How do Observers Assess Resolve?

Supplementary Appendix

Contents

1 Robustness checks

As we note in the main text, conjoint experimental designs rest on a small number of assumptions (Hainmueller, Hopkins and Yamamoto, 2014).[1] First, the *stability and carryover effects* assumption, which mandates that potential outcomes remain stable across periods — that is, a participant should choose a given country (conditional on its and the other country's attributes) regardless of what countries or choices they had seen previously. This is not only important to ensure that the AMCE is a meaningful quantity of interest, but also is important to guard against the potential for demand effects, in which participants might respond differently over multiple rounds as they become familiar with the purpose of the study. Second, the *no profile order effects* assumption, which posits that respondents' choices would be the same regardless of the order in which the two countries are presented within a choice task. Third is successful randomization, which simply assumes that the attributes of each profile are randomly generated. We test each of these assumptions in turn.

First, re-estimating the quantities of interest within each round suggest that the AMCEs are stable across rounds, except for in the sixth round, where the treatment effects are noticeably smaller in that round than in all of its counterparts. These temporarily smaller effect sizes are difficult to explain theoretically and are unlikely to be due to respondent fatigue, since the treatment effects in subsequent rounds return to their previous levels. We replicate the results from the main analysis in Figure 1, this time dropping the sixth round, and find that the results remain the same. Since including the sixth round doesn't substantively change our results (and if anything, renders them slightly more conservative), we include all eight rounds in all of the empirical results reported in the main text. This intertemporal stability also mitigates any potential concerns about demand effects, as our participants respond to the treatments in the last round much as they do in the first. We return to this point in Appendix §4.

Second, Figure 2 tests the profile order assumption, showing that the results do not appear to systematically differ based on whether a particular characteristic was presented as belonging to country A or country B. Third, Table 1 presents the results from our randomization check, showing that the treatment assignments are relatively well-balanced across a host of demographic characteristics. Finally, Figure 3 tests whether the row order in which attributes were presented to respondents within the experimental stimuli changes their results. Recall that although the order

[1] For examples of recent conjoint experiments in IR, see Ballard-Rosa, Martin and Scheve (2017); Huff and Kertzer (2018).

of attributes was randomized across respondents, it was held constant across rounds for any given respondent: that is, if a participant saw regime type in the first row of the table in the first round, that participant saw regime type in the first row of the table throughout all seven subsequent rounds. Figure 3 shows that there do not appear to be systematic differences in the AMCEs by attribute order: characteristics presented in the first row are not significantly stronger than those in the last row, for example, and characteristics presented in the first and last rows are not systematically different from those presented in the middle. Interestingly, although the row order randomization is designed to prevent treatments featured more prominently in the table from having a stronger effect, in the "real world", media outlets, political entrepreneurs, and political elites are likely to play a role in ensuring people are more likely to receive some treatments than others. Future work could therefore benefit from exploring these added layers to the information environment in which observers are situated.

Since the results in the main text are only presented graphically, Table 2 presents the results in tabular form instead; the coefficient estimates, standard errors, and 95% confidence intervals are derived from a pair of bootstrapped regression models ($B = 1500$), clustered at the respondent-level to account for the clustered structure of the data; the left half of the table presents the unweighted results, while the right half of the table presents weighted results (see Appendix §4 for details), although as shown in the main text, the results are nearly identical regardless of whether weights are used.

Finally, while the results in the main text focus on the question of how *observers* assess resolve rather than how participants do, since the original experiment also included observations where the US was itself a participant in the disputes, Figure 4 replicates the results from the main text but also including disputes where the US was a participant. Given the larger sample size, the confidence intervals narrow somewhat, but the pattern of results holds: assessments of resolve are heavily driven by capabilities, stakes, past actions, and to a lesser extent, costly signals. The interesting difference is that American respondents see the United States as significantly more resolved than other actors, an effect even larger than that of capabilities and stakes. Future research should examine whether foreign observers are as optimistic about American resolve as Americans themselves are.

Table 1: Randomization check

	(1)	(2)	(3)	(4)
	High capabilities	High stakes	New leader	Male leader
(Intercept)	[-0.184, 0.133]	[-0.18, 0.138]	[-0.067, 0.257]	[-0.189, 0.13]
Male	[-0.093, 0.034]	[-0.027, 0.096]	[-0.052, 0.073]	[-0.012, 0.113]
Age	[-0.004, 0.002]	[-0.001, 0.004]	[-0.003, 0.002]	[-0.003, 0.003]
Education	[-0.024, 0.024]	[-0.044, 0.001]	[-0.033, 0.011]	[-0.017, 0.029]
Party ID	[-0.039, 0.214]	[-0.144, 0.105]	[-0.128, 0.118]	[-0.158, 0.094]
Mil Assert	[-0.105, 0.227]	[-0.133, 0.198]	[-0.251, 0.074]	[-0.204, 0.137]

	(5)	(6)	(7)	(8)
	Adversary	Against adversary	Initiator	Backed down
(Intercept)	[0.008, 0.276]	[-0.173, 0.101]	[-0.114, 0.197]	[-0.14, 0.172]
Male	[-0.032, 0.091]	[-0.039, 0.078]	[-0.087, 0.041]	[-0.034, 0.094]
Age	[-0.003, 0.003]	[-0.002, 0.003]	[-0.004, 0.001]	[-0.006, 0]
Education	[-0.044, 0.001]	[-0.02, 0.025]	[-0.024, 0.022]	[-0.014, 0.033]
Party ID	[-0.216, 0.041]	[-0.102, 0.172]	[-0.267, -0.024]	[-0.12, 0.132]
Mil Assert	[-0.261, 0.067]	[-0.227, 0.106]	[0.033, 0.365]	[-0.074, 0.256]

	(9)	(10)	(11)	(12)
	Same leader	Democracy	Mixed	High service
(Intercept)	[-0.096, 0.215]	[-0.283, 0.088]	[-0.354, 0.008]	[-0.115, 0.262]
Male	[-0.051, 0.069]	[-0.019, 0.13]	[0.021, 0.165]	[-0.133, 0.026]
Age	[-0.005, 0]	[-0.001, 0.006]	[0, 0.007]	[-0.005, 0.002]
Education	[-0.021, 0.026]	[-0.019, 0.034]	[-0.018, 0.037]	[-0.03, 0.026]
Party ID	[-0.063, 0.188]	[-0.256, 0.057]	[-0.168, 0.145]	[0.005, 0.31]
Mil Assert	[-0.18, 0.149]	[-0.169, 0.202]	[-0.216, 0.18]	[-0.32, 0.114]

	(13)	(14)	(15)
	Some service	Mobilized troops	Public threat
(Intercept)	[-0.127, 0.234]	[-0.26, 0.112]	[-0.164, 0.189]
Male	[-0.071, 0.089]	[-0.012, 0.141]	[-0.022, 0.136]
Age	[-0.001, 0.005]	[-0.002, 0.004]	[-0.001, 0.005]
Education	[-0.051, 0.002]	[-0.055, 0]	[-0.068, -0.011]
Party ID	[-0.071, 0.241]	[-0.137, 0.174]	[-0.171, 0.139]
Mil Assert	[-0.424, -0.013]	[0.011, 0.404]	[-0.082, 0.338]

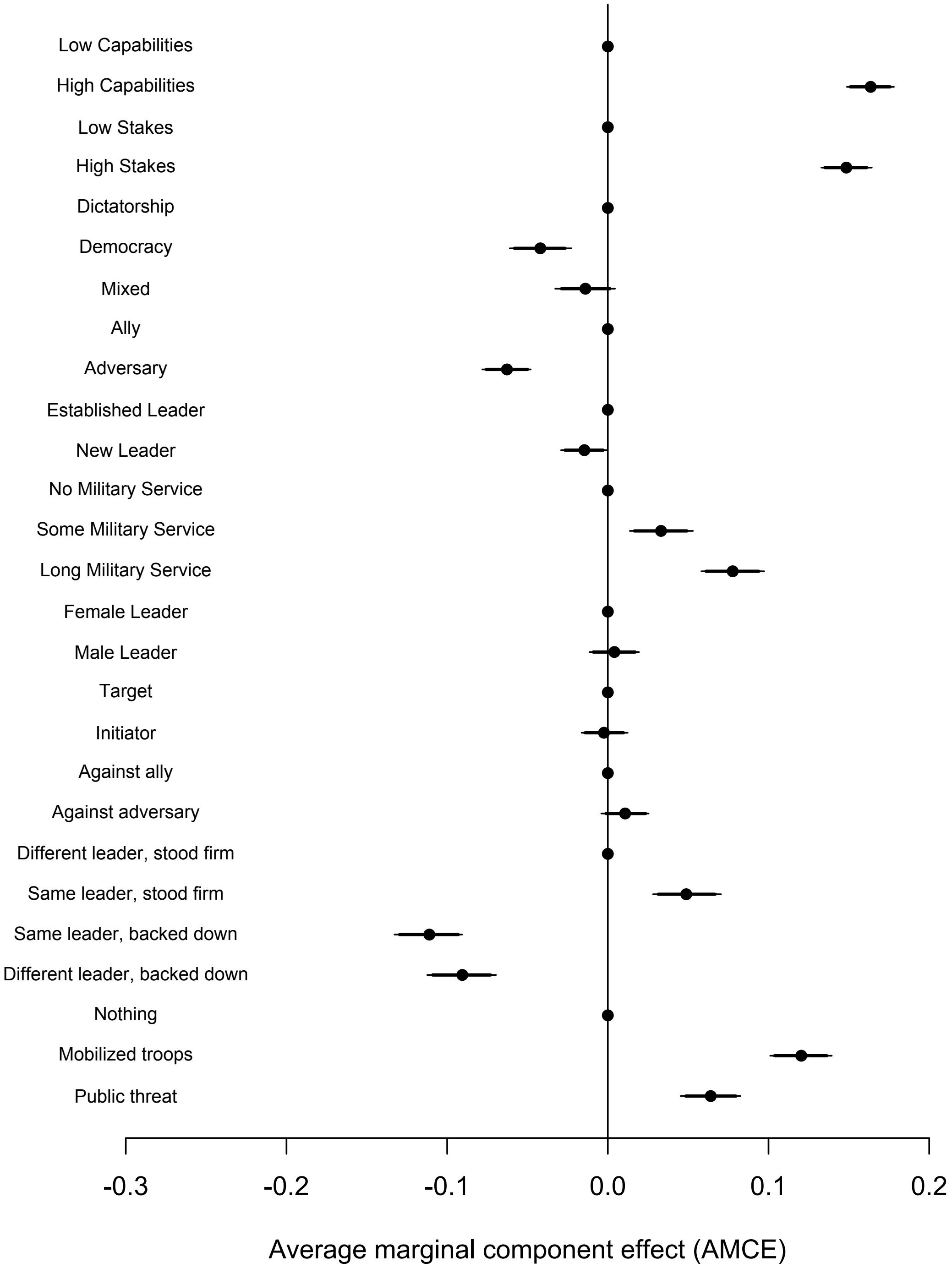
Note: Models 1- 9 depict quantile-based clustered boostrapped 95% confidence intervals from a series of logistic regression models, while models 10-15 depict the quantile-based clustered bootstrapped 95% confidence intervals from a series of multinomial logit models. The results show the treatment assignment is relatively well-balanced.

Table 2: Regression model specification of main results

	Unweighted			Weighted		
	β	SE	95% CI	β	SE	95% CI
(Intercept)	0.338	(0.015)	(0.307, 0.368)	0.341	(0.015)	(0.29, 0.393)
High Capabilities	0.163	(0.008)	(0.148, 0.179)	0.16	(0.008)	(0.136, 0.183)
High Stakes	0.148	(0.008)	(0.133, 0.164)	0.143	(0.008)	(0.12, 0.167)
Democracy	-0.042	(0.009)	(-0.061, -0.023)	-0.033	(0.009)	(-0.065, -0.002)
Mixed regime	-0.014	(0.009)	(-0.033, 0.005)	-0.017	(0.009)	(-0.047, 0.011)
Adversary	-0.063	(0.008)	(-0.078, -0.047)	-0.059	(0.008)	(-0.083, -0.037)
New Leader	-0.015	(0.008)	(-0.029, -0.002)	-0.024	(0.008)	(-0.048, 0)
Long Military Service	0.078	(0.009)	(0.059, 0.098)	0.08	(0.009)	(0.051, 0.108)
Some Military Service	0.033	(0.009)	(0.015, 0.051)	0.03	(0.009)	(0.003, 0.057)
Male Leader	0.004	(0.008)	(-0.011, 0.02)	0.019	(0.008)	(-0.005, 0.043)
Initiator	-0.003	(0.008)	(-0.017, 0.013)	-0.007	(0.008)	(-0.031, 0.016)
Against adversary	0.01	(0.008)	(-0.004, 0.025)	0.015	(0.008)	(-0.008, 0.038)
Different leader, backed down	-0.09	(0.011)	(-0.112, -0.068)	-0.095	(0.011)	(-0.127, -0.062)
Same leader, backed down	-0.111	(0.011)	(-0.131, -0.09)	-0.136	(0.011)	(-0.168, -0.104)
Same leader, stood firm	0.049	(0.011)	(0.028, 0.07)	0.052	(0.011)	(0.018, 0.085)
Mobilized troops	0.12	(0.009)	(0.102, 0.138)	0.118	(0.009)	(0.091, 0.145)
Public threat	0.065	(0.009)	(0.046, 0.084)	0.064	(0.009)	(0.036, 0.09)

Note: quantities of interest derived from clustered bootstrapped regression models. Reference categories: low capabililties, low stakes, dictatorship, ally, established leader, no military service, female leader, target, against ally, different leader stood firm, nothing.

Figure 1: Robustness check: dropping the sixth choice task



To test the stability and no carryover effects assumption, Figure 1 plots Average Marginal Component Effects (AMCEs) with 95% clustered bootstrapped confidence intervals, replicating Figure 1 from the main text, but without including data from the sixth choice task, which displayed violations of the caryover assumption. Importantly, the results hold, and are stronger than the more conservative estimates presented in the text. As before, positive values indicate that the attribute increases the perceived likelihood that an actor will stand firm, while negative values indicate that the attribute decreases the perceived likelihood of an actor standing firm. Thus, for example, democracies are perceived as 4% less likely to stand firm than dictatorships.

Figure 2: Model diagnostics: testing the profile order assumption

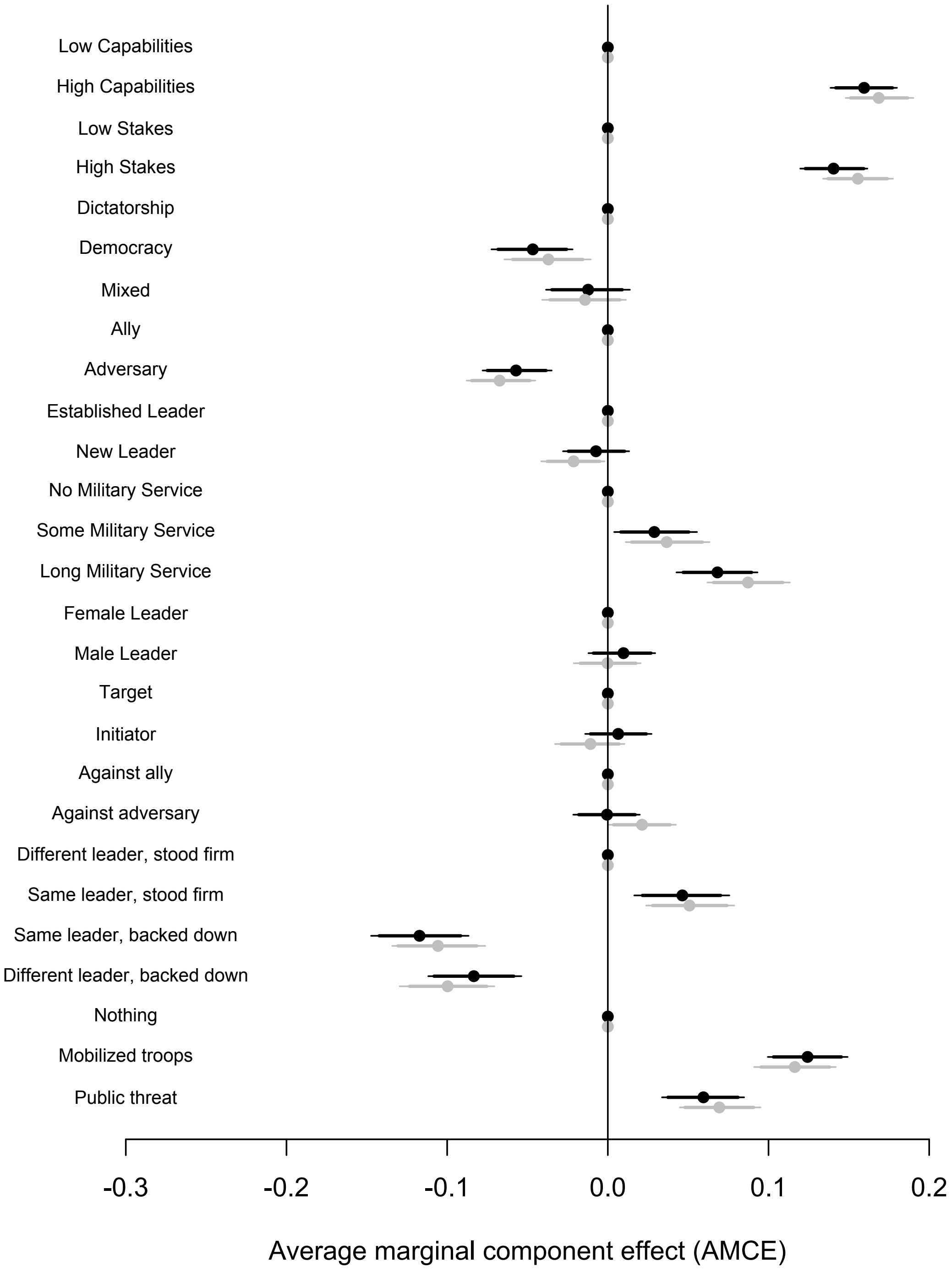


Figure 2 plots Average Marginal Component Effects (AMCEs) with 95% clustered bootstrapped confidence intervals, replicating the results from the main text, but conditional on whether each characteristic was presented as belonging to country A (in black) or country B (in grey). Importantly, the results do not appear to systematically differ based on whether a particular characteristic was presented first or second.

Figure 3: Model diagnostics: treatment effects do not systematically vary by attribute order

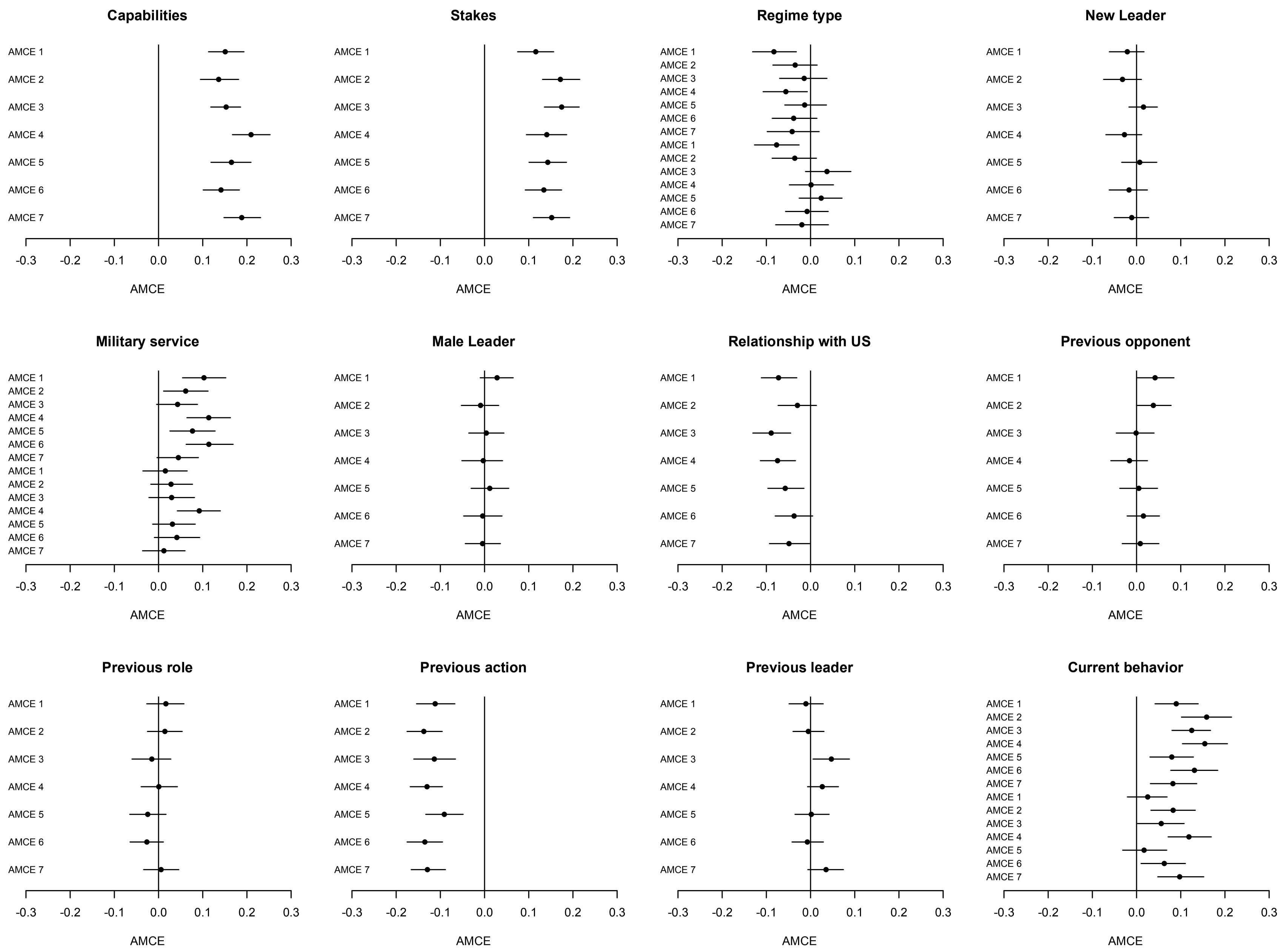
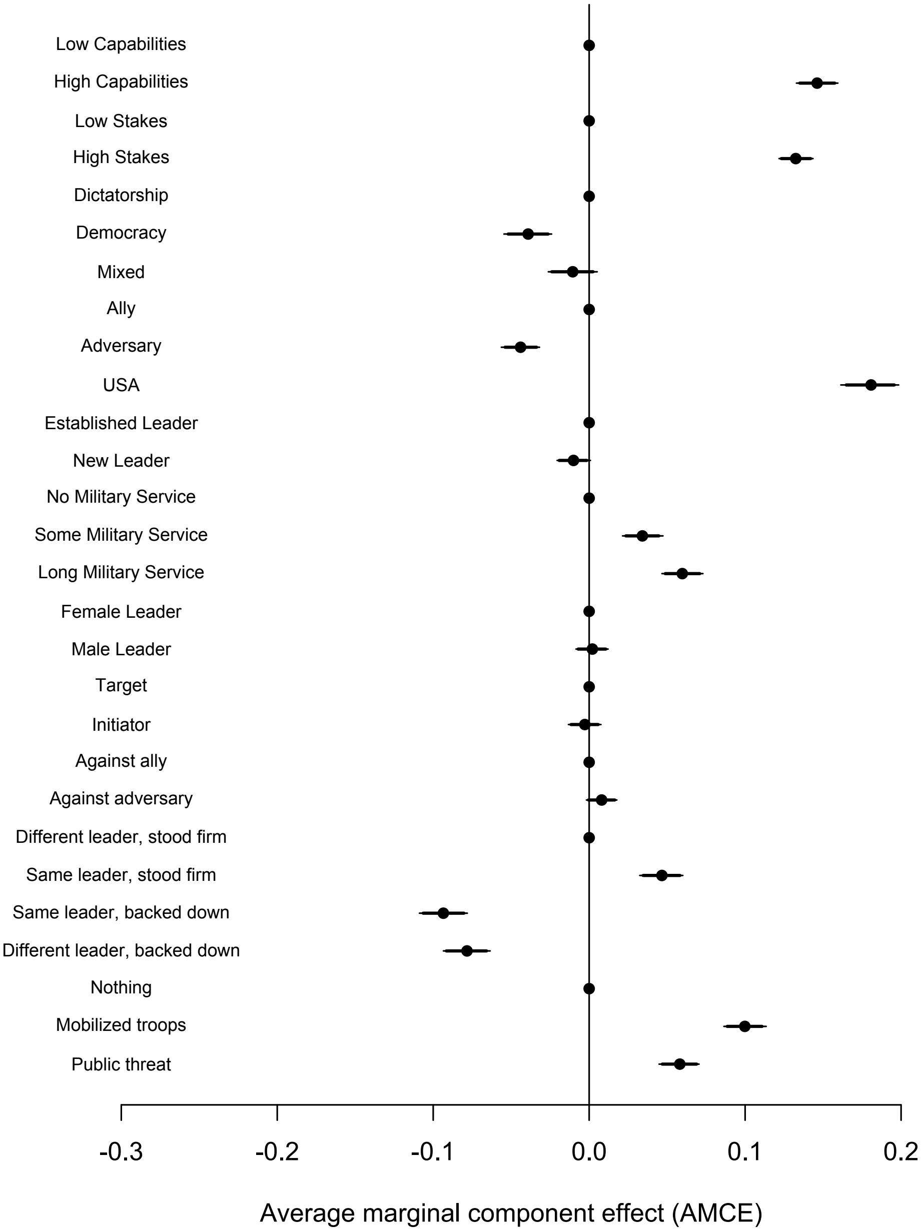


Figure 4: Results including the US as a participant



When we replicate the results from the main text but include disputes where the US itself is a participant, the confidence intervals shrink as a result of the larger number of observations, but the pattern of results is largely the same: assessments of resolve are heavily driven by capabilities, stakes, past actions, and to a lesser extent, costly signals. The interesting difference is that American respondents see the United States as significantly more resolved than other actors, an effect even larger than that of capabilities and stakes.

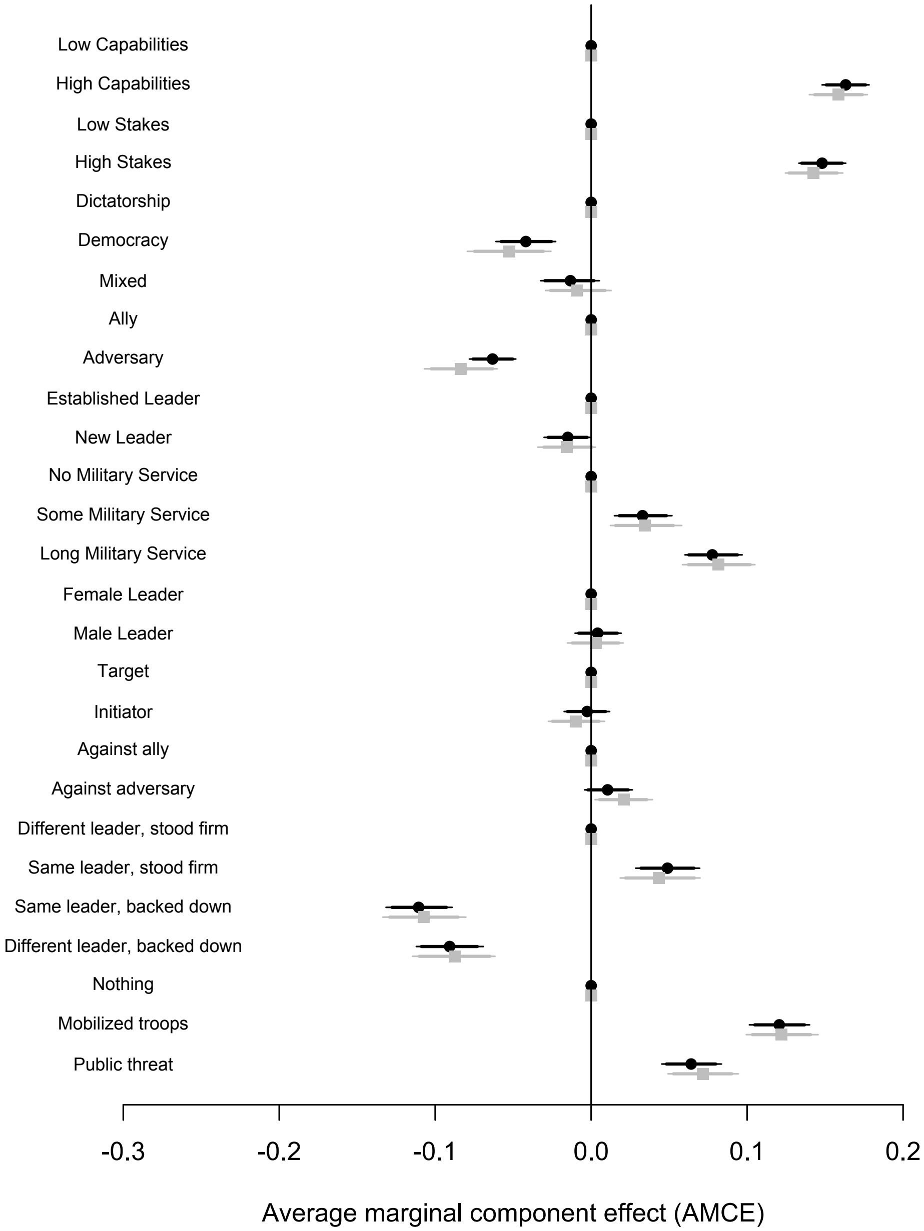
1.1 Dropping unusual dyadic combinations

As noted in the main text, the analyses presented above randomly generated characteristics of both countries in each dispute, such that the leader characteristics of one country, for example, are independent of the leader characteristics of the other. To prevent odd combinations of actor characteristics biasing the experimental results, we imposed the following randomization constraints *ex ante*:

- if a country was assigned to be the United States, it was always described as being a democracy and having a very powerful military
- if a country was assigned to be the United States, the identity of its opponent was constrained such that it could not also be the United States.

However, to preclude the possibility of other unusual dyadic combinations of characteristics skewing the results, Figure 5 below replicates the results from the main set of analyses, but dropping all disputes where two democracies faced off against one another, or two allies of the United States faced off against one another. A comparison of this restricted set of observations (in grey) with the unrestricted set from the main analyses (in black) shows that the results hold regardless of whether the dyads are included or not — it appears that allies are seen as relatively more resolute in the restricted model than the unrestricted one, but the difference is not statistically significant.

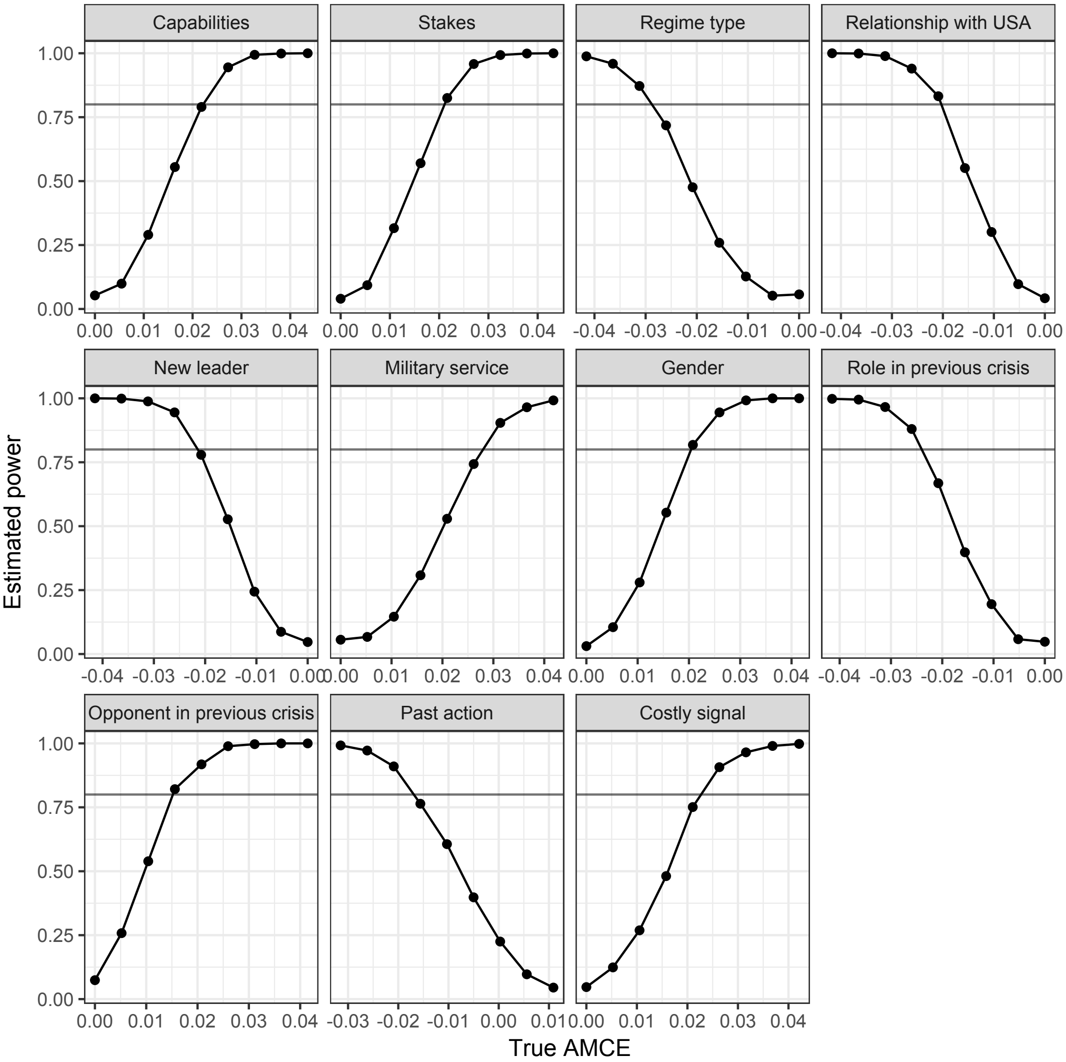
Figure 5: Dropping unusual dyadic combinations



1.2 Power simulations

Figure 6 presents the results from a power analysis, calculated using a simulation approach modeling the data-generating process with a conditional logit model. The power simulation varies the logit parameter for each quantity of interest while holding all others fixed, treating respondents' choices as random, and using a conditional logit model to calculate utilities based on the actual set of profiles assigned to respondents, clustering standard errors at the respondent-level, and recording the proportion of simulations in which the null hypothesis can be rejected at the $\alpha = 0.05$ level. Importantly, the analysis shows that the experiment is well-powered, retrieving AMCEs as small as 1-3% with 80% power.

Figure 6: Power analysis

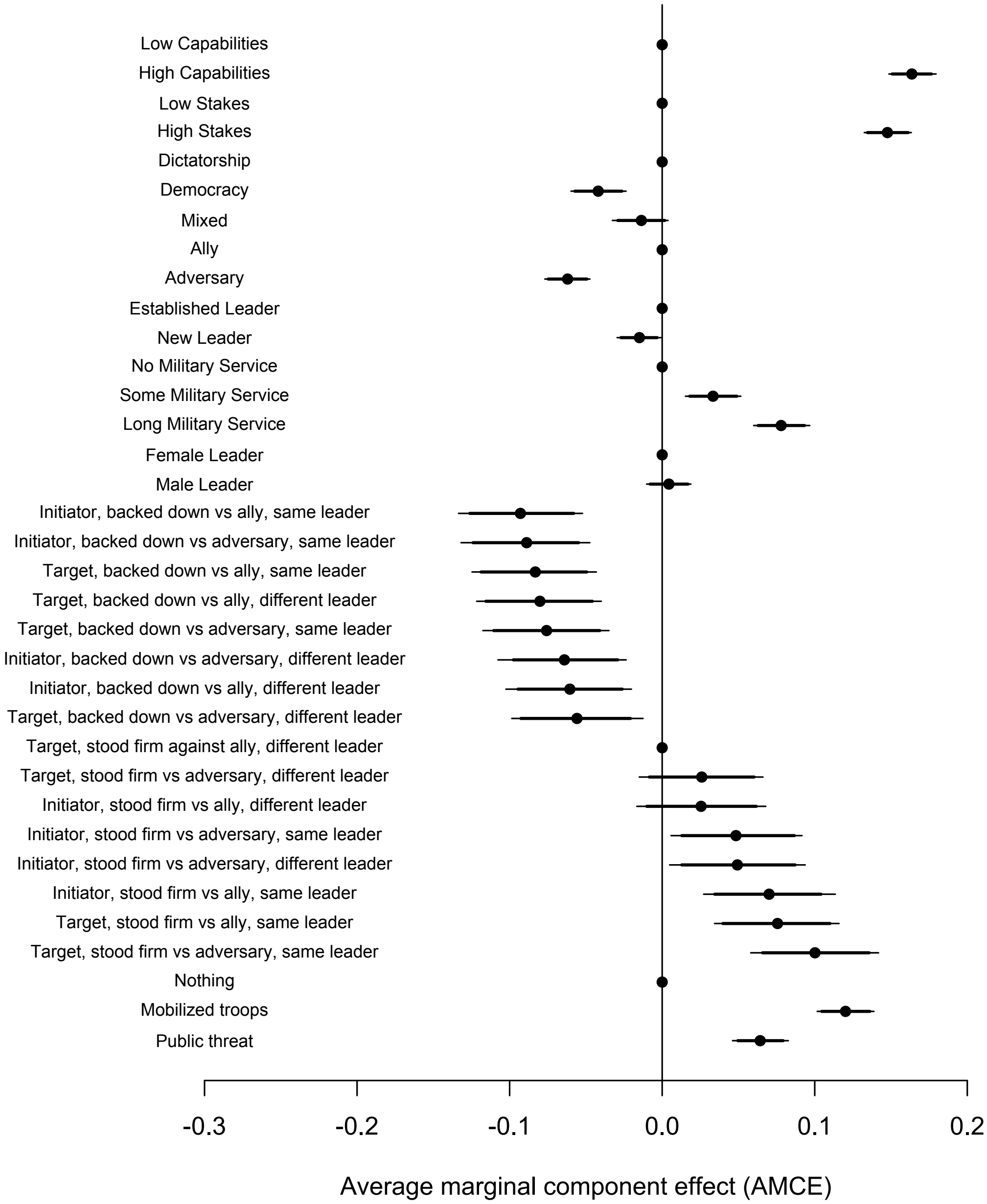


Power simulations for each of the factors manipulated in the experiment show the experiment is well-powered, recovering AMCEs as small as 1-3% with approximately 80% power, depending on the factor. Each panel represents simulations for a particular quantity of interest: for capabilities, it represents the effects of high capabilities, for stakes: high stakes, regime type: democracy, relationship with USA: adversary, new leader: new leader, military service: long military service, gender: male leader, role in previous crisis: initiator, opponent in previous crisis: adversary, past action: same leader backed down, costly signal: military mobilization.

1.3 Fully saturated interactions for past behavior

The analysis of the effects of past behavior in the main text presents interactions between past actions (stood firm versus backed down) and the identity of the leader responsible (a different leader than in the current dispute, versus the same leader as in the current dispute), but presents the other two past action treatments (whether the state was the target or initiator, and whether the opponent in the previous crisis was an ally or adversary of the US) as average effects rather than estimate a more complex four-way interaction. Figure 7 replicates the results from the main text, but this time with the fully-saturated four-way interaction. The results are harder to interpret because of the larger number of moving parts, but lends itself to the same substantive conclusion: behavior in a previous crisis is seen as more informative of present behavior if led by the same leader as in the current crisis than a different leader. In contrast, whether the state was the target or initiator, or was facing an ally or adversary of the US, does not appear to meaningfully interact with past actions in shaping assessments of resolve.

Figure 7: Estimating the full four-way interaction for past actions



1.4 Estimating heterogeneous treatment effects

1.4.1 By respondent characteristics

Figures 8-11 look for heterogeneous treatment effects across four different sets of characteristics: education (do respondents with a college degree rely on different cues than those without one?), militant internationalism (do hawks use different cues than doves?), partisanship (do Republicans use different cues than Democrats?), and interest in foreign affairs (do particularly engaged respondents employ different heuristics than respondents who are not as interested in international politics?).[2] The results find relatively little evidence of treatment heterogeneity: in Figure 8 the low-educated respondents (in grey) and high-educated respondents (in black) seem to use remarkably consistent heuristics: more educated respondents are slightly more pessimistic about democracies, for example, but the overall results remain the same. Similarly, in Figure 9, doves (in grey) and hawks (in black) also display strikingly similar results. Hawks appear less likely to give dictatorships the benefit of the doubt, and are more likely to see allies as reliable, but these differences are noticeably small given the vast differences between doves and haws we find in other areas of public opinion about foreign policy. We might expect stronger treatment heterogeneity with respect to militant internationalism if the United States was itself a participant in the disputes, rather than merely an observer.

In Figure 10, Republicans (in black) place slightly heavier emphasis on stakes than Democrats do, and interpret contextual factors relating to past actions slightly differently, but the overall configuration of results remains strikingly similar. Finally, although the confidence intervals around the point estimates in Figure 11 are wider for respondents with high levels of foreign policy interest (in black) than low levels of foreign policy interest (in grey), we see fairly similar results between the two: it appears that more interested respondents place a greater weight on military capabilities, and somewhat lesser weights on costly signals, but the differences for the latter are relatively small.

In general, then, there do not appear to be systematic differences across any of these characteristics: we never see that certain kinds of respondents are systematically more sensitive to leader-level variables rather than country-level ones, for example, or draw more information from current behavior and less from past behavior. The absence of these systematic differences is theoretically interesting, in that it shows a relatively uniform pattern of cue use across types of respondents, consistent with the notion of an "intuitive deterrence theory" articulated in the main text, in which

[2] We calculate the threshold for doves and hawks, and low and high interest in foreign policy by mean-splitting responses in the militant internationalism and foreign policy interest scales, shown in full below.

even those who think about world politics in fundamentally different ways nonetheless seem to place similar weights on the same set of indicators.

1.4.2 Average marginal treatment interaction effects (AMTIEs)

The analysis above estimates heterogeneous treatment effects with respect to respondent characteristics, but we can also look for interaction effects between the treatments themselves. Because interpreting cross-treatment interaction effects in conjoint experiments is sensitive to the choice of the baseline category, we follow Egami and Imai (2015) in presenting Average Marginal Treatment Interaction Effects (AMTIEs). Figures 12-13 thus present the full-range of one-way, two-way, and three-way AMTIEs for each of the treatment categories presented in the paper, letting us test for the possibility of higher-order interactive effects between sets of treatments. Importantly, just was we find relatively little evidence of heterogeneous treatment effects with respect to respondent characteristics, we also find relatively little causal heterogeneity between treatment combinations. Figure 12(a) reconfirms the findings from the main analysis, showing that the largest treatment effects are capabilities, stakes, past actions (operationalized here, as in the main text, based on whether the country backed down or stood firm in the previous crisis, conditioned on whether a different leader was in power at the time), and current behavior (costly signaling). In contrast, Figure 12(b) shows that the magnitude of the two-way effects are much smaller, suggesting we lose little by focusing solely on the ACMEs in the main text. Indeed, 30 of the 55 two-way AMTIEs have effect estimates of 0, and are thus omitted from the plot to save space. In Figure 13(a), the three-way AMTIE estimates are presented; not only are they extremely small, but only five estimates are presented, because the other 160 three-way AMTIEs have effect estimates of 0. Figure 13(b) illustrates the same pattern a different way, presenting the largest five one-way, two-way and three-way AMTIEs, showing once again that we can safely ignore higher-level interactions. In this sense, the weak higher-order interactions also reconfirm the lack of evidence in these experimental results in favor of the the current calculus and attribution theory hypotheses, both of which posit interactions between past actions and other variables. Finally, these findings also extend those from the previous subsection, in that individuals seem to anchor on the same cues regardless of the specific combination of treatments presented: it is not that particular higher-order interactions make capabilities and interests matter any less, for example. This offers further evidence in support of our “intuitive deterrence theory” model.

Figure 8: Heterogeneous treatment effects: low-educated (grey) versus high-educated (black)

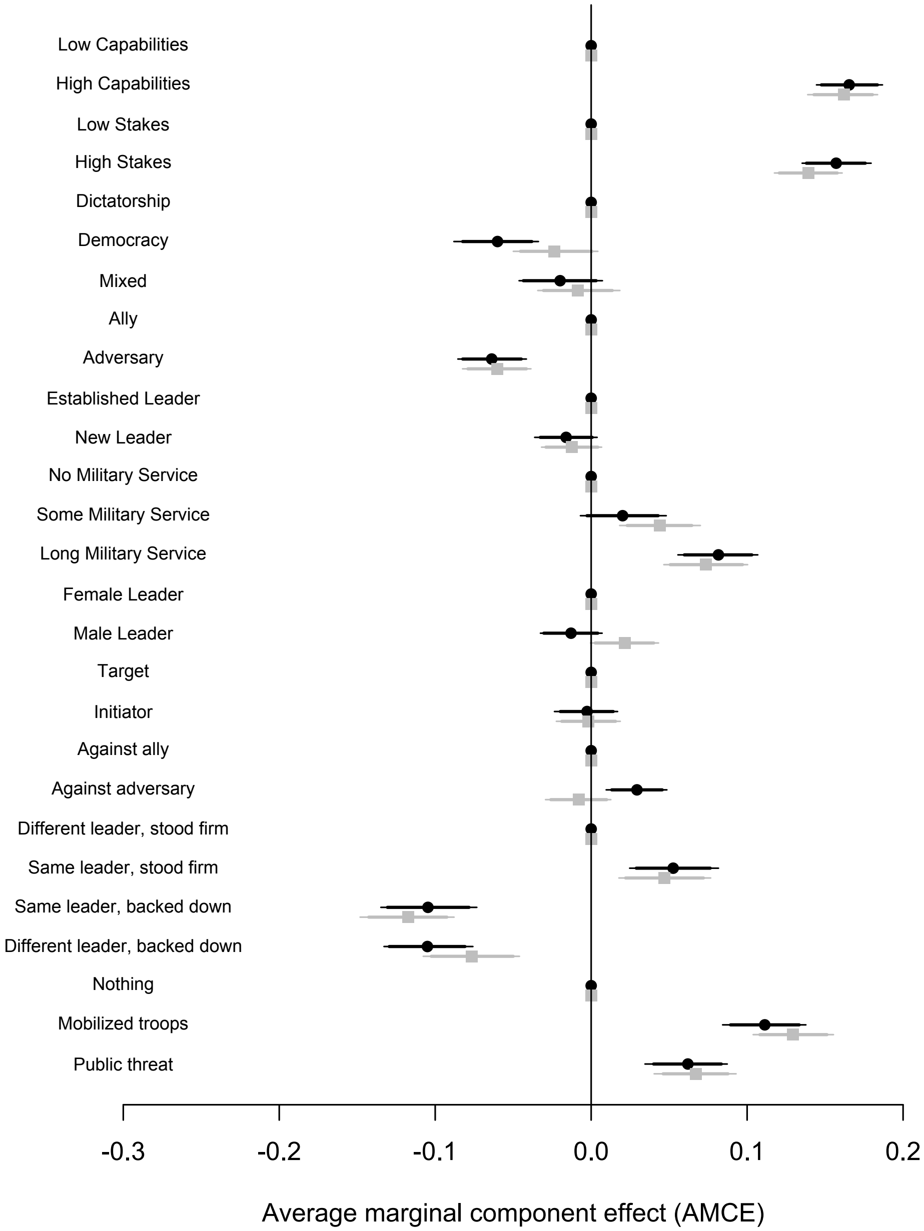


Figure 9: Heterogeneous treatment effects: hawks (black) versus doves (grey)

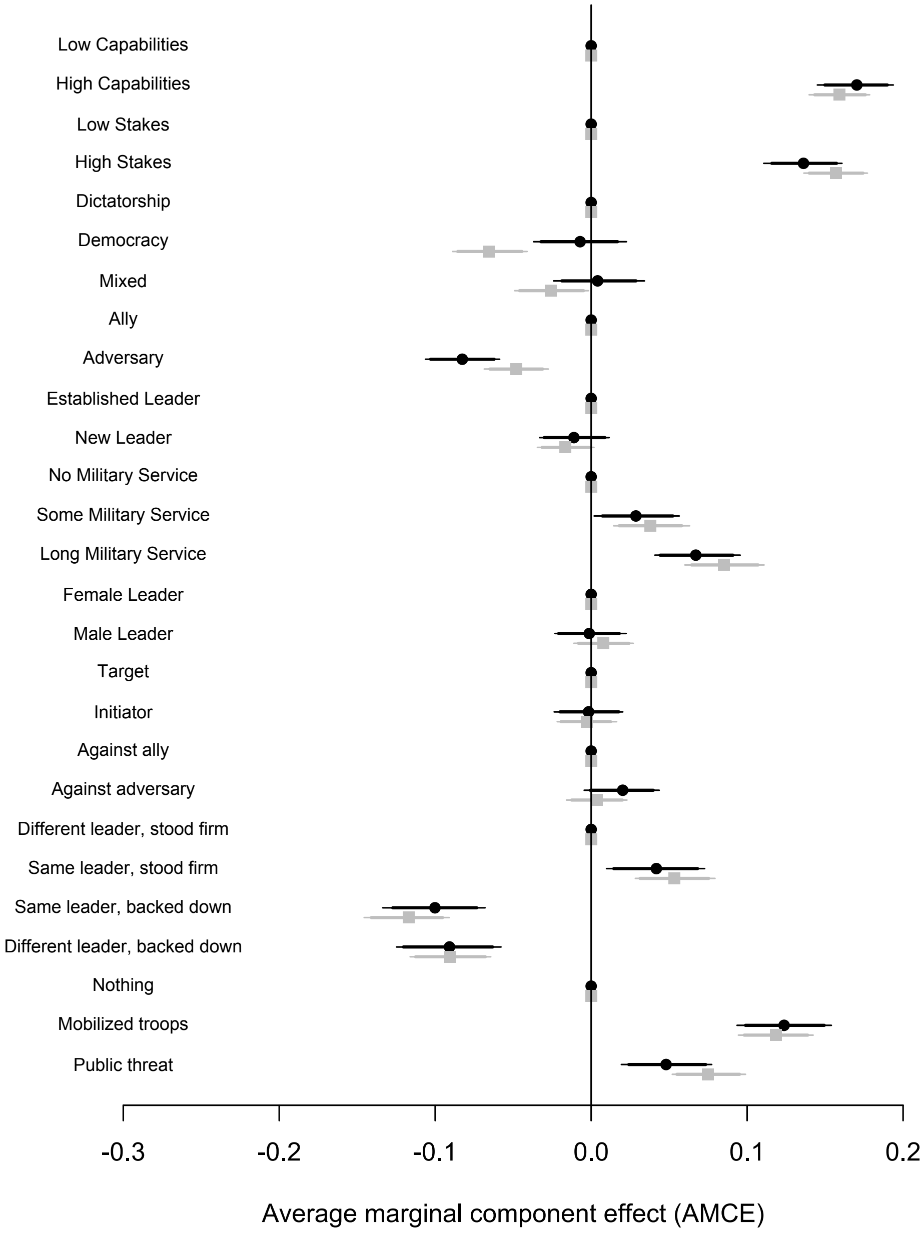


Figure 10: Heterogeneous treatment effects: Democrats (grey) versus Republicans (black)

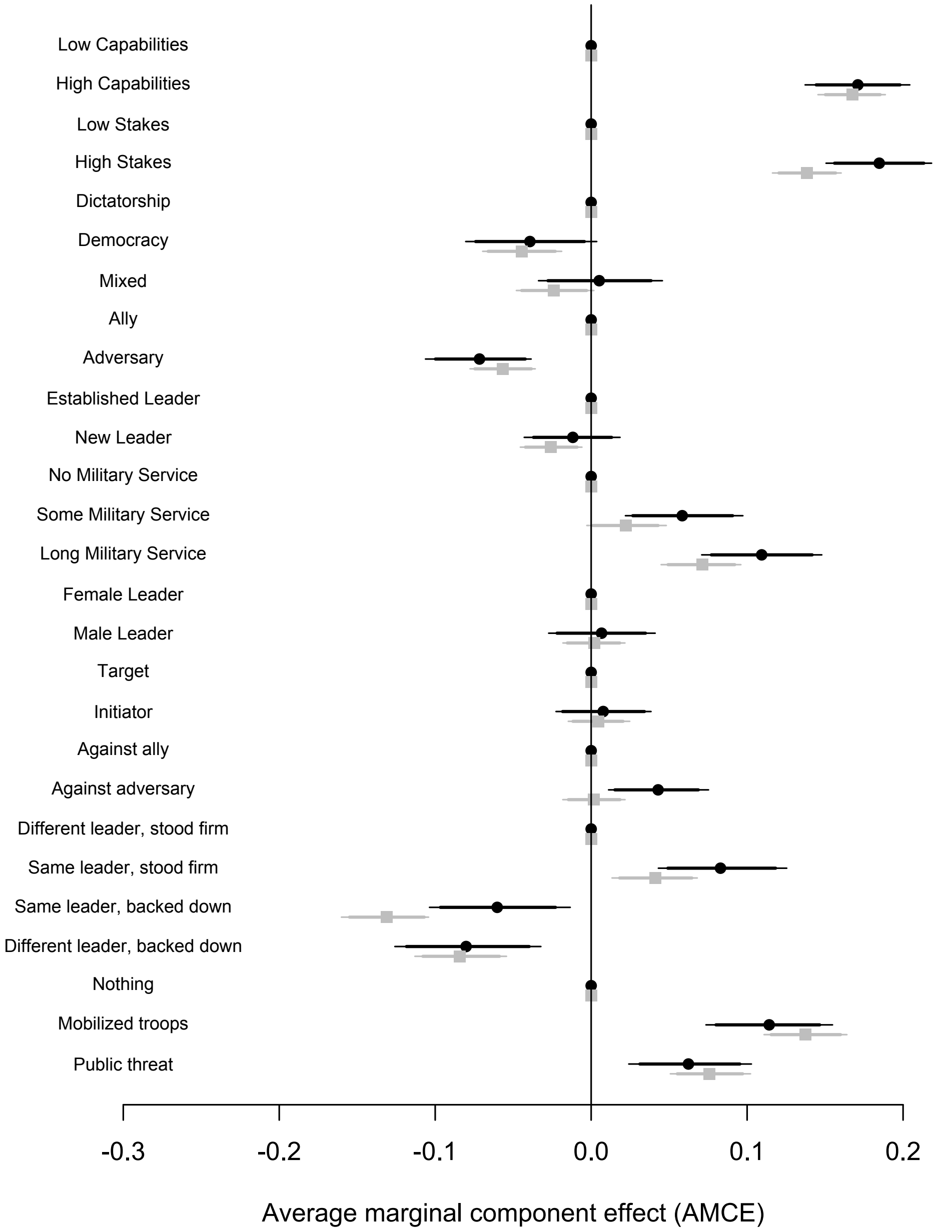


Figure 11: Heterogeneous treatment effects: low foreign policy interest (grey) versus high foreign policy interest (black)

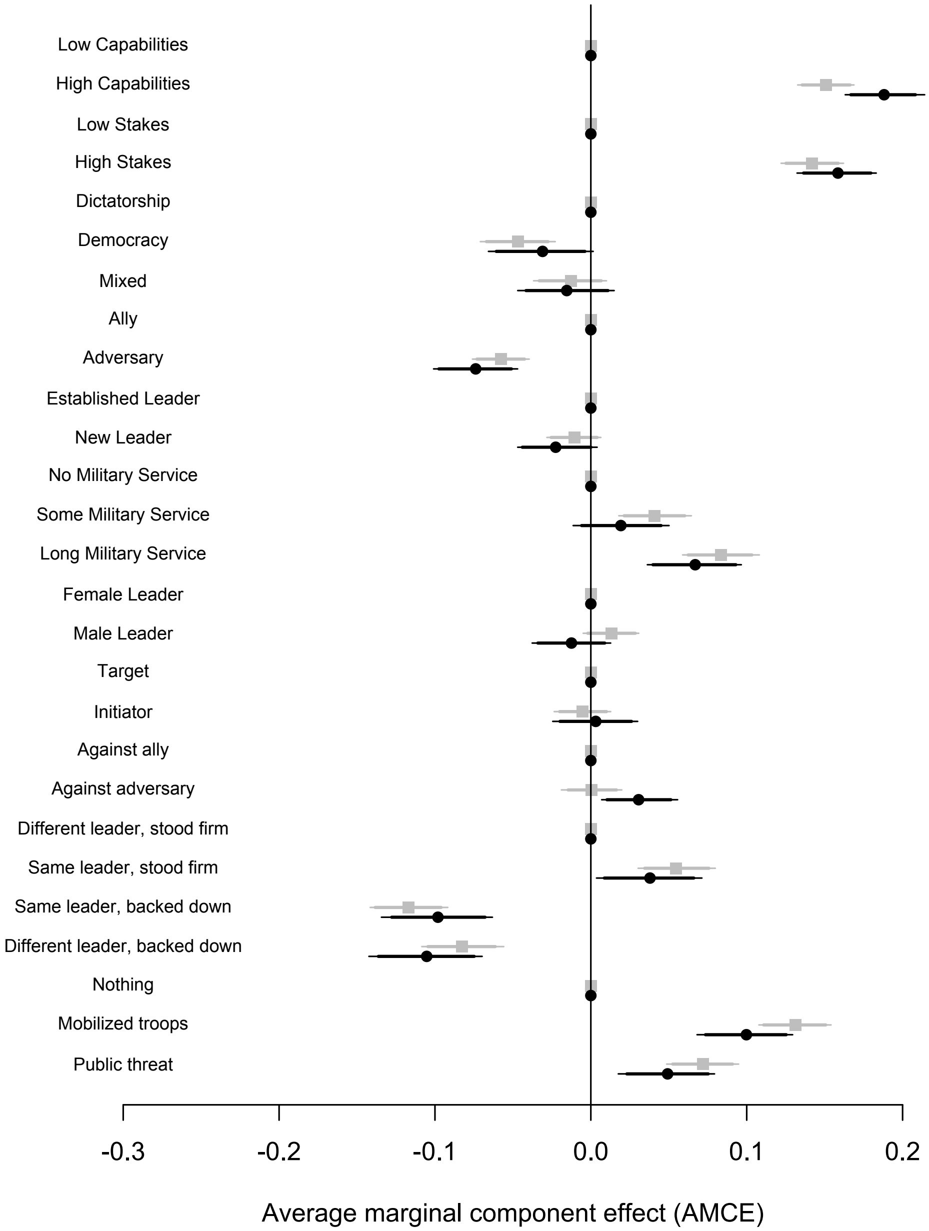
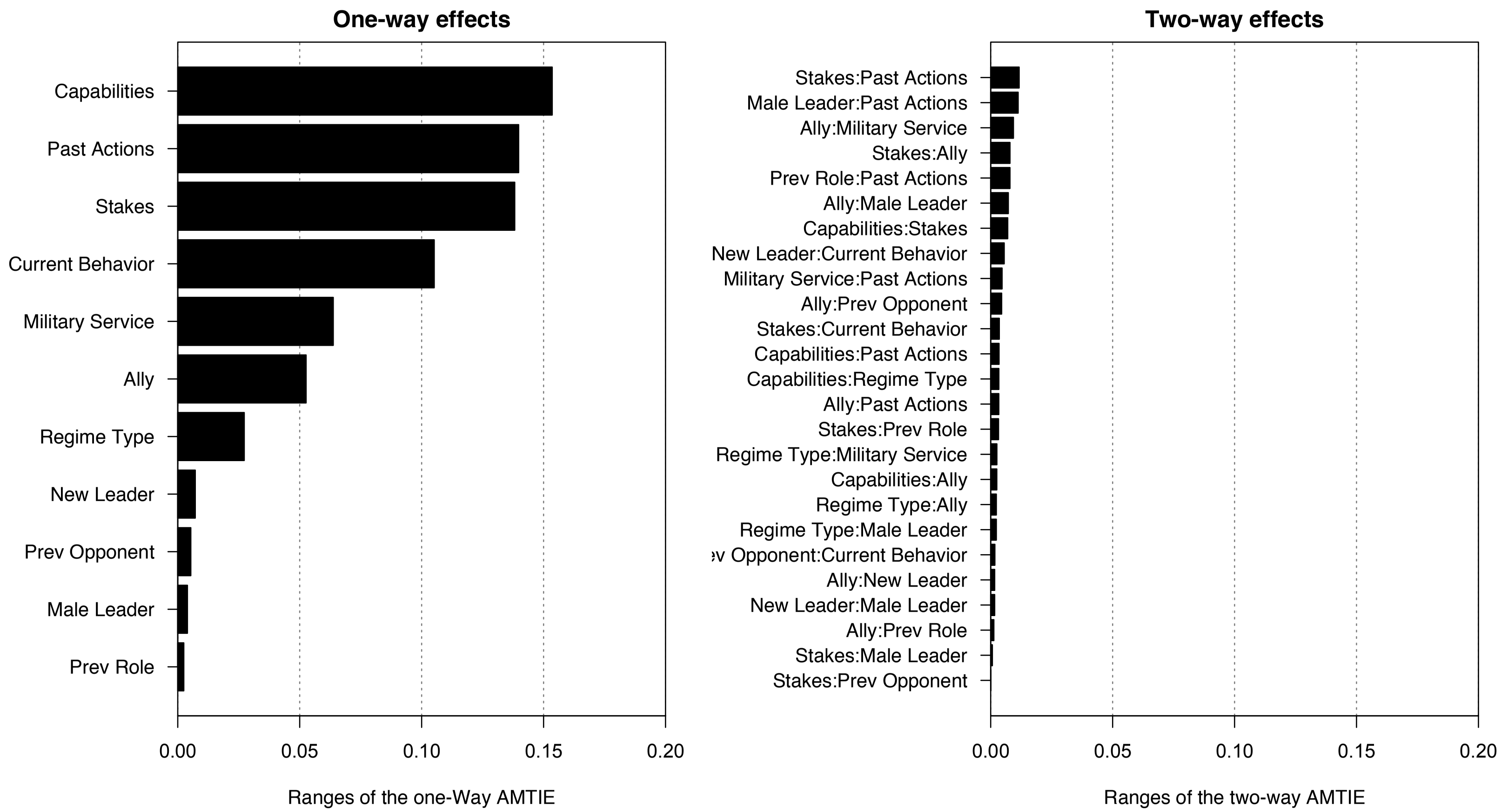


Figure 12: Estimated Ranges of AMTIEs (I)

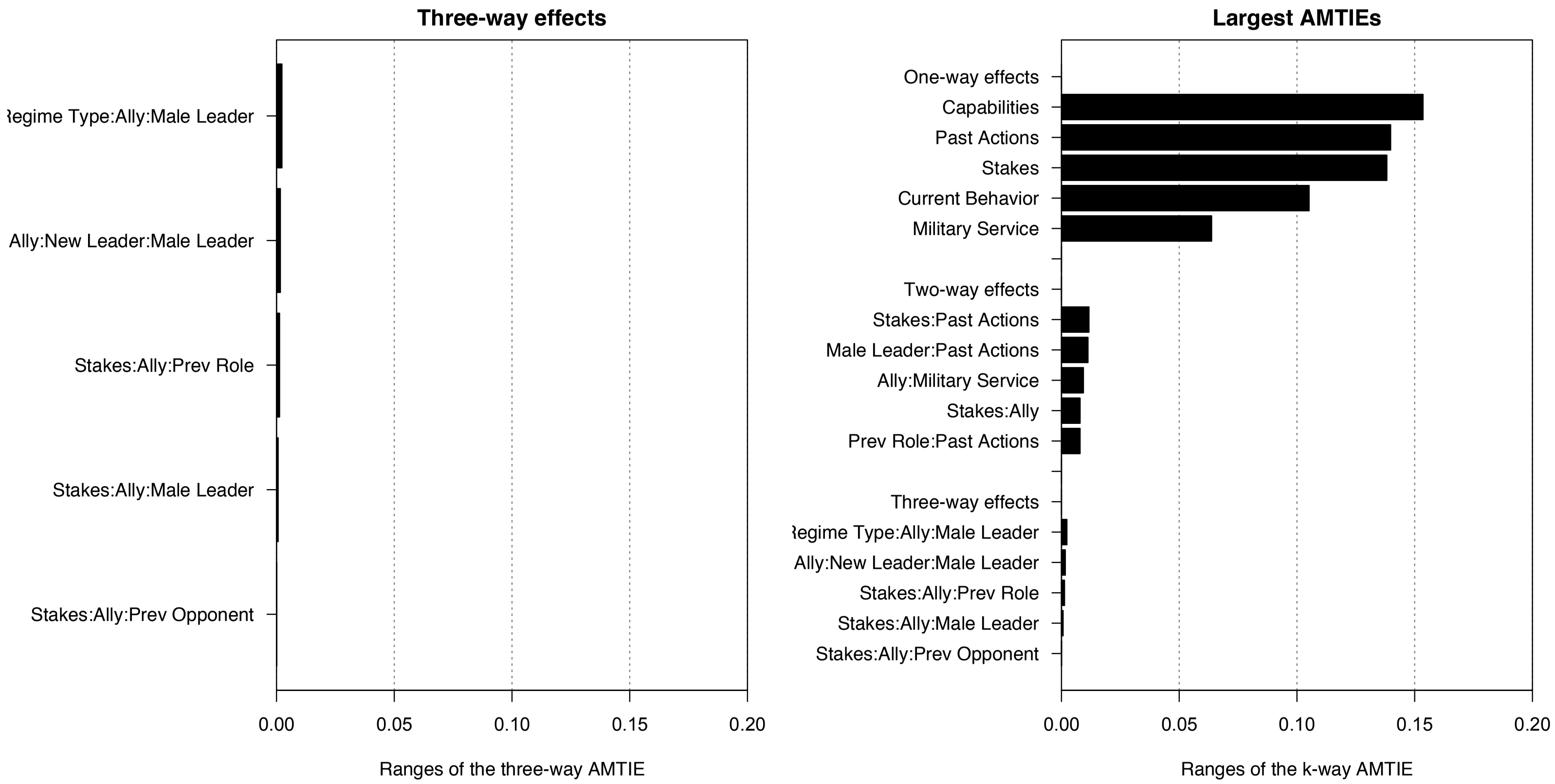


(a) One-way AMTIEs

(b) Two-way AMTIEs

Note: Panel (a) reconfirms the importance of capabilities, stakes, past actions, and current behavior (costly signaling), while other variables exert relatively weak effects. Panel (b) shows little evidence of significant two-way AMTIEs. 30 of the 55 two-way AMTIEs have effect estimates of 0, and are omitted here to save space.

Figure 13: Estimated Ranges of AMTIEs (II)



(a) Three-way AMTIEs

(b) Largest AMTIEs

Note: Panel (a) shows little evidence of significant three-way AMTIEs. 160 of the 165 three-way AMTIEs have effect estimates of 0, and are omitted here to save space. Panel (b) shows the five largest AMTIEs of each type.

1.5 Response latency results

The results presented above suggest the presence of a fairly ubiquitous mental model. Appendix §1.3.1 showed that when it comes to assessing resolve, participants with more education rely on a similar portfolio of cues as those with less education, participants more interested in foreign affairs assess resolve using a similar set of indicators as participants less interested in foreign affairs, and hawks utilize a similar set of cues as doves; Appendix §1.3.2 showed a similar absence of interactions between treatments, such that the effects of the cues with the largest effects in our results (capabilities and stakes, for instance) do not seem to systematically vary based on the presence of particular combinations of treatments. This absence of contingence is striking, in that it paints a picture of respondents as sharing a common schema (which in the main text we call "intuitive deterrence theory") in which participants focus their attention on a particular set of cues (particularly capabilities and stakes) to resolve the ill-structured problem of assessing resolve in disputes. Although the strong effects of these AMCEs in the experimental results document this pattern nicely, another way of further confirming it involves the use of response times.

The logic of response latency analysis (Mulligan et al., 2003) is straightforward: the question of how observers assess resolve is a ultimately a question of how individuals process information, and one way scholars of political behavior indirectly get at information processing involves looking at the speed at which individuals take to answer a question after it has been asked (e.g. Bolsen, Druckman and Cook, 2014). Although response latencies are inherently noisy measures, if on average respondents presented with certain combinations of treatments take systematically longer or shorter to respond than respondents presented with other combinations of treatments, it potentially sheds light on the cognitive mechanisms under investigation, whether processing efficiency or attitude accessibility (Fazio, 1990). For our purpose, two tests are potentially instructive. One would involve comparing how average response latency changes as the number of factors manipulated changes. For example, if participants assess resolve more quickly when capabilities information is presented than when it isn't (even though the amount of text participants are being presented with would actually increase!), this would offer suggestive evidence that participants rely on capabilities cues when assessing resolve. In the case of our experimental design, because none of our factors being manipulated has a pure control, we are unable to use such an estimation strategy here.[3] Instead, we use a different test, exploiting the choice-based nature of our conjoint design to test for the effects of

[3] For a broader discussion of this type of approach, see Acharya, Blackwell and Sen (2017).

conjunctive versus disjunctive treatment assignments between the two randomly generated profiles.

For purposes of simplicity, suppose four choice tasks, each between two country profiles, illustrated in Table 3.

Table 3: Hypothetical treatment assignments

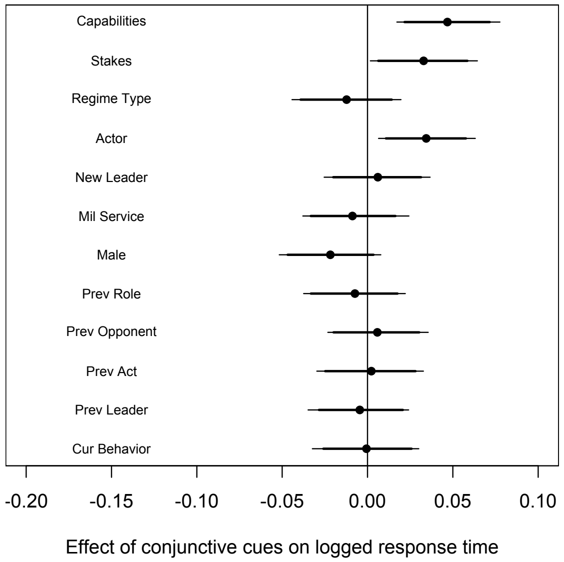
	Task 1		Task 2		Task 3		Task 4	
Choice	A	B	A	B	A	B	A	B
Factor 1	0	1	1	0	0	1	1	0
Factor 2	1	0	0	0	0	1	1	1
Factor 3	1	0	1	0	1	0	0	1
Factor 4	1	0	0	1	1	0	1	0

For each choice task, each choice profile consists of four randomly generated factors, each with multiple levels. For Task 1, the treatment assignments between each profile are *disjunctive*, in the sense that each of the four factors are assigned to different levels between country A and country B. The same is true for Task 3. For Task 2, however, the treatment assignment for factor 2 is *conjunctive*, in that the same level of the treatment is assigned to both choices in the set. The same is true for Task 4. By comparing how the average response latency differs between the conjunctive and disjunctive choice sets for each treatment (e.g. for factor 2, a comparison of the average response latency for tasks 1 and 3, versus tasks 2 and 4), we thus have a sense of how much respondents anchor on a particular factor as part of their decision-making process. For example, in a model of consumer choice, if individuals take much longer to decide between a pair of consumer goods if both options have the same price, this implies that price is an important consideration. The same holds here: if capabilities or stakes are important considerations, participants should assess resolve more quickly when given disjunctive capability or stake treatments than when given conjunctive ones.

In Figure 14, we present point estimates and 95% clustered bootstrapped confidence intervals from a regression model where each participant's logged response time is modeled as a function of conjunctive treatment assignments for each of the treatments from the conjoint design; each point estimate thus depicts a different coefficient estimate from the model. The plot shows that conjunctive cues for capabilities and stakes significantly increases response time, suggesting that a balance of power or balance of interests induces respondents to take longer when assessing resolve.[4] Interest-

[4] Additional analysis confirms this interpretation: the average response time when both countries have low capabilities is indistinguishable from when they both have high capabilities, but the average response time when one country has high capabilities and the other country has low is approximately two seconds faster, showing how an imbalance of power accelerates the assessment process. A similar pattern emerges with respect to stakes: the average response time when both countries have low stakes is similar to when they both have high stakes, but the average response time when one country has high stakes and the other country has low stakes is roughly 1.5 seconds faster.

Figure 14: Response latency effects



ingly, the other factor that displays conjunctive cue effects is the identity of the actor (an adversary or ally of the United States); this finding is striking given that the AMCE for the actor treatment is significant but relatively modest in size. The significant results with respect to logged response time are perhaps due to the centrality and automaticity of social categorization (Kurzban, Tooby and Cosmides, 2001) and importance of coalitional psychology (Lopez, McDermott and Petersen, 2011): even if people ultimately don't consider whether an actor is an ally or adversary to be a particularly strong indicator of how much resolve an actor will display in a crisis, participants take roughly 1.5 seconds faster to assess resolve in disputes between a member of the ingroup and a member of the outgroup than they do in disputes between outgroup members or between ingroup members.

2 Dispositional Instrument

As noted in the main text, in addition to the choice-based conjoint experiments, participants also completed a battery of dispositional and demographic measures. To avoid downstream effects, participants were randomly assigned to receive the dispositional questionnaire either before completing the conjoint tasks, or afterwards. The instrumentation used below is relatively standard in the public opinion about foreign policy literature, based off of classic work by Holsti and Rosenau (1988) and Wittkopf (1990); see, e.g. Kertzer et al. (2014).

Unless otherwise specified, all response options below are scaled from "strongly agree" (1) to "strongly disagree" (5).

2.1 Militant Internationalism

1. The best way to ensure world peace is through American military strength
2. The use of force generally makes problems worse [reverse-coded]
3. Rather than simply reacting to our enemies, it's better for us to strike first
4. Generally, the more influence America has on other nations, the better off they are

2.2 Cooperative internationalism

1. America needs to cooperate more with the United Nations in settling international disputes
2. It is essential for the United State to work with other nations to solve problems such as overpopulation, hunger and pollution

2.3 Isolationalism

1. The U.S. needs to play an active role in solving conflicts around the world [reverse-coded]
2. The U.S. government should just try to take care of the well-being of Americans and not get involved with other nations

2.4 International trust

1. Generally speaking, would you say that the United States can trust other nations, or that the United States can't be too careful in dealing with other nations? [the United States can trust other nations/ the United States can't be too careful in dealing with other nations]

2.5 Demographic Questions

1. What is your gender? [male/female]

2. What year were you born? [open text box]

3. What is the highest level of education you have completed [less than high school/ high school or GED/ some college/ 2 year college degree/ 4 year college degree/ Master's degree/ Doctoral degree/ Professional degree (e.g., JD or MD)]

4. Generally speaking, do you usually think of yourself as a Republican, Democrat, or as an independent (check the option that best applies)? [Strong Republican/ Republican/ Independent, but lean Republican/ Independent/ Independent, but lean Democrat/ Democrat/ Strong Democrat]

5. Below is a scale on which the political views that people might hold are arranged from "extremely conservative" to "extremely liberal." Where would you place yourself on this scale? [extremely conservative/ conservative/ slightly conservative/ moderate/ slightly liberal/ liberal/ extremely liberal]

6. How interested are you in information about what's going on in foreign affairs? [extremely interested/ very interested/ moderately interested/ slightly interested/ not interested at all]

7. How interested are you in information about what's going on in government and politics? [extremely interested/ very interested/ moderately interested/ slightly interested/ not interested at all]

3 Conjoint Instrument Screen

In this portion of the study, we will present you with information about a series of eight foreign policy disputes between different countries.

Countries often get into disputes over contested territories. These disputes receive considerable attention because of the risk they can escalate to the use of force. Thus, the kinds of disputes described here are ones that have occurred many times, and will likely occur again.

In each screen, we will present you with a pair of countries involved in a territorial dispute, tell you a bit about each of them, and ask you to make predictions about what you think will happen. There are no right or wrong answers, we're simply interested in the kinds of predictions you make.

4 Sample and weighting information, and demand effects

As noted in the main paper, the study was fielded on a sample of 2009 American adults recruited via Amazon Mechanical Turk in January 2015. Participants were paid $1 for their participation. One potential concern about the use of MTurk is the representativeness of the sample, as compared to other potential survey platforms. However, Berinsky, Huber and Lenz (2012, 366) show that MTurk samples are "often more representative of the general population and substantially less expensive to recruit" than other "convenience samples" often used in political science (see Huff and Tingley, 2015, for the latest and most definitive work on this subject).[5] They also demonstrate the ability to replicate results from nationally-representative samples — e.g., Kam and Simon's (2010) work on framing and risk and Tversky and Kahneman's (1981) classic "Asian Disease problem" — using MTurk workers.[6] As a result, MTurk is becoming increasingly widely used in experimental political science, and experiments using MTurk samples have been published in a variety of notable journals, including the *American Political Science Review* (Tomz and Weeks, 2013), the *American Journal of Political Science* (Healy and Lenz, 2014; Levy et al., 2015; Bishin et al., 2016; Huff and

[5] Though compared to nationally representative samples, MTurk workers tend to be younger and more ideologically liberal.

[6] While there has been some initial wariness regarding online experiments, many famous and well-known behavioral studies have been replicated using MTurk. For more on this, see Mason and Suri (2010); Buhrmester, Kwang and Gosling (2011); Rand (2012). For a different viewpoint, see Krupnikov and Levine (2014), though their caution applies only to very specific kinds of experimental studies, as we discuss below.

Kertzer, 2018), *International Organization* (Wallace, 2013), and the *Journal of Conflict Resolution* (Kriner and Shen, 2014). In keeping with "best practices" suggested by numerous researchers, we limited participation in the study to MTurk workers located in the United States, who had completed ≥ 50 HITs, and whose HIT approval rate was >95%.

Table 4: Sample characteristics

Sample Characteristic	Unweighted Sample	Weighted Sample	Population Target
Male	0.539	0.516	0.492
18 to 24 years	0.132	0.146	0.128
25 to 44 years	0.680	0.397	0.342
45 to 64 years	0.167	0.342	0.341
65 years and over	0.020	0.114	0.189
High School or less	0.109	0.343	0.420
Some college	0.287	0.223	0.194
College/University	0.491	0.325	0.282
Graduate/Professional school	0.113	0.109	0.104

As Huff and Kertzer (2018) note, there are typically two concerns about the use of MTurk. The first involves the composition of the sample. As noted above, MTurk samples, although more diverse than most convenience samples used in political science, are nonetheless not nationally representative of the US population as a whole, tending to skew somewhat younger and more educated (Huff and Tingley, 2015). To address this point, we employ a two-pronged strategy. First, like Huff and Kertzer (2018), we use entropy balancing (Hainmueller, 2012) to reweight our sample towards national population parameters, trimming the weights to reduce the impact of extreme values. Table 4 presents the weighted and unweighted sample characteristics, showing that the weighted data hews closely towards population targets. Figure 1 in the main text presents both the weighted and unweighted AMCEs, showing that the effects themselves do not significantly differ between the two datasets. Second, we present a series of models testing for heterogeneous treatment effects (Figures 8-11), which include a number of additional characteristics where we might expect our sample to differ but which are not explicitly being accounted for in the reweighting. For example, if participants recruited on MTurk also happen to be systematically more liberal, or more interested in politics, than the population at large, it is helpful to test whether more conservative participants, or less politically engaged respondents, rely on systematically different heuristics. As the results in Appendix §1.3.1 show, we fail to find evidence that this is the case.

A second concern about MTurk might involve the potential for demand effects (Chandler, Mueller and Paolacci, 2014), the tendency for experimental participants to decipher the purpose of the study, and act in a way either consistent with the experimenters' wishes (e.g. *experimenter bias* - Orne, 1962; Rosenthal and Fode, 1963) or contrary to them (e.g. the *screw-you effect* - Masling, 1966). As is the case with participants in other online survey platforms, MTurk users often participate in a large number of studies, which not only raises concerns that study participants might be more susceptible to demand effects, but raises particular concerns for studies that require naive participants that have not encountered a particular experimental paradigm before (Paolacci and Chandler, 2014; Krupnikov and Levine, 2014; Huff and Kertzer, 2018). Given the purpose of our study, these concerns are mitigated here. First, in an information-rich conjoint experiment with countervailing indicators, demand effects are less relevant than they might be in a less elaborate experimental design; it is presumably easier to determine which factor the experimenters are interested in in an experiment that manipulates one factor than an experiment that manipulates seventeen of them. Second, the within-subject component of conjoint designs minimizes the need for naive participants, since participants are being exposed to multiple treatment conditions across multiple rounds. Third, as noted in Appendix §1, we can test for demand effects explicitly with conjoint designs by validating the stability and carryover effect assumption. If demand effects are present, we should expect this assumption to be violated, as participants decipher the goals of the experimenters by taking multiple rounds of the experiments, and thus respond to the treatments differently over time. Instead, the results reported in Appendix §1 suggest the AMCEs are relatively stable over time, such that participants do not appear to respond any differently to the treatments in the last round as they do in the first. Thus, we have little reason to be concerned about demand effects here.

5 Elite experimental benchmarks

Below we present additional information about the pair of elite experiments we use as benchmarks to evaluate the mass public results, drawn from a sample of 89 current and former members of the Israeli Parliament (i.e., the Knesset) from the beginning of the 14th Knesset in June 1996 through the 20th Knesset, sworn in in March 2015. Each experiment is itself the focus of a separate paper; readers interested in additional information about the recruitment and participant verification protocol, see Yarhi-Milo, Kertzer and Renshon (2018).

We present the complete instrumentation for each experiment below (translated from Hebrew to English), as well as a set of supplementary analyses. Table 5 presents basic descriptive statistics for our Knesset sample, showing that the sample is unusually "elite" by the standards of many experiments conducted in international relations. For example, two-thirds of the sample has experience on the foreign affairs and defense committee, and over 40% served as deputy minister or higher.[7] Table 6 tests for the representativeness of our elite respondents, in two different ways. First, it compares our respondents to the complete population of individuals who served in the Knesset from 1996 to 2015. Second, it compares our respondents to the sampling frame (a measure of non-response, since it excludes MKs who had passed away, etc.) The results show that, perhaps unsurprisingly, current members of the Knesset were less likely to participate in the survey than former members, but that importantly, our participants are not significantly less "elite", and if anything, are slightly more experienced than the universe of decision-makers. Table 7 provides balance checks for the two elite benchmark experiments, showing the experiments are well-balanced along a host of demographic characteristics.

Finally, one of the patterns in Figure 5 in the main paper is that while the results from our mass public conjoint experiment are similar to the results from the elite benchmark experiments, the ATE distributions in the elite sample have much heavier tails. There are two potential interpretations of this pattern worth differentiating. In one, the wider spread of the distributions in the elite sample is simply due its relatively small sample size. In another, the wider spread of the distributions in the elite sample is due to systematic differences between elites and masses, or perhaps to divergent

[7] There is minimal missingness in the data: all 89 respondents participated in the costly signaling experiment, and 88 of 89 in the regime type experiment. Of the demographic and dispositional variables from Table 5, the only demographic variable with any missingness is military experience (84/89 completed); of the dispositional variables, 89/89 completed military assertiveness, 87/89 competed ideology, 87/89 completed attitudes towards the Arab-Israeli conflict, and 86/89 completed international trust.

experimental designs between the two studies. If this alternate explanation were the case, we would expect that if the mass public sample was a similarly small size as the elite samples, the distributions would look similar to one another. We test this assumption in Figure 15, which downsamples the mass public sample so as to match the elite sample in terms of size; the results show that the greater spread in the ATE distributions for the elite results in the main paper are due to the relatively small sample size rather than oddities about the sample.

Table 5: Knesset Sample Characteristics (N=89)

	Proportion of respondents
Knesset Member:	
Current	25%
Former	75%
Exp. on Foreign Affairs/Defense Committee . . .	
. . . as backup or full member	67%
. . . as full member	54%
Highest level of experience:	
. . . not a Minister	58%
. . . Deputy Minister	29%
. . . Cabinet Member or higher	12%
Male	84%
Served in military	95%
Active combat experience	64%

	Mean	SD
Age	61.4	10.7
Terms in Knesset	3.0	2.1
Military Assertiveness	0.61	0.20
Right Wing Ideology	0.45	0.24
Hawkishness (Arab-Israeli conflict)	0.39	0.25
International Trust	0.40	0.25

Note: individual differences in bottom four rows scaled from 0-1.

Table 6: Sample representativeness tests

	Compared to…			
	All Knesset members		Sampling frame	
	(1)	(2)	(3)	(4)
Current member	−0.043	−0.049	-0.210^{***}	-0.184^{***}
	(0.045)	(0.057)	(0.054)	(0.065)
Highest level of experience:				
… Deputy minister	0.017	0.044	0.035	0.079
	(0.054)	(0.071)	(0.072)	(0.088)
… Cabinet member or higher	−0.044	−0.098	−0.075	−0.096
	(0.076)	(0.098)	(0.093)	(0.114)
Male	0.025	0.081	0.072	0.097
	(0.053)	(0.063)	(0.067)	(0.076)
Terms in office	0.011	0.021	0.008	0.013
	(0.012)	(0.016)	(0.015)	(0.018)
Left-right party membership		-0.070^{**}		−0.063
		(0.030)		(0.038)
Constant	0.177^{***}	0.312^{***}	0.320^{***}	0.436^{***}
	(0.054)	(0.087)	(0.070)	(0.108)
N	415	295	288	225
R^2	0.007	0.043	0.063	0.080

*p < .1; **p < .05; ***p < .01

Table 7: Elite experiment balance checks

	Regime type:		Costly signal:	
	Dictatorship	Democracy	Mobilization	Threat
Current member	0.18	0.32	0.23	0.27
Foreign affairs experience	0.67	0.68	0.64	0.71
Highest level of experience	0.58	0.50	0.48	0.60
Male	0.87	0.82	0.82	0.87
Age	62.24	60.48	61.91	60.84
Active combat experience	0.62	0.67	0.66	0.63
Military assertiveness	0.59	0.62	0.58	0.63
Right wing ideology	0.44	0.46	0.41	0.49
Hawkishness (Arab-Israeli conflict)	0.38	0.40	0.36	0.41
International Trust	0.39	0.40	0.41	0.38

Figure 15: Comparing our downsampled mass public results with those from elite benchmarks

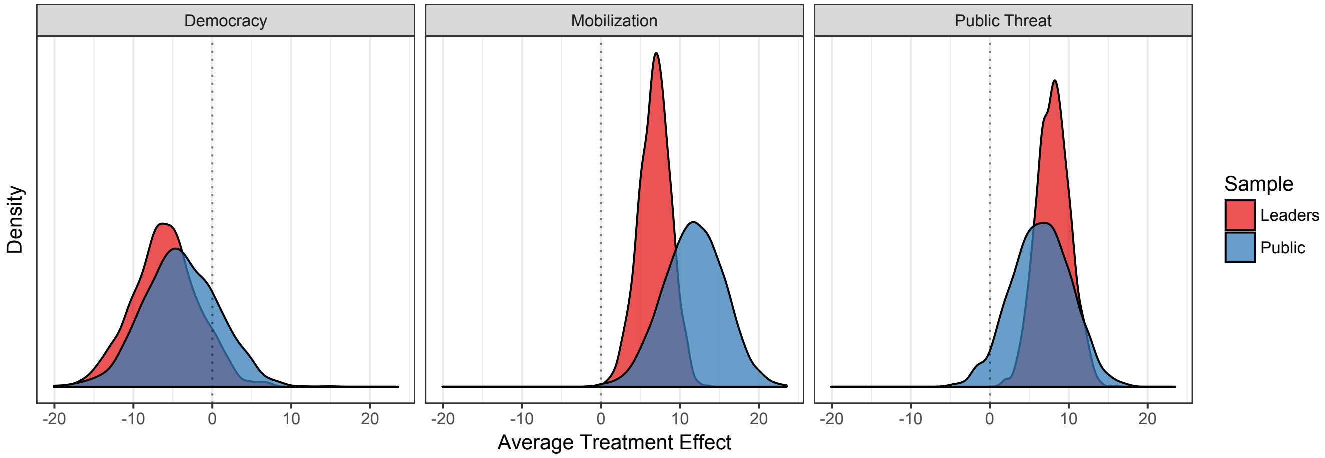


Figure 15 compares the bootstrapped average treatment effects for three factors from the conjoint experiment, in blue (calculated using B=2500 clustered bootstraps), with the bootstrapped average treatment effects for the same three factors in a pair of survey experiments fielded on an elite sample of members of the Israeli Knesset, in red (calculated using $B = 2500$ bootstraps). The results for regime type are shown in the left-hand panel, and results for costly signals in the middle and right-hand panels. Collectively, they show that the heavier tails in the ATE distributions for the Knesset results in the main paper is due to the relatively small sample size rather than oddities about the sample; when we downsample the public results, we find the distributions have similar spread. For downsampling the public results for regime type in the left-hand panel, we sample $N = 89$ observations with replacement; because of the costly signaling experiment's between and within-subjects design (in which all participants are also administered the control condition, unlike in the public conjoint experiment), we sample $N = 135$ observations from the public sample with replacement, to ensure that in expectation there are as many sampled observations in each costly signaling condition in the public sample as there would be in its elite counterpart.

5.1 Experiment Instrumentation

5.1.1 Regime Type Experiment

Here is the situation:

- Two countries are currently involved in a public dispute over a contested territory. The dispute has received considerable attention in both countries, because of the risk that disputes like these can escalate to the use of force.
- Country **A** is a [democracy/dictatorship]. Country **B** is a dictatorship.
- Both countries have moderately powerful militaries, with large armies, moderate sized-navies, and well-trained air forces.
- Neither country is a close ally of the United States.
- Country **A** is slightly larger than Country **B**, though their economies are approximately the same size.
- Country **A** has moderate levels of trade with the international community. Country **B** has high levels of trade with the international community.
- The last time the two countries were involved in an international dispute, different leaders were in power.

1. Given the information available, what is your best estimate about whether Country A will stand firm in this dispute, ranging from 0% to 100%?
2. If the dispute were to escalate and war were to break out, what is your best estimate about whether Country A will win, ranging from 0% to 100%?

5.1.2 Costly Signals Experiment

Here is the situation:

- Your country — Israel — is involved in a dispute with Country B, a strong military dictatorship.
- The dispute began with a collision between an Israeli shipping vessel and a ship registered to Country B.

- During the collision, injuries were reported on both sides.
- Additionally, both countries maintain that their ship was carrying sensitive military technology, and are suspicious of the motives of the other side, leading to a tense standoff at sea.
- Currently, because of the remote location, the public is not aware of the incident.

[*Outcome 1 (Baseline)*] Given the information available, what is your best estimate about whether Country B will stand firm in this dispute, ranging from 0% to 100%?

[NEW SCREEN]

Now we would like to ask you a question about a different, alternative version of the scenario you just read. Suppose the basic details remain the same:

- Israel is involved in a dispute with a dictatorship with a strong military, Country B.
- The dispute began with a collision between an Israeli shipping vessel and a ship registered to Country B. During the collision, injuries were reported on both sides.
- Both countries maintain that their ship was carrying sensitive military technology, and are suspicious of the motives of the other side, leading to a tense standoff at sea.
- Currently, because of the remote location, the public is not aware of the incident.

But this time, suppose that...

×

[Tying Hands]: The President of Country B has issued a public statement through the news media warning that they will "do whatever it takes" to win this dispute.
[Sinking Costs]: Country B has mobilized their military and sent additional gunboats to the location of the dispute at sea.
Outcome 2 (Treatment)] Given the information available, what is your best estimate about whether Country B will stand firm in this dispute, ranging from 0% to 100%?

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. Forthcoming. “Analyzing Causal Mechanisms in Survey Experiments.” *Political Analysis* .

Ballard-Rosa, Cameron, Lucy Martin and Kenneth Scheve. 2017. “The Structure of American Income Tax Policy Preferences.” *Journal of Politics* 79(1):1–16.

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(3):351–368.

Bishin, Benjamin G., Thomas J. Hayes, Matthew B. Incantalupo and Charles Anthony Smith. 2016. “Opinion Backlash and Public Attitudes: Are Political Advances in Gay Rights Counterproductive?” *American Journal of Political Science* 60(3):625–648.

Bolsen, Toby, James N. Druckman and Fay Lomax Cook. 2014. “The Influence of Partisan Motivated Reasoning on Public Opinion.” *Political Behavior* 36(2):235–262.

Buhrmester, M., T. Kwang and S.D. Gosling. 2011. “Amazon’s mechanical turk.” *Perspectives on Psychological Science* 6(1):3–5.

Chandler, Jesse, Pam Mueller and Gabriele Paolacci. 2014. “Nonnaivete among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers.” *Behavioral Research Methods* 46(1):112–130.

Egami, Naoki and Kosuke Imai. 2015. “Causal Interaction in High-Dimension.” Working paper.
URL: *http://imai.princeton.edu/research/files/int.pdf*

Fazio, Russell H. 1990. “A practical guide to the use of response latency in social psychological research.” *Research methods in personality and social psychology* 11:74–97.

Hainmueller, Jens. 2012. “Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies.” *Political Analysis* 20(1):25–46.

Hainmueller, Jens, Daniel J Hopkins and Teppei Yamamoto. 2014. “Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments.” *Political Analysis* 22(1):1–30.

Healy, Andrew and Gabriel S Lenz. 2014. “Substituting the End for the Whole: Why Voters Respond Primarily to the Election-Year Economy.” *American Journal of Political Science* 58(1):31–47.

Holsti, Ole R. and James N. Rosenau. 1988. “The domestic and foreign policy beliefs of American leaders.” *Journal of Conflict Resolution* 32(2):248–294.

Huff, Connor and Dustin H. Tingley. 2015. “Who Are These People?: Evaluating the Demographics Characteristics and Political Preferneces of MTurk Survey Respondents.” *Research and Politics* 2(3):1–12.

Huff, Connor and Joshua D. Kertzer. 2018. “How The Public Defines Terrorism.” *American Journal of Political Science* 62(1):55–71.

Kam, Cindy D and Elizabeth N Simas. 2010. “Risk orientations and policy frames.” *The Journal of Politics* 72(2):381–396.

Kertzer, Joshua D., Kathleen Powers, Brian C. Rathbun and Ravi Iyer. 2014. “Do moral values shape foreign policy preferences?” *Journal of Politics* 76(3):825–840.

Kriner, Douglas L. and Francis X. Shen. 2014. "Reassessing American Casualty Sensitivity: The Mediating Influence of Inequality." *Journal of Conflict Resolution* 58(7):1174–1201.

Krupnikov, Yanna and Adam Seth Levine. 2014. "Cross-Sample Comparisons and External Validity." *Journal of Experimental Political Science* 1(1):59–80.

Kurzban, Robert, John Tooby and Leda Cosmides. 2001. "Can race be erased? Coalitional computation and social categorization." *Proceedings of the National Academy of Sciences* 98(26):15387–15392.

Levy, Jack S., Michael K. McKoy, Paul Poast and Geoffrey PR Wallace. 2015. "Commitments and Consistency in Audience Costs Theory." *American Journal of Political Science* 59(4):988–1001.

Lopez, Anthony C., Rose McDermott and Michael Bang Petersen. 2011. "States in Mind: Evolution, Coalitional Psychology, and International Politics." *International Security* 36(2):48–83.

Masling, Joseph. 1966. "Role-related Behavior of the Subject and Psychological Data." *Nebraska Symposium on Motivation* 14:67–103.

Mason, W. and S. Suri. 2010. "Conducting behavioral research on Amazon's Mechanical Turk." *Behavior Research Methods* pp. 1–23.

Mulligan, Kenneth, J. Tobin Grant, Stephen T. Mockabee and Joseph Quin Monson. 2003. "Response Latency Methodology for Survey Research: Measurement and Modeling Strategies." *Political Analysis* 11(3):289–301.

Orne, Martin T. 1962. "On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications." *American Psychologist* 17(11):776–783.

Paolacci, Gabriele and Jesse Chandler. 2014. "Inside the turk: understanding mechanical turk as a participant pool." *Current Directions in Psychological Science* 23(3):184–188.

Rand, D.G. 2012. "The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments." *Journal of Theoretical Biology* 299(21):172–179.

Rosenthal, Robert and Kermit L. Fode. 1963. "Psychology of the Scientist: V. Three Experiments in Experimenter Bias." *Psychological Reports* 12(2):491–511.

Tomz, Michael R and Jessica Weeks. 2013. "Public opinion and the democratic peace." *American Political Science Review* 107(4):849–865.

Tversky, Amos and Daniel Kahneman. 1981. "The framing of decisions and the psychology of choice." *Science* 211(4481):453–458.

Wallace, Geoffrey PR. 2013. "International law and public attitudes toward torture: An experimental study." *International Organization* 67(01):105–140.

Wittkopf, Eugene R. 1990. *Faces of internationalism: Public opinion and American foreign policy.* Duke University Press.

Yarhi-Milo, Keren, Joshua D Kertzer and Jonathan Renshon. 2018. "Tying Hands, Sinking Costs, and Leader Attributes." *Journal of Conflict Resolution* Forthcoming.