Models of Escalation and Reputation Concerns

Allan Dafoe and Devin Caughey

10/26/2015

This document contains the analyses of Models of Escalation and Reputation Concerns. These are referenced in the paper: Dafoe, Allan and Devin Caughey. 2016. "Honor and War: Southern U.S. Presidents and the Effects of Concern for Reputation." *World Politics.* 68(2). The paper, and associated files, are available at www.allandafoe.com/southernhonor.

Contents

A Summary and Analysis of the Formal Models **2**
A.1 Summary of the Models . . . 2
A.2 Equilibrium Strategies and Beliefs . . . 3
A.3 Analysis of the Model . . . 4
A.3.1 Round 3 (Final Round) . . . 4
A.3.2 Round 2 . . . 5
A.3.3 Round 1 . . . 5
A.4 Limiting the Parameter Space Under Consideration . . . 5
A.5 Only C Indifference Point Interior? (Denoted Situation C) . . . 6
A.6 Solving the Model under Situation C . . . 6
A.6.1 Probability P_1 Escalates in First Round (p_1) . . . 6
A.6.2 Probability of P_2 Escalating (q_2) . . . 6
A.6.3 Probability of P_1 Escalating in Final Round (p_3) . . . 6
A.7 Existence and Uniqueness of Equilibrium . . . 7
A.8 Private Honor . . . 7
A.8.1 Model F with Private Honor . . . 7
A.8.2 Model R with Private Honor . . . 9
A.8.3 Conclusion . . . 9
A.9 Model F and R with Observable Honor . . . 9
A.9.1 Probability of P1 Escalating in Last Round (p_1) . . . 10
A.9.2 Probability of P2 Escalating (q_2) . . . 10
A.9.3 Probability of P1 Escalating in Final Round (p_3) . . . 10
A.9.4 Comparative Statics for Models with Observable Honor . . . 10
A.10 Signing $\frac{dp_1}{dR_1}$ for Specific Distributions . . . 12
A.10.1 Uniformly Distributed Types . . . 12
A.10.2 Logistically Distributed Types . . . 13
A.11 Summary of Comparative Statics . . . 14
A.12 Empirical Implications of the Model . . . 15
A.13 Numerical Examination of Hypotheses . . . 16
A.13.1 Duration in Model F with Observable Honor: $\frac{d\phi}{dR_1}$ . . . 17
A.13.2 Victory in Model R with Observable Honor: $\frac{w_R}{dR_2}$ . . . 17

A Summary and Analysis of the Formal Models

A.1 Summary of the Models

We develop and solve four members of a family of models. These four models all build from the same basic game, but vary on key assumptions. We examine this family of models, rather than an individual model, to ensure that our predictions are robust to modification of certain key assumptions. The first two models, referred to as models *with private honor*, assume that the difference in honor between Southerners and non-Southerners is private information; these models are consistent with the view that others do not realize that there is a systematic difference in concern for reputation between Southerners and non-Southerners. The second two models, referred to as models *with observable honor*, assume that the difference in honor between Southerners and non-Southerners is observed and commonly known.

Within each of these groups, we consider a model in which the focal-agent (denoted **Model F** for focal-agent) and a model in which the rival-agent (denoted **Model R** for rival-agent) makes the first move. The actor who makes the first move (denoted P_1) is the actor who decides whether to carry out the action that engages the reputation of both parties. We assume that current theory and datasets (such as the dataset of Militarized Interstate Disputes) are unable to empirically identify which country is responsible for the reputation-engaging event. Therefore, we look for comparative statics that are not sensitive to which actor engages the reputation of both parties.[1]

Two agents, P_1 and P_2, both want to resolve some issue, commonly valued at $v > 1$, in their favor (e.g., claim territory, restructure the interstate system, renegotiate their strategic relationship). The game is three rounds long, as illustrated in Figure A.1. P_1 first decides whether to escalate the conflict ($\omega_1^1 = E_1$, where ω_i^j denotes the strategy of player i in round j), such as through a verbal commitment or threat, or to concede the issue ($\omega_1^1 = C_1$). If P_1 escalates then both parties' reputations become engaged to the dispute. P_2 then similarly decides whether to concede ($\omega_2^2 = C_2$) or to further escalate the issue through a use of force ($\omega_2^2 = E_2$). If P_2 concedes, then neither side pays any material cost of conflict, but P_2 pays a reputational cost $R_2 + r_2$, where $R_2 \geq 0$, $r_2 \geq 0$. If P_2 uses force, P_1 pays an immediate cost of c_1/k, where $k > 1$. Following escalation by P_2, P_1 decides whether to concede ($\omega_1^3 = C_3$), and absorb the reputational cost of $R_1 + r_1$ ($R_1 \geq 0, r_1 \geq 0$), or escalate one last time by using force ($\omega_1^3 = E_3$). If P_1 uses force, this portion of the game ends and the issue in dispute is resolved by some contest function that yields the prize to P_1 with probability π and P_2 with probability $(1 - \pi)$, P_1 pays an additional cost of c_1, and P_2 pays a cost of c_2. The material costs of conflict c_1 and c_2 are private information drawn from continuous cumulative distribution functions $F_1(x)$ and $F_2(x)$, where $\forall \epsilon > 0, i \in \{1, 2\} : F_i(\epsilon) > 0$. These have corresponding density functions $f_1(x)$ and $f_2(x)$, defined so that their support is a convex set.[2] A component of the reputational cost is common knowledge (R_1 and R_2), and a component is private information (r_1 and r_2) drawn from continuous cumulative distribution functions $G_1(x)$ and $G_2(x)$ with density functions $g_1(x)$ and $g_2(x)$. In models with observable honor, we set $r_1 = r_2 = 0$ and examine comparative statics with respect to observable honor: R_1 or R_2. In models with private honor, we consider comparative statics with respect to private honor: r_1 or r_2. All other aspects of the game are common knowledge.

Let p_1, q_2, and p_3 denote the probability that the relevant agent escalates in Round 1, 2, or 3, respectively, in equilibrium from the perspective of an observer who is unable to see the player's

[1] Other models are possible in which reputation-engagement not dichotomous and symmetric.
[2] That is, $\forall y, x, z$, $i \in \{1, 2\}$, where $y < x < z$, $f_i(y) > 0$, and $f_i(z) > 0$, then $f_i(x) > 0$

private costs (c_1 or c_2) but can see the player's private reputational concern (r_1 or r_2). These parameters capture the key observable implications of the model for an analyst who can detect differences in r, as we presume to do. Let $\hat{p}_3$ denote P_2's beliefs about the probability that P_1 will escalate in Round 3 (and similarly for $\hat{q}_2$). Note that, from Assumption 1, bargaining behavior prior to the dispute becoming militarized occurs in Round 1. It is in this sense that we model, and hence account for, selection into a MID. Since extant datasets do not provide systematic data on bargaining behavior before disputes have become militarized, predictions on p_1 can not be tested on the MID or other datasets.

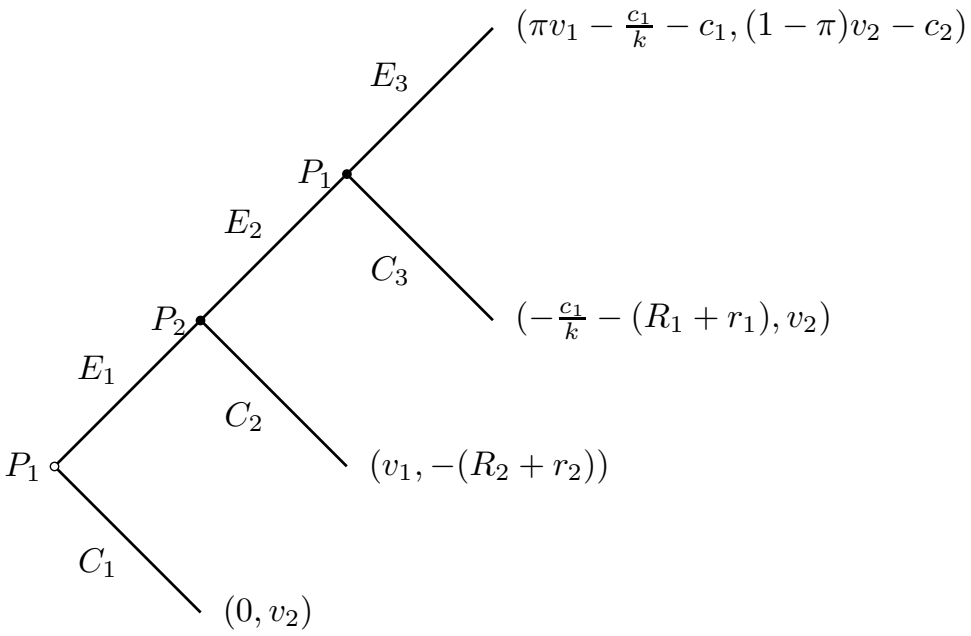

Figure 1: Sequential 3-Move Escalation Game.
E_1 denotes a threat or commitment, E_2 and E_3 a use of force, C_i a concession on the issue in dispute in Round i. $p_1 = Pr(\omega_1^1 = E_1|r_1)$; $q_2 = Pr(\omega_2^2 = E_2|r_2)$; $p_3 = Pr(\omega_1^3 = E_3|r_1)$

A.2 Equilibrium Strategies and Beliefs

We restrict our attention to portions of the parameter space that result in "interior" equilibrium solutions, which best reflect the logic of crisis bargaining.[3] Given these conditions, in any perfect

[3]Specifically, we impose the following two conditions:
Condition C: $\hat{q}_2 < q_2^C \equiv \frac{vk}{(R_1+r_1)(1+k)+v(p+k)}$. Condition C guarantees that we only consider conflicts in which there is some probability that P_1 will back down in Round 3. In a model with rational expectations, $\hat{q}_2 = q_2^*$, therefore Condition C requires $q_2^* < q_2^C$ (where q_2^* denotes the equilibrium level of q_2 from the perspective of an observer who doesn't see r_2).
Condition I: The upper bound on the support of the distribution for c_1 is greater than $c_1^C \equiv \frac{k(v(1-\hat{q}_2)-(R_1+r_1)\hat{q}_2)}{\hat{q}_2}$. Condition I ensures that some P_1 types will not escalate in the first round.

Bayesian equilibrium the strategies of P_1 and P_2 can be implicitly[4] defined as:

$$\omega_1 = \begin{cases} \{C_1, \cdot\} & \text{if } c_1 > c_1^C \equiv \frac{k(v(1-\hat{q}_2)-(R_1+r_1)\hat{q}_2)}{\hat{q}_2} \\ \{E_1, C_3\} & \text{if } \pi v + R_1 + r_1 < c_1 < c_1^C \\ \{E_1, E_3\} & \text{if } c_1 < \pi v + R_1 + r_1 \end{cases} \quad \text{and} \quad \omega_2 = \begin{cases} \{C_2\} & \text{if } c_2 > \frac{v(1-\pi\hat{p}_3)+(R_2+r_2)}{\hat{p}_3} \\ \{E_2\} & \text{otherwise} \end{cases}$$

It follows that

$$\begin{aligned} p_1 &= Pr(c_1 < c_1^C(r_1)|r_1) = Pr\left(c_1 < \frac{k(v(1-\hat{q}_2)-(R_1+r_1)\hat{q}_2)}{\hat{q}_2}\Big| r_1\right) \\ q_2 &= Pr\left(c_2 < \frac{v(1-\pi\hat{p}_3)+(R_2+r_2)}{\hat{p}_3}\Big| r_2\right) \\ p_3 &= \frac{Pr(c_1-(R_1+r_1) < \pi v|r_1)}{Pr(c_1+(R_1+r_1) < k(v(1-\hat{q}_2))/\hat{q}_2|r_1)} \end{aligned}$$

Given rational expectations, $\hat{q}_2 = \bar{q}_2(\bar{p}_3)$, and $\hat{p}_3 = \bar{p}_3(\bar{q}_2)$, where $\bar{q}_2$ and $\bar{p}_3$ denote the expected probability of escalation for a given set of beliefs from the perspective of an observer who cannot see the private information of the other player: r_i and c_i. For example, $\bar{p}_3(\bar{q}_2) = \int_0^\infty p_3(r_1, \bar{q}_2)g_1(r_1)dr_1$. In Round 1, both players' beliefs about the other player are those given by the structure of the game. In Round 2, P_2 believes that c_1 and r_1 are drawn from their joint density function truncated above when $c_1 + kr_1 > \bar{c}_1^C \equiv c_1^C|r_1 = 0$.[5] Since $\bar{q}_2(\bar{p}_3)$ and $\bar{p}_3(\bar{q}_2)$ cross each other once and only once, there is a unique equilibrium.

A.3 Analysis of the Model

This section works through the model to demonstrate that the above strategies and beliefs represent the unique perfect Bayesian equilibrium.

A.3.1 Round 3 (Final Round)

P_1's best response function is:

$$\omega_1^3 = \begin{cases} E_3 & \text{if } c_1 - (R_1 + r_1) < \pi v_1 \\ C_3 & \text{if } c_1 - (R_1 + r_1) \geq \pi v_1 \end{cases} \tag{1}$$

Conditional on reaching round 3, P_1 escalates with probability:

$$p_3 = Pr(c_1 < \pi v_1 + R_1 + r_1 | r_1, \omega_1^1 = E_1)$$

This probability statement is written conditional on r_1 to indicate that p_3 is conditional on the realization of r_1.

[4] It was not possible to solve this model for general distributions explicitly for the exogenous parameters. Other modelers have also relied on implicit solutions, when necessary, for finding comparative statics (e.g. Werner, 2000; Slantchev, 2011).

[5] Denote P_2's beliefs about c_1 and r_1 in Round 2 by the joint density function $h_1^2(c_1, r_1)$:

$$h_1^2(c_1, r_1) = \begin{cases} \frac{f_1(c_1)g_1(r_1)}{\int_0^{\bar{c}_1^C}\int_0^{(\bar{c}_1^C - c_1)/k} f_1(c_1)g_1(r_1)dr_1dc_1} & \text{if } c_1 + kr_1 \leq \bar{c}_1^C \\ 0 & \text{otherwise.} \end{cases}$$

A.3.2 Round 2

P_2's best response function is:

$$\omega_2^2 = \begin{cases} E_2 & \text{if } \hat{p}_3((1-\pi)v_2 - c_2) + (1-\hat{p}_3)v_2 > -(R_2 + r_2) \\ C_3 & \text{if } \hat{p}_3((1-\pi)v_2 - c_2) + (1-\hat{p}_3)v_2 \leq -(R_2 + r_2) \end{cases} \tag{2}$$

P_2 therefore escalates in Round 2 with probability:

$$q_2 = Pr(\omega_2^2 = E_2) = Pr\left(c_2 < \frac{v_2(1-\pi\hat{p}_3) + (R_2 + r_2)}{\hat{p}_3} \Big| r_2\right)$$

A.3.3 Round 1

$EU_1(C_1) = 0$ (since no reputation costs)
$EU_1(E_1, C_3) = (1-\hat{q}_2)v_1 + \hat{q}_2(-(R_1 + r_1) - c_1/k)$
$EU_1(E_1, E_3) = (1-\hat{q}_2)v_1 + \hat{q}_2(\pi v_1 - c_1 - c_1/k)$

We know from Round 3 that

$EU_1(E_1, E_3) > EU_1(E_1, C_3) \iff c_1 - (R_1 + r_1) < \pi v$. Denote **Case E** (for escalate) when $c_1 - (R_1 + r_1) < \pi v_1$ and **Case C** (for concede) when $c_1 - (R_1 + r_1) > \pi v_1$. Denote the value of c_1 for which P_1 will be indifferent between E_1 and C_1 under Case E for a given r_1 as $c_1^E \equiv \frac{kv_1(1-\hat{q}_2+\hat{q}_2\pi)}{\hat{q}_2(k+1)}$, and the value of c_1 for which P_1 will be indifferent between E_1 and C_1 under Case C as $c_1^C(R_1 + r_1) \equiv \frac{k(v_1(1-\hat{q}_2)-(R_1+r_1)\hat{q}_2)}{\hat{q}_2}$.

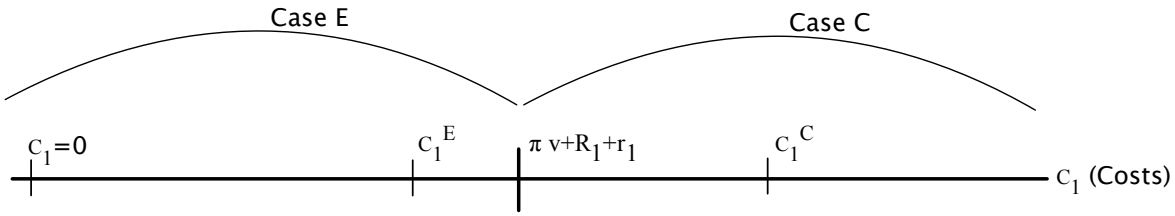

Figure 2: Cases and Indifference Points for P_1's Costs

Within case E, P_1 will prefer E_1 when $c_1 < c_1^E$ and not otherwise. Within case C, P_1 will prefer E_1 when $c_1 < c_1^C$.

Therefore, the probability that P_1 will escalate initially is:

$$p_1 = Pr(c_1 < \pi v_1 + (R_1 + r_1))Pr(c_1 < c_1^E) + Pr(c_1 > \pi v_1 + (R_1 + r_1))Pr(c_1 < c_1^C(R_1 + r_1))$$

A.4 Limiting the Parameter Space Under Consideration

There are three possibilities to consider: (1) **Situation EC**: c_1^E is interior to Case E and c_1^C is interior to Case C; (2) **Situation E**: c_1^E is interior to Case E and c_1^C is not interior to Case C (so that for all values in case C, P_1 prefers to concede); (3) **Situation C**: c_1^E is not interior to Case E and c_1^C is interior to Case C. Situation E is problematic because a change in R_1 or r_1 will have no effect on p_1, q_2, or p_3 since in this equilibrium Player 1 never backs down after initially

escalating. Situation E is more representative of the dynamic of pure force, rather than the logic of crisis bargaining. Situation EC consists of a combination of the results from Situation E and Situation C. To simplify our analysis, we exclude formal consideration of Situation E and Situation EC.

A.5 Only C Indifference Point Interior? (Denoted Situation C)

In order for just c_1^C to be interior to Case C we need that $c_1^C > \pi v_1 + (R_1 + r_1)$ and $\pi v_1 + (R_1 + r_1) < c_1^E$. This implies:

$$k(v_1(1-\hat{q}_2) - (R_1+r_1)\hat{q}_2)/\hat{q}_2 > \pi v_1 + (R_1+r_1)$$

and

$$\pi v_1 + (R_1+r_1) < \frac{1-\hat{q}_2+\pi\hat{q}_2}{\hat{q}_2}\frac{v_1 k}{k+1}$$

Which imply:

$$\hat{q}_2 < \frac{v_1 k}{(R_1+r_1)(1+k)+v_1(\pi+k)} = q_2^C$$

A.6 Solving the Model under Situation C

Situation C requires that:

Condition C: $\hat{q}_2 < q_2^C \equiv \frac{v_1 k}{(R_1+r_1)(1+k)+v_1(\pi+k)}$

In a model with rational expectations, $\hat{q}_2 = q_2^*$, therefore Condition C requires that $q_2^* < q_2^C$ (where q_2^* denotes the equilibrium level of q_2 from the perspective of an observer who doesn't see r_2).

And in order to have an "interior" solution so that some P_1 types don't escalate in the first round, we need that:

Condition I: The upper bound on the support of the distribution for c is greater than

$$c_1^C \equiv \frac{k(v_1(1-\hat{q}_2) - (R_1+r_1)\hat{q}_2)}{\hat{q}_2}$$

A.6.1 Probability P_1 Escalates in First Round (p_1)

$$p_1 = Pr(c_1 < c_1^C(r_1)|r_1) = Pr\left(c_1 < \frac{k(v_1(1-\hat{q}_2) - (R_1+r_1)\hat{q}_2)}{\hat{q}_2}\Big|r_1\right) \tag{3}$$

A.6.2 Probability of P_2 Escalating (q_2)

$$q_2 = Pr\left(c_2 < \frac{v_2(1-\pi\hat{p}_3) + (R_2+r_2)}{\hat{p}_3}\Big|r_2\right) \tag{4}$$

A.6.3 Probability of P_1 Escalating in Final Round (p_3)

$$p_3 = \frac{Pr(c_1 - (R_1+r_1) < \pi v_1|r_1)}{Pr(c_1 + (R_1+r_1) < k(v_1(1-\hat{q}_2))/\hat{q}_2|r_1)} \tag{5}$$

A.7 Existence and Uniqueness of Equilibrium

Given Conditions C and I, there is a unique perfect Bayesian equilibrium. To see this, let $\bar{q}_2$ denote the probability of P_2 escalating from the perspective of an observer who cannot observe r_2, and $\bar{p}_3$ the probability of P_1 escalating in Round 3 from the perspective of an observer who cannot observe r_1. These can be thought of as the "population" best response function since it maps beliefs about the other's strategy into an expected response.

That is,

$$\bar{q}_2(\hat{p}_3) = \int_{r_2=0}^{\infty} q_2(r_2, \hat{p}_3) g_2(r_2) dr_2$$

$$\bar{p}_3(\hat{q}_2) = \int_{r_1=0}^{\infty} p_3(r_1, \hat{q}_2) g_1(r_1) dr_1$$

In equilibrium the beliefs of the agents must be correct: $\hat{q}_2 = \bar{q}_2$, and $\hat{p}_3 = \bar{p}_3$. From 4 and 5 we see that, in equilibrium, $\frac{dq_2(\bar{p}_3)}{d\bar{p}_3} < 0$ and $\frac{dp_3(\bar{q}_2)}{d\bar{q}_2} > 0$ (and therefore $\frac{d\bar{q}_2(\bar{p}_3)}{d\bar{p}_3} < 0$ and $\frac{d\bar{p}_3(\bar{q}_2)}{d\bar{q}_2} > 0$. These expressions have the natural interpretation that P_2 will be less likely to escalate as P_1 is more likely to escalate in Round 3, and P_1 will be more likely to escalate in Round 3 as P_2 is more likely to escalate in Round 2 (because the types of P_1 that select into Round 2 will be more likely to be resolved). We are also able to deduce where the curves $\bar{p}_3(\bar{q}_2)$ and $\bar{q}_2(\bar{p}_3)$ touch their extreme points, which allows us to show that the curves cross each other once and, by monotonicity, only once (see figure 3). Specifically, we know that when $\bar{q}_2 = 0$, there will be a proportion of P_1 types that escalate in Round 3 greater than 0 and less than 1; as $\bar{q}_2$ approaches 1, $\bar{p}_3$ will also approach 1 (and will hit it when $\bar{q}_2 = q_2^C$); as $\bar{p}_3$ approaches 0, $\bar{q}_2$ approaches 1 (and reaches 1 if the distribution on c_2 has an upper bound); when $\bar{p}_3 = 1$, $\bar{q}_2$ equals some positive value (since there are always cost types sufficiently low, that is $c_2 < v_2(1-\pi) + (R_2 + r_2)$). Thus, given continuity in $\bar{q}_2$ and $\bar{p}_3$ and **Condition C** being satisfied, there will be a unique equilibrium.

A.8 Private Honor

A.8.1 Model F with Private Honor

We now consider specific models. Model H with Private Honor refers to the setup in which P_1 is the honor-agent, so that some types of P_1 have high realized values of r_1 ("Southerners") and others have low realized values. Thus, we are looking for comparative statics in values of r_1.

First we note that since P_2 doesn't observe the realized value of r_1, $\frac{d\hat{p}_3}{dr_1} = 0$. By equation 4 this implies

$$\frac{dq_2}{dr_1} = 0$$

This then implies (by equations 3 and 5) that:

$$\frac{dp_1}{dr_1} < 0$$

and

$$\frac{dp_3}{dr_1} > 0$$

In words, these results state that for conflicts in which the focal-agent performs the reputation engaging action the focal-agent will be more likely to escalate above the reputation-engagement

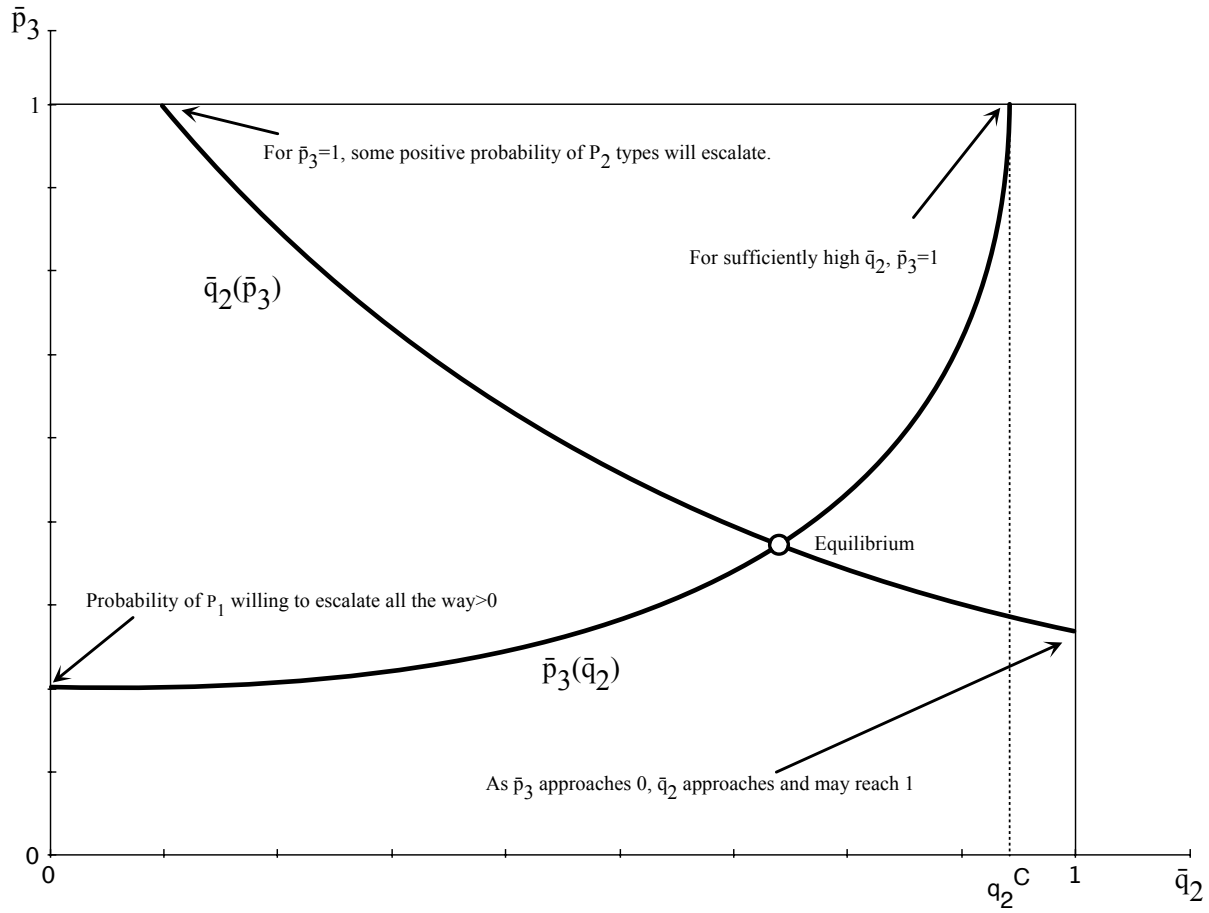

Figure 3: Unique Equilibrium

threshold, and less likely to escalate below the reputation-engagement threshold. The behavior of the rival-agent will not vary since, by assumption, they are unable to perceive a difference between focal-agents with greater or less concern for honor.

A.8.2 Model R with Private Honor

Model R refers to the setup in which P_2 is the focal-agent, so that we are looking for comparative statics in realized values of r_2.

Since P_1 doesn't observe the realized value of r_2, we have (by 3 and 5):

$$\frac{dp_1}{dr_2} = 0$$

$$\frac{dp_3}{dr_2} = 0$$

This then implies that (by 4):

$$\frac{dq_2}{dr_2} > 0$$

In words, this result states that for conflicts in which the rival-agent performs the action that crosses the focal-agent's reputation-engagement threshold, the focal-agent will be more likely to escalate. As before, the rival-agent's behavior will not vary since, by assumption, they are unable to perceive a difference between focal-agents with greater or less concern for honor.

A.8.3 Conclusion

In summary, the selection dynamic in models with private honor is straightforward: honor-agents pay higher costs for making a failed bluff and so are less willing to engage their reputation to a dispute. Both the higher costs of backing down and the biased selection of tough types into conflicts implies that honor-agents are more likely to escalate once their reputation is engaged. These comparative statics form the basis for the hypotheses tested in this paper.

Consideration of models with observable honor complicates these predictions because now the rival-agent will tend to select out of disputes in which the honor-agent is likely to be resolved. In these models, it matters far more which agent is the one responsible for leading the conflict over the reputation-engagement threshold since their equilibrium behavior will then be shaped by selection effects. With suitable auxiliary assumptions, however, we show that these core comparative statics are probably reasonable even in situations with observable honor.

A.9 Model F and R with Observable Honor

Under these models we set $r_1 = 0$ and $r_2 = 0$, so that given rational expectations we have $\hat{p}_3 = p_3$ and $\hat{q}_2 = q_2$. As above, let the cumulative distribution function of cost types for P_1 be $F_1(x)$ and for P_2 be $F_2(x)$. Let $E_1(x) = F_1(x)/F_1(c_1^C)$ for $x < c_1^C$ and $E_1(x) = 1$ for $x \geq c_1^C$. This, then, is the rescaled cumulative distribution function for the portion of $F_1(x)$ below c_1^C. We can then rewrite equations 3, 4, 5 as:

A.9.1 Probability of P1 Escalating in Last Round (p_1)

$$p_1 = F_1(c_1^C) = F_1\left(\frac{k(v(1-q_2) - R_1 q_2)}{q_2}\right) \tag{6}$$

A.9.2 Probability of P2 Escalating (q_2)

$$q_2 = F_2\left(\frac{v(1-\pi p_3) + R_2}{p_3}\right) \tag{7}$$

A.9.3 Probability of P1 Escalating in Final Round (p_3)

$$p_3 = E_1(\pi v + R_1) = \frac{F_1(\pi v + R_1)}{F_1\left(\frac{k(v(1-q_2)-R_1 q_2)}{q_2}\right)} \tag{8}$$

Note that by Condition C, $\pi v + R_1 < c_1^C$, which implies $p_3 < 1$.

A.9.4 Comparative Statics for Models with Observable Honor

$$\frac{dp_3(R_1)}{dR_1} = \frac{dE_1(\pi v + R_1)}{dR_1} = \frac{F_1(c_1^C(R_1))\frac{\partial F_1(\pi v + R_1)}{\partial(\pi v + R_1)} - F_1(\pi v + R_1)\frac{\partial F_1(c_1^C(R_1))}{\partial c_1^C(R_1)}\frac{d(c_1^C(R_1))}{dR_1}}{(F_1(c_1^C(R_1)))^2} \tag{9}$$

$$\frac{dq_2(R_1)}{R_1} = -\frac{\partial F_2\left(\frac{v(1-\pi p_3)+R_2}{p_3}\right)}{\partial\left(\frac{v(1-\pi p_3)+R_2}{p_3}\right)}\frac{\frac{dp_3(R_1)}{dR_1}(R_2+v)}{(p_3(R_1))^2} \tag{10}$$

$$\frac{dp_1(R_1)}{dR_1} = \frac{\partial F_1(c_1^C(R_1))}{\partial(c_1^C(R_1))}\frac{dc_1^C(R_1)}{dR_1} = \frac{\partial F_1(c_1^C(R_1))}{\partial(c_1^C(R_1))}(-k)\left(1 + \frac{v\frac{dq_2(R_1)}{dR_1}}{(q_2(R_1))^2}\right) \tag{11}$$

$$\frac{dp_3(R_2)}{dR_2} = \frac{dE_1(\pi v + R_1)}{dR_2} = \frac{kvF_1(\pi v + R_1)\frac{\partial F_1(c_1^C(R_2))}{\partial c_1^C(R_2)}\frac{dq_2(R_2)}{dR_2}}{(F_1(c_1^C(R_2)))^2(q_2(R_2))^2} \tag{12}$$

$$\frac{dq_2(R_2)}{R_2} = \frac{\partial F_2(\frac{v(1-\pi p_3(R_2))+R_2}{p_3(R_2)})}{\partial\left(\frac{v(1-\pi p_3(R_2))+R_2}{p_3(R_2)}\right)}\left(\frac{p_3(R_2) - \frac{dp_3(R_2)}{dR_2}(R_2+v)}{(p_3(R_2))^2}\right) \tag{13}$$

$$\frac{dp_1(R_2)}{R_2} = -\frac{kv\frac{\partial(F_1(c_1^C(R_2))}{\partial(c_1^C(R_2))}\frac{d(q_2(R_2))}{dR_2}}{(q_2(R_2))^2} \tag{14}$$

Proposition 1.

$$\frac{dp_3}{dR_1} > 0$$

Proposition 2.

$$\frac{dq_2}{dR_2} > 0$$

Propositions 1 and 2 state that as an agent's concern about his reputation increases, the agent will be more likely to escalate conditional on being above the reputation-engagement threshold. That is, agents that are more concerned about reputation that have either selected into a conflict or have been surprised by a reputation-engaging challenge will be more likely to escalate.

Proof of $\frac{dp_3}{dR_1} > 0$ *by contradiction.* Suppose that $\frac{dp_3}{dR_1} \leq 0$. But then equation 10 and the fact that all the terms therein, except $\frac{dp_3}{dR_1}$ and the negative sign, are necessarily positive, implies:

$$\frac{dq_2}{dR_1} \geq 0$$

$$\Rightarrow \left(1 + \frac{v\frac{dq_2(R_1)}{dR_1}}{(q_2(R_1))^2}\right) \geq 0$$

This implies, by equation 11, that

$$\frac{\partial c_1^C(R_1)}{\partial R_1} < 0 \text{ and } \frac{dp_1(R_1)}{dR_1} < 0$$

But this implies by equation 9 that

$$\frac{dp_3(R_1)}{dR_1} > 0$$

contradicting our starting assumption.

□

Proof of $\frac{dq_2}{dR_2} > 0$ *by contradiction.* Suppose that $\frac{dq_2}{dR_2} \leq 0$. Then, by equation 14, $\frac{dp_1}{dR_2} \geq 0$ and by equation 12 $\frac{dp_3}{dR_2} \leq 0$. But then, by equation 13, $\frac{dq_2}{dR_2} > 0$ (so long as $p_3 > 0$ which will be true for Situation C so long as the lower bound on the support of $F(\cdot)$ is lower than $\pi v + R_1$ which we assume to be the case). Contradiction. □

Proposition 3.

$$\frac{dq_2}{dR_1} < 0$$

Proposition 4.

$$\frac{dp_1}{dR_2} < 0$$

Propositions 3 and 4 state that rival-agents are less likely to escalate over the reputation-engagement threshold when facing an honor-agent.

Proof of $\frac{dq_2}{dR_1} < 0$. Equation 10 and Proposition 1 ($\frac{dp_3}{dR_1} > 0$) implies $\frac{dq_2}{dR_1} < 0$. □

Proof of $\frac{dp_1}{dR_2} < 0$. Equations 14 and Proposition 2 imply that $\frac{dp_1}{dR_2} < 0$

□

Proposition 5. $\frac{dp_3}{dR_2} > 0$

Proof of $\frac{dp_3}{dR_2} > 0$. Equations 12 and Proposition 2 imply that $\frac{dp_3}{dR_2} > 0$

□

Proposition 5 states that rival-agents who have selected into the conflict will be more likely to escalate against honor-agents. The logic here is that as P_2 becomes more concerned about honor, P_2 is less likely to back down in Round 2. Consequently, fewer P_1's are willing to gamble that P_2 will back down, whereas a greater proportion of P_1's who escalate in Round 1 intend to escalate again in Round 3.

Proposition 6.

$$\frac{dp_1}{dR_1} > 0 \iff \left(1 + \frac{v\frac{dq_2(R_1)}{dR_1}}{(q_2(R_1))^2}\right) < 0$$

$$\iff \frac{dq_2(R_1)}{dR_1} < -\frac{(q_2(R_1))^2}{v}$$

This follows immediately from equation 11.

Proposition 6 states that an increase in concern for honor may lead an honor-agent, below the reputation-engagement threshold, to escalate more or less. The honor-agent will escalate more at the pre-engagement stage if the rival is sufficiently sensitive to an increase in the honor-agent's concern for honor, if v is sufficiently large, and/or if q_2 is sufficiently low.

A.10 Signing $\frac{dp_1}{dR_1}$ for Specific Distributions

To sign $\frac{dp_1}{dR_1}$ we need to determine the actual equilibrium solutions. For uniform uncertainty this is done analytically. For logistic uncertainty this is done numerically.

A.10.1 Uniformly Distributed Types

We now assume that the costs are drawn from a uniform distribution. $c_1 \sim U[L_1, H_1]$, $c_2 \sim U[L_2, H_2]$, where $H_1 > 0$, $H_2 > 0$. In order to make this model tractable, an additional simplifying assumption is later made that $L_1 = L_2 = 0$. For convenience we also assumed that $\pi = 1/2$.

$$F_1(x) = \int_{L_1}^{x} \frac{1}{H_1 - L_1} = \frac{x - L_1}{H_1 - L_1}$$

$$F(c_1^C) = \frac{c_1^C - L_1}{H_1 - L_1}$$

$$E_1(x) = \begin{cases} \frac{x-L_1}{H_1-L_1}\frac{H_1-L_1}{c_1^C-L_1} & \text{if } x < c_1^C \\ 1 & \text{if } x \geq c_1^C \end{cases}$$

$$\Rightarrow E_1(x) = \begin{cases} \frac{x-L_1}{c_1^C-L_1} & \text{if } x < c_1^C \\ 1 & \text{if } x \geq c_1^C \end{cases}$$

$$F_2(x) = \frac{x - L_2}{H_2 - L_2}$$

Therefore, setting $L_1 = L_2 = 0$ and remembering that in equilibrium $c_1^C = \frac{k(v(1-q_2)-R_1q_2)}{q_2}$ we have that:

$$p_3 = \frac{q_2(v/2 + R_1)}{k(v(1-q_2) - R_1q_2)}$$

$$q_2 = \frac{v(2-p_3) + 2R_2}{H_2 2p_3}$$

Let $Z = -v(2R_1 + v) + 4k(R_1 + v)(R_2 + v)$. $Z > 1$ for the permitted values of the parameters. Solving for p_3 gives.

$$p_3 = \frac{Z - \sqrt{16kv(2H_2 + R_1 + v)(2R_1 + v)(R_2 + v) + (-Z)^2}}{4kv(2H_2 + R_1 + v)}$$

or

$$p_3 = \frac{Z + \sqrt{16kv(2H_2 + R_1 + v)(2R_1 + v)(R_2 + v) + (-Z)^2}}{4kv(2H_2 + R_1 + v)}$$

This shows that, within the numerator, the term under the radical sign is larger than the term outside the radical. Therefore, the first solution is always negative and should be discarded. The second solution is always positive. It is now possible to solve for q_2 and p_1:

$$q_2 = \frac{-4k(R_1 + v)(R_2 + v) - v(2R_1 + v) + \sqrt{16kv(2H_2 + R_1 + v)(2R_1 + v)(R_2 + v) + (v(2R_1 + v) - 4k(R_1 + v)(R_2 + v))^2}}{4H_2(2R_1 + v)}$$

$$p_1 = \frac{k(v(1-q_2) - R_1q_2)}{H_1q_2}$$

$$= -\frac{1}{16H_1{H_2}^2(2R_1 + v)^2}k$$

$$\Big(v(2R_1 + v) + 4k(R_1 + v)(R_2 + v) - \sqrt{16kv(2H_2 + R_1 + v)(2R_1 + v)(R_2 + v) + (v(2R_1 + v) - 4k(R_1 + v)(R_2 + v))^2}\Big)$$

$$\Big(4H_2v(2R_1 + v) + 4k(R_1 + v)^2(R_2 + v) + (R_1 + v)$$

$$\Big(2R_1v + v^2 - \sqrt{v^2(2R_1 + v)^2 + 8kv(4H_2 + R_1 + v)(2R_1 + v)(R_2 + v) + 16k^2(R_1 + v)^2(R_2 + v)^2}\Big)\Big)$$

Proposition 7. *For Uniform Uncertainty,* $\frac{dp_1}{dR_1} > 0$

The proof of proposition 7 involves signing the derivative of p_1, which is too cumbersome to report here. Please contact the authors for more information.

A.10.2 Logistically Distributed Types

The model was not analytically tractable for logistically distributed uncertainty, but through numerical methods it was possible to establish that $\frac{dp_1}{dR_1} < 0$ for many parameter values with logistic uncertainty (and $\frac{dp_1}{dR_1} > 0$ for some). Let μ_1 and μ_2 be the means of the logistic distribution for player 1 and player 2, respectively, and σ_1 and σ_2 the corresponding standard deviation of the distributions. Specifically, the probability of escalation given a best response to a correct conjecture about the probability of the other player escalating was calculated ($p_3(q_2)$ and $q_2(p_3)$). We then started with $q_2 = 0.5$ and iterated through both functions until we found the fixed point. The values tended to converge, so long as σ_1 and σ_2 were not too small (which reflects how the conflict equilibrium is unstable when there is low uncertainty). We then plotted q_2, $p_3(q_2)$, and $p_1(q_2)$ as

a function of R_1 and R_2 and calculated the change in each as a function of small changes in r_1 and r_2. We selected particular realizations of the parameter values, varied R_1 through 40 values between 0 and 2, examined the proportion of those in which $dp_1/dR_1 < 0$. We report these in table A.10.2. To summarize, the comparative statics in the model with logistically distributed costs for the parameter values examined were identical to those for the uniform case (as they should be), with one exception: for most parameter values, $\frac{dp_1}{dR_1} < 0$.

Table 1: Numerical Examination of $\frac{dp_1}{dR_1}$ for Logistic Distribution

μ_1	μ_2	σ_1	σ_2	v	k	r_2	Proportion with $\frac{dp_1}{dR_1} < 0$
4	4	4	4	1	4	0	1.0
4	4	4	4	1.5	6	0	0.78
4	4	5	5	1	6	0	1.0
4	4	4	4	1	6	1.0	1.0
4	4	4	5	2	3	0	0.9

A.11 Summary of Comparative Statics

Table 2: Comparative statics for Models F and R with private and observable honor.

Model	**Round**	**Agent**	**Action**	**Private Honor**	**Observable Honor**
F	1	Focal	Engage reputation	$dp_1/dr_1 < 0$	Depends on uncertainty: Uniform $\implies dp_1/dR_1 > 0$ Logistic $\implies$ (mostly) $dp_1/dR_1 < 0$
F	2	Rival	Use force	$dq_2/dr_1 = 0$	$dq_2/dR_1 < 0$
F	3	Focal	Use force	$dp_3/dr_1 > 0$	$dp_3/dR_1 > 0$
R	1	Rival	Engage reputation	$dp_1/dr_2 = 0$	$dp_1/dR_2 < 0$
R	2	Focal	Use force	$dq_2/dr_2 > 0$	$dq_2/dR_2 > 0$
R	3	Rival	Use force	$dp_3/dr_2 = 0$	$dp_3/dR_2 > 0$

Table 2 summarizes the comparative statics of the four models. The results are most straightforward for the models with private honor: honor-agents are less likely to escalate before their reputation is engaged ($dp_1/dr_1 < 0$) because the costs of a failed bluff are greater. The increased reputational costs of a concession and the greater reluctance of honor-agents to commit their reputation when costs are high implies that honor-agents whose reputation does become engaged tend to be more willing to escalate further ($dp_3/dr_1 > 0$ and $dq_2/dr_2 > 0$).

The models with observable honor involve more complicated dynamics because others anticipate honor-agents' greater concern for coercive reputation and adjust their behavior accordingly. As is the case in models with private honor, when honor is observable honor-agents are more likely to escalate once their reputation is engaged ($dp_3/dR_1 > 0$ and $dq_2/dR_2 > 0$). In addition, so long as they have not selected into a conflict in which reputation is engaged, players facing an honor-agent are less likely to escalate the dispute ($dq_2/dR_1 < 0$ and $dp_1/dR_2 < 0$). However,

rival-agents who have selected into a conflict in which reputation is engaged will, unlike the models of private honor, be more likely to escalate against an honor-agent ($dp_3/dR_2 > 0$). Moreover, the comparative statics of honor-agents' probability of escalation before their reputation is engaged are indeterminate ($dp_1/dR_1 =?$).

A.12 Empirical Implications of the Model

The models yield the following predictions:

1. **Use of Force:** The clearest prediction that emerges from our models concerns the use of force. Following from the fact that $dp_3/dr_1 > 0$, $dp_3/dR_1 > 0$, $dq_2/dr_2 > 0$ and $dq_2/dR_2 > 0$, we can conclude that for models with either private and observable honor, for all possible distributions of types that satisfy Conditions C and I, honor-agents will be more likely to use force in a militarized dispute. Furthermore, this result is present for both Models F and R, which is to say that the result holds irrespective of which party is responsible for the action that engages reputation: selection effects do not muddy this prediction.

   > **Hypothesis 1**: *Conditional on a MID occurring, leaders from a culture of honor will be more likely to use force.*

2. **Duration of Disputes:** Let ϕ_2 and ϕ_3 be the length of a MID that ends, respectively, with a concession in the second or third round. Let ϕ_4 be the duration of a MID that involves escalation to the highest level (escalation in Round 3). Let $\phi_4 > \phi_3 > \phi_2 > 0$. Then, the average length of a MID is

$$\begin{aligned}\phi = \phi_2 Pr(&\text{conflict ends at Round 2}|\text{Round 2 has been reached})\\ +\phi_3 Pr(&\text{conflict ends at Round 3}|\text{Round 2 has been reached})\\ +\phi_4 Pr(&\text{conflict escalates in Round 3}|\text{Round 2 has been reached})\\ &= \phi_2(1-q_2) + \phi_3 q_2(1-p_3) + \phi_4 q_2 p_3\\ \phi &= q_2\left(p_3(\phi_4-\phi_3)+\phi_3-\phi_2\right)+\phi_2\end{aligned}$$

   Differences in the reputational concern of a leader will be associated with differences in the average MID length according to the following equations:

$$\frac{d\phi}{dr_i} = \frac{dq_2}{dr_i}(\phi_3 - \phi_2 + p_3(\phi_4-\phi_3)) + \frac{dp_3}{dr_i}q_2(\phi_4-\phi_3) \tag{15}$$

$$\frac{d\phi}{dR_i} = \frac{dq_2}{dR_i}(\phi_3 - \phi_2 + p_3(\phi_4-\phi_3)) + \frac{dp_3}{dR_i}q_2(\phi_4-\phi_3) \tag{16}$$

   The comparative statics in 15 and 16 are positive for all type distributions, with both private and observable honor, under Model R and Model F, with one exception: under Model F with observable honor, the sign of equation 16 could be positive or negative. In words, this result says that disputes experienced by Southerners should last longer than those experienced by non-Southerners. In Online Appendix A.13 we show that for reasonable parameter values these results hold also for Model F with observable honor.

Hypothesis 2: *The durations of MIDs experienced by leaders from a culture of honor will tend to be greater.*

3. **Outcome of Disputes:** Let w_F and w_R denote the proportion of MIDs which the focal-agent wins in Model F and Model R, which we interpret as occurring when the rival agent concedes: $w_F = 1 - q_2$; $w_R = q_2(1 - p_3)$. Similarly, let l_F and l_R be the proportion of MIDs the focal-agent loses: $l_F = q_2(1 - p_3)$ and $l_R = 1 - q_2$. Then:

$$\frac{dw_F}{dr_1} = -\frac{dq_2}{dr_1} = 0; \quad \frac{dw_F}{dR_1} = -\frac{dq_2}{dR_1} > 0; \quad \frac{dw_R}{dr_2} = \frac{dq_2}{dr_2}(1 - p_3) - \frac{dp_3}{dr_2}q_2 > 0; \quad \frac{dw_R}{dR_2} =? \quad (17)$$

$$\frac{dl_F}{dr_1} = \frac{dq_2}{dr_1}(1 - p_3) - \frac{dp_3}{dr_1}q_2 < 0; \quad \frac{dl_F}{dR_1} < 0; \quad \frac{dl_R}{dr_2} = -\frac{dq_2}{dr_2} < 0; \quad \frac{dl_R}{dR_2} < 0 \quad (18)$$

In words, this says that for both private and observable honor, under models F and R, for all type distributions, honor agents will (weakly) win more (expressions in 17 are weakly positive) and lose less (expressions in 18 are negative), with one exception ($\frac{dw_R}{dR_2}$): under Model R with observable honor honor-agents may win more or less. In Online Appendix A.13 we show that for reasonable parameter values this comparative static is also consistent with the others ($\frac{dw_R}{dR_2} > 0$).

Hypothesis 3: *Leaders from cultures of honor will be more likely to win and less likely to lose their MIDs.*

A.13 Numerical Examination of Hypotheses

Predictions from the family of models on duration and victory are consistent in all of the models with one ambiguous exception for each prediction. This section will examine these ambiguous exceptions to confirm that there exist parameter values where the prediction is consistent with the other ones, as well as to see if there are parameter values where the prediction goes contrary to the other ones. The formal theoretic study of interstate conflict is not at the point where we could credibly calibrate our model on real world data in order to estimate reasonable values of parameters, as is done for example in macroeconomics. Instead, we select values of the parameters so that Condition I and Condition C are satisfied, and the slopes of the response functions are such that the computational methods can discover the solution readily.[6] Absent any guiding principle about how best to search the parameter space, we simply select a number of parameter sets and demonstrate that the prediction could be signed positive or negative.

These findings slightly qualify our formal predictions on duration and victory. We found that three out of four models yield consistent predictions, and the fourth model yields predictions consistent with these under many parameter values. From this we conclude that the models predict that honor-agents will experience longer MIDs and will be more likely to win, with the qualification that these predictions are weaker than that regarding the use of force. As it turns out, the empirical association on use of force is the most significant. Mathematica code for replicating these analyses is available upon request.

[6] For some distributions on costs the equilibrium is unstable in the sense that an iterated mapping through the response functions from a point close to the equilibrium diverges.

A.13.1 Duration in Model F with Observable Honor: $\frac{d\phi}{dR_1}$

$$\frac{d\phi}{dR_1} = \frac{dq_2}{dR_1}(\phi_3 - \phi_2 + p_3(\phi_4 - \phi_3)) + \frac{dp_3}{dR_1}q_2(\phi_4 - \phi_3)$$

where $\frac{dq_2}{dR_1} < 0$ and $\frac{dp_3}{dR_2} > 0$. Therefore $\frac{d\phi}{dR_1}$ is more likely to be positive when ϕ_4 is large relative to ϕ_3 and ϕ_2. In table 3, we report numerical analyses in which we solve for the equilibrium for particular values of the parameters, and report the results by stating how large ϕ_4 would have to be, relative to ϕ_3 and ϕ_2.

Table 3: Numerical Examination of $\frac{d\phi}{dR_1}$

μ_1	μ_2	σ_1	σ_2	v	k	R_1	R_2	Condition for $\frac{d\phi}{dR_1} > 0$
4	4	3	3	1	6	0	0	$\phi_4 < 7.6\phi_2 - 6.7\phi_3$
4	4	4	4	1	6	0	0	$\phi_4 > 2.5\phi_3 - 1.5\phi_2$
4	4	5	5	1	6	0	0	$\phi_4 > 1.9\phi_3 - 0.9\phi_2$
4	4	4	4	1.5	6	0	0	$\phi_4 > 3.3\phi_3 - 2.3\phi_2$
4	4	4	4	1.5	1.1	0	0	$\phi_4 > 1.8\phi_3 - 0.8\phi_2$
4	4	4	4	1.5	2	0	0	$\phi_4 > 2.2\phi_3 - 1.2\phi_2$
4	4	4	4	1.5	4	0	0	$\phi_4 > 4\phi_3 - 3\phi_2$
4	4	4	4	1.5	6	0	0.5	$\phi_4 > 0.3\phi_3 - 8.3\phi_2$
4	4	4	4	1.5	6	0	1.5	$\phi_4 > 3.7\phi_3 - 2.7\phi_2$
4	4	4	4	1.5	6	0.5	0	$\phi_4 > 2.4\phi_3 - 1.4\phi_2$

A.13.2 Victory in Model R with Observable Honor: $\frac{w_R}{dR_2}$

Table 4 reports the direction of the prediction on victory for honor-agents in Model R with observable honor for different values of the parameters.

Table 4: Numerical Examination of $\frac{w_R}{dR_2}$

μ_1	μ_2	σ_1	σ_2	v	k	R_1	R_2	$\frac{w_R}{dR_2}$
4	4	3	3	1	6	0	0	+
4	4	4	4	1	6	0	0	+
4	4	5	5	1	6	0	0	+
4	4	4	4	1.5	6	0	0	+
4	4	4	4	1.5	1.1	0	0	−
4	4	4	4	1.5	2	0	0	−
4	4	4	4	1.5	4	0	0	+
4	4	4	4	1.5	6	0	0.5	+
4	4	4	4	1.5	6	0	1.5	−

References

Slantchev, Branislav L. 2011. *Military Threats: The Costs of Coercion and the Price of Peace*. New York: Cambridge UP. 4

Werner, Suzanne. 2000. "The Effects of Political Similarity on the Onset of Militarized Disputes, 1818-1985." *Political Science Quarterly* 53(2):343–374. 4