2, the costly threat received an average score of 3.83, while the costless threat had an average score of 2.81 ($n = 32$, $p = 0.12$). When they were asked to provide a retrospective judgment after the crisis ended, the MIT sample assessed the sunk-cost signal as more credible, whereas the national sample assessed the signals with or without sunk cost as equally credible.

The non-behavioral measure, of course, does not directly test or refute our behavioral expectations stated in the previous section. By eliciting retrospection after the crisis is over and in a non-incentivized environment, the non-behavioral measure is also less targetted and does not satisfy the demands of an economic experimental test under induced value theory

(Smith 1976). It will also be necessary for future studies to tease apart whether the "better" retrospective judgment displayed in Experiment 2 is caused by differences in the subject samples, experimental settings, or response elicitation modes.[18] Despite its limitations, the non-behavioral retrospective evidence cautions us against the strong claim that a sunk-cost signal has no credibility effect whatsoever. This is part of the value of a stress-test design that sets up multiple replication tests. Conducting one experiment with a single sample, setting and outcome measure would have produced unambiguous results that tell a "simple story" (Experiment 1). But is the simple story robust? A stress-test design that seeks to falsify the hypothesis in different ways can complicate the story, but may also allow for a more nuanced understanding that is robust across different samples, settings, and measures.

Sunk-Cost Signaling in a Concrete Context

I further extend the stress-test design with a third experiment. Experiments 1 and 2 were designed in an abstract deterrence setting. Do our conclusions change with a concrete crisis scenario? In addition, Experiments 1 and 2 were implemented on a game-theoretic structure. It is useful to test if consistent conclusions are also obtained in a non-game-theoretic setting, which would suggest a more general phenomenon in human cognition.

Design

Experiment 3 contextualizes sunk-cost signaling in an international crisis scenario free from a game-theoretic structure. The experiment elicits credibility estimates from respondents directly without strategic interaction. Unlike Experiments 1 and 2, there is no signaling game and respondents are not assigned into a specific role. Instead, they are presented with a scenario and asked to assess the credibility of a threat made by a country in the scenario. Experiment 3 was embedded in a time-shared survey conducted over the Internet on the

[18] As discussed earlier, Experiment 2 used a one-stage elicitation of retrospective credibility estimates, as two-stage elicitation might be more cumbersome and less effective with repeated rounds.

Qualtrics platform. 635 U.S. adults were recruited from 6 to 11 April 2012 through AMT. Each participant received $0.51 for completing the survey.

To construct the sunk-cost signal in its purest form, the signal is deliberately decontextualized and disassociated from specific examples in Experiments 1 and 2. However, it is possible that credibility estimates can change when we move from a pure and abstract sunk-cost signal to an impure but concrete real-world example. Experiment 3 is designed as a stress test with a concrete example of a sunk-cost signal (military mobilization) located in a specific crisis context (territorial dispute). While there are few pure cases of sunk-cost signals in international politics, military mobilization is frequently cited as a classic example of a sunk-cost signal (Fearon 1994; 1997), though it is admittedly not a pure case (Fearon 1997, 70).

To increase the rigor of the stress-test, I design two parallel versions of Experiment 3: a simple non-factorial experiment and a richer 2x2 factorial experiment. In both versions, respondents began by reading about a foreign crisis scenario, in which two states had staked their claims on an important piece of territory. One state ("Country X")[19] threatened to fight a war if the other state moved into the territory. Respondents were told that Country X had mobilized its military.

In the simple non-factorial experiment, 210 respondents were randomly divided into control and treatment groups that differed only on one dimension: the treatment group was told that military mobilization was "very costly," while the control group was told that military mobilization was "not very costly." A "high-cost versus low-cost" treatment design is used instead of a "mobilization versus no-mobilization" design. This is to focus the subject's attention on the cost of the signal, as well as control for the within-signal confounds that may arise when the signaling effect of the signal is contaminated by the substantive effect of

[19] Countries are anonymized to prevent the results from being driven by the respondent's thoughts and feelings about specific countries.

the action involved (see Section I). “Very costly” and “not very costly” are used instead of “costly” and “costless,” since the physical deployment of troops cannot plausibly be costless.

Military mobilization as a sunk-cost signal may still be confounded if respondents infer a correlation between the cost of military mobilization and the expected cost of fighting a war. To control for this potential confounder, I design a 2x2 factorial version of Experiment 3 using the remainder of the respondent pool ($n = 425$). By design, this version is exactly the same as the earlier version except for an additional sentence that highlights whether fighting a war at this time “[will be / will not be] very costly to X.” Respondents were randomly assigned into one of four experimental conditions that differed along two dimensions: whether military mobilization was “very costly” or “not very costly” to Country X, and whether a war would be “very costly” or “not very costly” to Country X. The two sentences were presented in random order. Interacting the two possibilities across the two dimensions creates the four experimental conditions. This allows me to isolate the effect of costly mobilization (the sunk-cost signal) on credibility estimates, while explicitly controlling for the effect driven by the expected cost of war.

Results

The dependent variable is measured on a seven-point credibility scale based on the respondent’s perceived likelihood that Country X would fulfill its threat to fight Country Y. The scale ranges from “Very unlikely” (0) to “Very likely” (6), with “Neither likely nor unlikely” (3) at the midpoint. The treatment group in the simple non-factorial experiment gave an average credibility score of 5.00 whereas the control group gave 5.23, with no significant difference between the two (two-tailed t-test, $p = 0.29$, $n = 210$), corroborating our conclusions from Experiments 1 and 2.

Table 4 shows the average credibility scores in the separate 2x2 factorial experiment that controls for the effect driven by the expected cost of war. It is interesting to note that there are significant differences in credibility scores between respondents who saw that the expected cost of war is high for Country X compared to those who saw that the cost of war is low, given high-cost and low-cost military mobilizations (two-tailed t-tests, $p = 0.0032$ with $n = 213$ and $p < 0.0001$ with $n = 212$ respectively). In contrast, there is no significant difference in credibility scores between high-cost and low-cost military mobilizations when the expected cost of war is high for Country X (two-tailed t-test, $p = 0.44$, $n = 211$). Similarly, there is also no significant difference when the cost of war is low (two-tailed t-test, $p = 0.54$, $n = 214$). Taken together, the results from the 2x2 factorial experiment replicate the findings from the simple non-factorial version, and resonate with the general conclusion from Experiments 1 and 2 that the logic of sinking costs does not operate as expected at the receiver's end.

[**Table 4 here**]

Caveats

Given the difficulties in identifying signaling effects with observational data, the experimental method is a particularly useful tool for testing signaling mechanisms. External validity, however, can be an important concern. One limitation is that experiments cannot put human lives at stake, unlike decisions for war and peace. It is useful to note, however, that costly signaling theory does not depend on "high-stake effects," and that the relevant sunk-cost signaling model is rooted in a specific incentive structure that *is* replicated in the experiments.

Another limitation is that without an experimental sample of national leaders caught in the midst of an actual international dispute, we cannot know for sure if leaders in crisis would respond to costly signaling in a similar way. This is an empirical question that cannot be resolved without a series of specific tests. Conceptually, however, one may observe that the theory of costly signaling is not limited to national leaders, but is instead scoped to apply in general to all human actors. And although costly signaling effects specifically among leaders-in-crises remain an open question, the experiments help to shed light on how human individuals generally respond to the logic of costly signaling – which is in itself a non-trivial question. Experimental replication in different settings and with different samples has allowed us to test if the findings are generalizable beyond a particular subject pool. A third experiment also provides contextualization based on an international crisis scenario to enhance the concreteness of the experimental context.

Both internal and external validity are crucial in science, but the external validity of a theory cannot be established without first establishing its internal validity. At the first instance, it is ideal to test a theory with confounders minimized. Experiments can distill the sunk-cost signaling mechanism in its idealized form, to shed light on whether and how the mechanism operates in a ceteris-paribus setting. In the end, external validity can only be established by comparing findings from different internally-valid studies conducted across different contexts. To this end, the results here provide only an initial baseline for future studies, which may reinforce or challenge the findings that emerged from our specific settings and samples.

CONCLUSION

Costly signaling is central to international politics in an asymmetric-information world, but its mechanisms are difficult to test with observational data. The difficulty arises from the twin problems of observability and confounding. The signaler's true resolve and the receiver's credibility belief are private information that cannot be easily observed. The effect of a sunk-cost signal is also almost always confounded with the effects of other previous or simultaneous signals and information, at least some of which cannot be known to the researcher.

This paper presents a series of controlled experiments to overcome these problems. Observability is achieved with random assignment of resolve and direct measurement of credibility, while confounding is minimized with signaling experiments conducted in *tabula-rasa* and ceteris-paribus settings. The purpose is to provide a confound-cleansed baseline for our theoretical understanding, and for future investigations of the mechanism in different contexts, by studying how sunk-cost signaling operates under ideal conditions. A novel feature is to impose a rigorous stress-test design with three different experiments, testing the same mechanism in different ways, on different samples, and with different measures.

Three main conclusions are extracted. First, signalers randomly assigned with high resolve are much more likely to sink costs or "burn money." This result suggests that the sunk-cost signaling mechanism operates well at the signaler's end. Second, while there is evidence that the mechanism works well at the signaler's end, there is no clear evidence that the same mechanism also works well at the receiver's end. As emphasized earlier, this should not be taken as a generic conclusion that a sunk-cost signal has no credibility effect whatsoever. What it suggests is that the sunk-cost signaling mechanism has a strong and clear effect at the signaler's end, but a relatively weak and unclear effect at the receiver's end. The signaler component of the mechanism is validated and replicated in the experiments, but not the receiver component of the same mechanism. These results underscore the importance of simultaneously measuring responses on both sides of the signaler-receiver relationship. Finally, signalers and receivers do not appear to respond to the logic of sinking costs in the same way. This suggests a sender-receiver gap in how the same deterrence interaction is perceived at the two ends of the signaling mechanism, contrary to what the theory of costly signaling automatically assumes.

It is useful to note that it is generally unsurprising to find human players deviating from the equilibrium predictions of a game-theoretic model. Such deviations are common in experimental economics, where in many cases even probabilistic directional expectations (e.g. that most people should behave in the predicted way rather than its opposite) are contradicted by the observed behavior, not to mention deterministic point predictions (e.g. that 100% will behave according to the model). What is surprising here is that those randomly assigned to the role of the signaler behaved in a way consistent with the logic of sunk-cost signaling, whereas those in the role of the receiver behaved in a way that contradicted the same logic.[20]

The unexpected results have interesting implications. First, they uncover an anomaly that is more difficult to dismiss with the standard alternative explanations. It is typically easy to explain away a null finding in economic experiments by speculating that subjects might not have understood the game; were given limited opportunity to learn the "correct" strategic behavior; or were simply unmotivated due to the insufficient stakes. These standard speculations, however, are less satisfying here because they cannot explain the anomaly uncovered over repeated experiments. Signalers and receivers were randomly assigned; they played the same game; they operated with similar knowledge of the possible payoffs; and both made only one simple binary decision in the game. Yet why is it that across separate experiments the behavior of the signaler matched the costly signaling logic, but not the behavior of the receiver? This is precisely the puzzle leading to the second implication that needs to be addressed in future research. What can explain why those in the role of the signaler can see and behave according to the costly signaling logic, whereas those in the role of the receiver do not see and cannot behave according to the same logic?

The third implication touches more broadly on the classic rationalist/non-rationalist debate in the deterrence literature.[21] The results suggest that the behavioral reality may be more nuanced than the polar positions defined in the debate, resting somewhere between the two extremes. The conclusion that "[rational deterrence] is inadequate as an explanatory theory

[20] I thank Dave Clark for highlighting this point.

[21] While there appears to be no systematic discussion in the literature specifically on the failures of sunk-cost signaling, there is much discussion on deterrence failure in international relations, which relates to signaling. Of particular importance is the rational deterrence debate published in a special issue of *World Politics* in 1989 (Vol. 41, No. 2).

. . . [because] neither leaders contemplating challenges nor leaders seeking to prevent them necessarily act as the theory predicts" (Lebow 1985, 203) may be too loose. Yet the general point on misperception at the receiver's end is well-taken, that "commitments by one actor that are objectively clear and credible (as measured by the perceptions of disinterested third parties) may not be perceived by another" (Jervis 1989, 198). This general point resonates beyond sunk-cost signaling, and may be relevant to other forms of signaling such as tied-hands signals and audience costs. At the same time, the systematic asymmetry in signaler-receiver behaviors suggests there may be something deeper and more structural in the causes of signaling failure, beyond the "ideological, political, or cultural differences in cognitive contexts" indicated in Stein's (1988, 251) review.

The fourth implication is how the anomaly detected in the experiments resonates at the international level, suggesting that it may not simply be an individual-level phenomenon or artefact of the lab. The experimental results have an uncanny connection to the real-world puzzle of why sunk-cost signaling is frequently used despite its costs and inefficiencies. Particularly, between 1950 and 2000, great powers had deployed their nuclear assets to more than 20 countries to enhance the extended deterrence of their alliances, yet a recent study shows that these deployments have surprisingly little deterrent effect: stationing nuclear weapons on allied territory does not affect the likelihood that the ally would be targeted in a militarized dispute (Fuhrmann and Sechser 2014). Great-power states believed that sinking costs would help to communicate the credibility of their resolve to potential challengers of their allies, but in fact it did not. This observational finding clicks with the central results from our experiments: the sunk-cost signaling mechanism exerts a strong effect at the signaler's end – hence signalers willingly sink costs to demonstrate credibility. But the mechanism is weaker at the receiver's end – thus receivers do not assign equivalent credibility to the signal despite the costs and risks incurred by the signaler. Sinking costs may be a poor way to sharpen credibility, yet signalers sink costs nonetheless. Consequently, we observe large-scale deployments of sunk costs to demonstrate credibility in international politics, by leaders both in history and in contemporary times,[22] despite the risks and inefficiencies involved.

Several questions open from this study for future research. First, this study is specifically scoped in the context of deterrence. Future studies may find it very interesting to investigate how sunk-cost signaling operates in the reversed context of reassurance. Several theoretical hints suggest that threat-based signaling differs significantly from promise-based signaling.[23] One particularly intriguing hypothesis is whether the *removal* or discontinuation of the sunk costs – insofar as it is costly in itself – can credibly reveal the signaler's resolve to cooperate. Even if sinking costs does not credibly communicate the willingness to fight, can "un-sinking" the sunk costs communicate the willingness to cooperate? For instance, does the removal of costly military or nuclear installations abroad or the retrenchment of costly alliance structures signal effectively one's intent for peace? Some historical examples suggest that it might, as shown by the Soviet Union's unilateral reduction of 500,000 troops in central Europe at the end of the Cold War (Kydd 2000), and Egyptian President Anwar Sadat's politically costly visit to Jerusalem in 1977 that eventually culminated into the Egypt-Israel Peace Treaty (Stein 1991).

Second, do sunk costs at the domestic level differ from those at the international level? Discussions of sunk costs have largely focused on the international level with the assumption of rational unitary states. Yet it is plausible – through a backward inference from collective action theory (Olson 1965) – that irrecoverable domestic political costs individually borne by a leader can convey a seriousness that differs from that conveyed through the sunk costs

[22] See Slantchev (2011, 67-75).

[23] These hints derive from both psychological and rationalist sources. Prospect theory, for instance, suggests that leaders facing a threat operate in the loss-frame whereas those facing a promise operate in the gain-frame (see Davis 2000; Kahneman and Tversky 1979). In the rationalist literature, Glaser (1994) and Kydd (2000) argue that costly signals can reduce mistrust. Research on tied-hands signaling also highlights how "Type II audience costs" (the domestic political losses suffered by a leader for entering into a conflict after promising to stay out) generate implications that differ from that of threat-based audience costs (Quek 2012).

collectively borne by the state. If this is the case, then it is not so much the Soviet costs of international retreat or the Egyptian costs of regional defection that convinced their foes, but the irrecoverable political costs and risks incurred by Gorbachev and Sadat that moved the skeptics' credibility estimates. Future research should fill the gaps in our understanding of sunk-cost signaling in its different forms.

Last but not least is the puzzle raised at the start. The experiments found that the logic of sinking costs is more straightforward and distinct to the signaler than it is to the receiver. Why this is the case is an interesting question. One possibility is "projection bias" – the psychological tendency to overestimate the degree to which one's own characteristics, beliefs, or perceptions are shared by others (Allport 1924; Holmes 1968; Ross, Greene and House 1977; Ashenfelter and Krueger 1994; Hogset and Barrett 2010). In signaling, projection bias may operate with signalers projecting their own beliefs and perceptions on receivers, and vice versa. This projection bias can be further aggravated by empathy-deficiency in mutual perception (Fiske and Taylor 1984; Jervis 1985; Stein 1988). Citing examples from international history, Stein (1988, 250) argued that the "inability to empathize intrudes at both ends of the signaling process to confound the perception of threat. The leaders who issue the threat are frequently insufficiently sensitive to the way their adversary sees them ... [and the] target interprets the threat from a different cognitive context and deduces meaning that is unintended."[24] Empathy deficiency in conjunction with projection bias may offer one plausible explanation for why the logic of sinking costs fails.[25]

[24] An interesting question for future research is whether and how we can improve the sunk-cost signaling mechanism. Are there refinements or "twists" to the mechanism that can improve its efficacy? For instance, does sunk-cost signaling work better if we use alternative or unconventional ways to trigger, enhance, or block type-separation reasoning? Along this line, I attempted a preliminary exploration that arbitrarily blocked all information to receivers about the payoffs for the two types of signalers, except that signalers would get a higher payoff if the receiver stayed out. The outcomes seemed to suggest that arbitrarily blocking type-separation reasoning could improve or at least not worsen performance, though the effects are statistically ambiguous. More directly, however, strategies that weaken the projection bias and empathy-deficiency may be particularly promising to investigate in future research.

[25] It should be highlighted that there is a long menu of cognitive biases in the psychological literature (Kunda 1999), and not all of them are necessarily supportive of the null expectations. For instance, the "proportionality bias" (Komorita 1973; Jervis 1985) appears to contradict the null expectations – insofar as

Another plausible explanation is the existence of a biased asymmetric-information environment where the receiver always knows less than the signaler. The signaler's type is known to the signaler but opaque to the receiver. For the receiver, the signaler's type is a mystery and deception always a possibility. There is thus greater uncertainty at the receiver's end than at the signaler's end, resulting in greater cognitive complexity for the receiver than for the signaler. While beyond the scope of this paper, it would be interesting to further replicate the experiments with extremely high stakes to see if they can motivate receivers to overcome the cognitive load and improve their judgment. The replication is interesting because some experimental literature suggests the reverse of this intuitive expectation: experiments have shown that very large stakes increase mistakes rather than decrease them (Ariely, Gneezy, Loewenstein and Mazar 2009). It appears that while large stakes can boost the incentive for subjects to *try* to behave more rationally, they can also shake the stability of the subjects' cognitive judgment. More broadly, if very large stakes, high stress and environmental complexity impair rather than improve judgment, then the effects uncovered in this study may well be an *under*-estimate of what we should expect in international politics, given the larger stakes, higher stress and greater complexity involved.

In the experiments here, many sources of noise that are likely in real life are shut down by design: the type-separation logic is made unusually transparent; reputational concerns and the shadow of the future are removed; and uncertainty over the signaler's reaction to the receiver's choice is by design eliminated. The results suggest, however, that the receiver may not be able to respond optimally to the sunk-cost signal even in an idealized environment. This points to a potential tragedy in costly signaling: the signalers choose to suffer the sunk cost because they assume it makes their signal more credible, but the receivers do not

it leads leaders to "make inferences about the intentions of others from the costs and consequences of the actions they initiate, [such that] [w]hen a state incurs high costs, others assume that important objectives were at stake for the leadership" (Stein 1988, 254). As Stein (1988, 264) admits, despite a relatively well-established set of cognitive biases, their "relationship to one another and to substantive and situational factors remains as yet unexplicated." With this caveat, it remains useful to hypothesize the interactions of cognitive biases that relate most closely to signaling failure, and which best support the null expectations.

respond in line with the signalers' logic, despite the sunk cost suffered. The consequences are wasted resources and a suboptimal outcome for everyone.

REFERENCES

Allport, Floyd Henry. 1924. *Social Psychology.* Cambridge, MA: Riverside Press.

Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar. 2009. "Large Stakes and Big Mistakes." *Review of Economic Studies* 76: 451-69.

Ashenfelter, Orley, and Alan Krueger. 1994. "Estimates of the Economic Return to Schooling from a New Sample of Twins." *American Economic Review* 84: 1157-73.

Banks, Jeffrey S., Colin Camerer, and David Porter. 1994. "An Experimental Analysis of Nash Refinements in Signaling Games." *Games and Economic Behavior* 6: 1-31.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20: 351-68.

Blainey, Geoffrey. 1988. *The Causes of War.* New York: Free Press.

Brandts, Jordi, and Charles A. Holt. 1992. "An Experimental Test of Equilibrium Dominance in Signaling Games." *American Economic Review* 82: 1350-65.

Cho, In-Koo, and David M. Kreps. 1987. "Signaling Games and Stable Equilibria." *Quarterly Journal of Economics* 102: 179-221.

Costa-Gomes, Miguel A., and Vincent P. Crawford. 2006. "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study." *American Economic Review* 96: 1737-68.

Cooper, David, Susan Garvin, and John Kagel. 1997. "Signaling and Adaptive Learning in an Entry Limit Pricing Game." *RAND Journal of Economics* 28: 662-83.

Cooper, David, and John Kagel. 2005. "Are Two Heads Better than One? Team versus Individual Play in Signaling Games." *American Economic Review* 95: 477-509.

Crawford, Vincent P., and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50: 1431-51.

Davis, James W. 2000. *Threats and Promises: The Pursuit of International Influence.* Baltimore, MD: Johns Hopkins University Press.

Farrell, Joseph. 1987. "Cheap Talk, Coordination, and Entry." *RAND Journal of Economics* 18: 34-9.

Fearon, James. 1994. "Domestic Political Audiences and the Escalation of International Disputes." *American Political Science Review* 88: 577-92.

Fearon, James. 1997. "Signaling Foreign Policy Interests: Tying Hands Versus Sinking Costs." *Journal of Conflict Resolution* 41: 68-90.

Filson, Darren, and Suzanne Werner. 2002. "A Bargaining Model of War and Peace: Anticipating the Onset, Duration, and Outcome of War." *American Journal of Political Science* 46: 819-37.

Fiske, Susan, and Shelley Taylor. 1984. *Social Cognition.* Reading, MA: Addison-Wesley.

Fuhrmann, Matthew, and Todd S. Sechser. 2014. "Signaling Alliance Commitments: Hand-Tying and Sunk Costs in Extended Nuclear Deterrence." *American Journal of Political Science* 58: 919-35.

Gartzke, Erik, Quan Li and Charles Boehmer. 2001. "Investing in Peace: Economic Interdependence and International Conflict." *International Organization* 55: 391-38.

Glaser, Charles L. 1994. "Realists as Optimists: Cooperation as Self-Help." *International Security* 19: 50-90.

Hogset, Heidi, and Christopher B. Barrett. 2010. "Social Learning, Social Influence, and Projection Bias: A Caution on Inferences Based on Proxy Reporting of Peer Behavior." *Economic Development and Cultural Change* 58: 563-89.

Holmes, David S. 1968. "Dimensions of Projection." *Psychological Bulletin* 69: 248-68.

Horton, John J., David G. Rand, and Richard J. Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14: 399-425.

House Committee on Armed Services, United States Senate. 2013. Statement of Donald L. Cook, Deputy Administrator for Defense Programs, National Nuclear Security Administration, U.S. Department of Energy, On the B61 Life Extension Program and Future Stockpile Strategy, October 29, 2013.

Huber, Gregory A., Seth J. Hill, and Gabriel S. Lenz. 2012. "Sources of Bias in Retrospective Decision Making: Experimental Evidence on Voters' Limitations in Controlling Incumbents." *American Political Science Review* 106: 720-41.

Jervis, Robert. 1970. *The Logic of Images in International Relations.* Princeton, NJ: Princeton University Press.

Jervis, Robert. 1985. "Perceiving and Coping with Threat." In *Psychology and Deterrence*, ed. Robert Jervis, Richard Ned Lebow, and Jane G. Stein. Baltimore, MD: Johns Hopkins University Press.

Jervis, Robert. 1989. "Rational Deterrence: Theory and Evidence." *World Politics* 41: 183-207.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47: 263-91.

Knops, Raymond. 2010. "U.S. Non-Strategic Nuclear Weapons in Europe: A Fundamental NATO Debate." NATO Parliamentary Assembly Committee Report for the 2010 Annual Session in Warsaw.

Komorita, Stuart S. 1973. "Concession-Making and Conflict Resolution." *Journal of Conflict Resolution* 17: 745-62.

Kristensen, Hans M. 2005. *U.S. Nuclear Weapons in Europe: A Review of Post-Cold War Policy, Force Levels, and War Planning.* Washington, DC: Natural Resources Defense Council.

Kunda, Ziva. 1999. *Social Cognition: Making Sense of People.* Cambridge, MA: MIT Press.

Kydd, Andrew. 2000. "Trust, Reassurance and Cooperation." *International Organization* 54: 325-57.

Kydd, Andrew. 2005. *Trust and Mistrust in International Relations.* Princeton, NJ: Princeton University Press.

Lebow, Richard Ned. 1985. "The Deterrence Deadlock." In *Psychology and Deterrence*, ed. Robert Jervis, Richard Ned Lebow, and Jane G. Stein. Baltimore, MD: Johns Hopkins University Press.

Lebow, Richard Ned. 1985. "Conclusions." In *Psychology and Deterrence*, ed. Robert Jervis, Richard Ned Lebow, and Jane G. Stein. Baltimore, MD: Johns Hopkins University Press.

Lebow, Richard Ned. 1987. "Deterrence Failures Revisited." *International Security* 12: 197-213.

Lektzian, David J., and Christopher M. Sprecher. 2007. "Sanctions, Signals, and Militarized Conflict." *American Journal of Political Science* 51: 415-31.

Miller, Ross M., and Charles R. Plott. 1985. “Product Quality Signaling in Experimental Markets.” *Econometrica* 53: 837-72.

Morton, Rebecca B., and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab.* New York: Cambridge University Press.

Morrow, James D. 1989. “Capabilities, Uncertainty, and Resolve: A Limited Information Model of Crisis Bargaining.” *American Journal of Political Science* 33: 941-72.

Morrow, James D. 1994. “Alliances, Credibility, and Peacetime Costs.” *Journal of Conflict Resolution* 38: 270-97

Morrow, James D. 2000. “Alliances: Why Write Them Down?“ *Annual Review of Political Science* 3: 63-83.

Nalebuff, Barry. 1991. “Deterrence in an Imperfect World.” *World Politics* 43: 313-35.

O’Neill, Barry. 1990. “The Intermediate Nuclear Force Missiles: An Analysis of Coupling and Reassurance.” *International Interactions* 15: 345-63.

Olson, Mancur. 1965. *The Logic of Collective Action.* Cambridge, MA: Harvard University Press.

Potters, Jan, and Frans Van Winden. 1996. “Comparative Statics of a Signaling Game: An Experimental Study.” *International Journal of Game Theory* 25: 329-53.

Powell, Robert. 1987. “Crisis Bargaining, Escalation, and MAD.” *American Political Science Review* 81: 717-36.

Quek, Kai. 2012. “Do Domestic Audience Costs Really Exist?” Paper for the Annual Meeting of the American Political Science Association, New Orleans. http://ssrn.com/abstract=2107187 (accessed January 13, 2016).

Ross, Lee, David Greene, and Pamela House. 1977. "The False Consensus Phenomenon: An Attributional Bias in Self-Perception and Social Perception Processes." *Journal of Experimental Social Psychology* 13 (3): 279-301.

Sagan, Scott D. 2003. "More Will Be Worse." In *The Spread of Nuclear Weapons: A Debate Renewed*, ed. Scott D. Sagan and Kenneth N. Waltz. New York: Norton.

Schelling, Thomas. 1960. *The Strategy of Conflict.* Cambridge, MA: Harvard University Press.

Schelling, Thomas. 1966. *Arms and Influence.* New Haven, CT: Yale University Press.

Slantchev, Branislav L. 2005. "Military Coercion in Interstate Crises." *American Political Science Review* 99: 533-47.

Slantchev, Branislav L. 2011. *Military Threats: The Costs of Coercion and the Price of Peace.* New York: Cambridge University Press.

Smith, Vernon L. 1976. "Experimental Economics: Induced Value Theory," *American Economic Review* 66: 274-79.

Spence, Michael. 1973. "Job Market Signaling." *Quarterly Journal of Economics* 87: 355-74.

Stein, Janice G. 1988. "Building Politics into Psychology: The Misperception of Threat." *Political Psychology* 9: 245-71.

Stein, Janice G. 1991. "Reassurance in International Conflict Management." *Political Science Quarterly* 106: 431-51.

Tingley, Dustin H., and Barbara F. Walter. 2011. "Can Cheap Talk Deter? An Experimental Analysis." *Journal of Conflict Resolution* 55: 996-1020.

Figure 1: Sunk-Cost Signaling Game

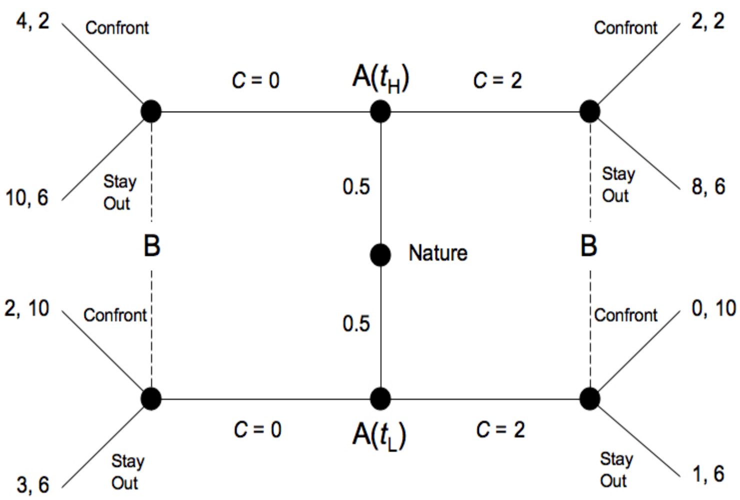

Figure 2: Proportion Sending Costly Signal (Experiments 1 and 2)

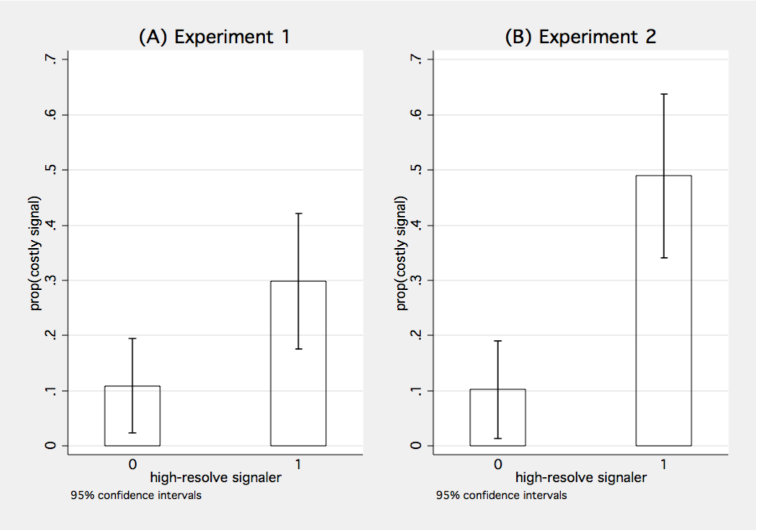

Figure 3: Proportion Staying Out (Experiments 1 and 2)

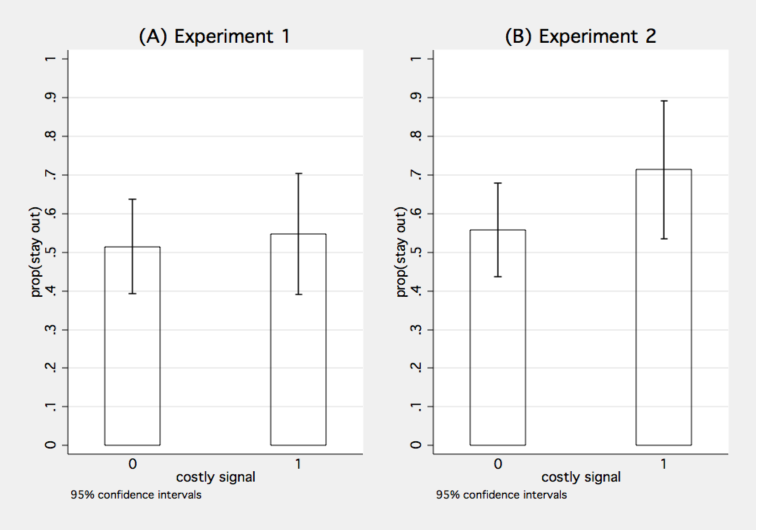

Table 1: Summary of Experiments

	Experiment 1	Experiment 2	Experiment 3
Type	Internet-based game	Laboratory game	Internet-based survey
Activity	Signaling game	Signaling game	Credibility rating
Incentivized	Yes	Yes	No
Sample	U.S. adults / AMT	MIT students	U.S. adults / AMT
Rounds	One-shot	Three rounds	One-shot
Subjects	253	64	635

Table 2: Logit Estimates – Determinants of Costly Threat

	Experiment 1		Experiment 2	
	(1)	(2)	(3)	(4)
High-resolve	1.244	1.242	2.138	2.142
	(0.523)*	(0.529)*	(0.613)**	(0.616)**
Risk-preference		0.183		0.012
		(0.152)		(0.353)
Constant	-2.100	-2.630	-2.676	-2.668
	(0.434)**	(0.671)**	(0.661)**	(1.198)*
Round & session FE	No	No	Yes	Yes
Log-likelihood	-53.688	-52.974	-46.316	-45.928
Prob>Chi2	0.017	0.044	0.002	0.005
Pseudo-R2	0.056	0.069	0.201	0.198
N	112	112	96	94

Notes: ** $p< 0.01$; * $p< 0.05$. In parentheses are robust standard errors, which are corrected for clustering at the subject level for Experiment 2. Round and session dummies are used to control for round and session fixed effects.

Table 3: Logit Estimates – Determinants of Staying Out

	Experiment 1		Experiment 2	
	(1)	(2)	(3)	(4)
Costly threat	0.132	0.229	0.694	0.648
	(0.395)	(0.414)	(0.476)	(0.481)
Risk-preference		-0.323		0.269
		(0.157)*		(0.245)
Constant	0.059	1.050	0.932	0.303
	(0.244)	(0.519)*	(0.455)*	(0.648)
Round & session FE	No	No	Yes	Yes
Log-likelihood	-76.026	-73.266	-58.786	-57.750
Prob>Chi2	0.738	0.119	0.084	0.156
Pseudo-R2	0.001	0.037	0.088	0.097
N	110	110	96	95

Notes: ** $p < 0.01$; * $p < 0.05$. In parentheses are robust standard errors (corrected for clustering at subject level for Experiment 2).

Table 4: Credibility Scores (Experiment 3)

		Sunk Cost	
		High	**Low**
Cost of War	**High**	4.72 (1.85)	4.51 (1.97)
	Low	5.37 (1.34)	5.48 (1.13)

Note: Standard deviations in parentheses