Appendices

Appendix A: Experimental Material

Appendix B: Internal Validity & Sample analysis

Appendix C: Main results, screening phase

Appendix D: Main results, policy selection phase

Appendix E: Main results, contextual preference reversal

Appendix A

In this section, I detail the experimental instrument used for collecting the data.

A.1: Factorial design

The experimental is based on a 2 (time horizon main) x 2 (reciprocal/other outcomes) x 2 (human casualties) design. In addition, I add two control (baseline) conditions consisting of only the Short/Long-term orientation treatment and no 'cost' or 'other outcomes' treatments.

	Short term Orientation		Long term orientation	
	Other version 1	Other version 2	Other version 1	Other version 2
Costs Low				
Costs High				

Figure 1: Factorial Design

A.2: Instructions

All respondents read an instructions page that describes the main objective and expectations.

This experiment explores public opinion regarding government policy in a foreign conflict that is significant to US interests. All of your responses will be anonymous.
You will be presented with a hypothetical scenario about an international conflict involving US interests. Then, a list of possible policy options will be presented. You will be asked, given all the information provided, to choose whether to accept or reject each alternative for further evaluation by senior decision makers. The objective for you is to create a 'short list' of options that you see as more useful to address the situation.
Following that, you will be asked to answer a few follow-up questions about your selection of alternatives and the experiment as a whole.
While this specific scenario is fictional, similar situations have occurred in the past. Your thoughtful responses will help researchers better understand the complex deliberations concerning decisions such as those involved in these kinds of situations.

A.3: Experimental treatments

Respondents are randomly assigned to one of the scenarios below (10 total scenarios).

Baseline conditions scenarios (total of two)
"Bagram Air base - US military headquarters in Afghanistan has been under constant attacks over the last 6 months (resulting in both casualties and damages). The administration has to address this situation since a high ranking US government official (the vice president) is planning an official visit [*short time horizon condition*: at the end of this month / *long time horizon*

condition: early next year (about 8 months from now)]. Military decision makers need to implement a policy that will provide a secure environment for the visit [*short time horizon condition*: as soon as possible / *long time horizon condition*: that is planned about 8 months from now]. Officials emphasize that immediate solution is [*short time horizon condition*: critical and you should be less concerned about future prospects of the situation (i.e. continued insurgency in later time) / *long time horizon condition*: not that critical and thus you should be less concerned with current insurgency attacks. You must be more concerned about future prospects of the situation and the planned visit next year.]

Below are 7 potential policy options to address the situation. For each option, experts estimate its chances of success in the [*short time horizon condition*: short term / *long time horizon condition*: long term]. We ask that you choose which of the options should be accepted for further consideration by senior decision makers and which to reject. Only accept the options that you consider to be worthy for further deliberation."

After the scenario, respondents face the first decision task - screening of policy options. The 7 options are described using a table format.

	Chances of success in (Short / Long) term	Choose one option
Policy 1	85%	Accept / Reject
Policy 2	80%	Accept / Reject
Policy 3	75%	Accept / Reject
Policy 4	70%	Accept / Reject
Policy 5	65%	Accept / Reject
Policy 6	60%	Accept / Reject
Policy 7	55%	Accept / Reject

Full conditions scenarios (total of eight)
"Bagram Air base - US military headquarters in Afghanistan has been under constant attacks over the last 6 months (resulting in both casualties and damages). The administration has to address this situation since a high ranking US government official (the vice president) is planning an official visit [*short time horizon condition*: at the end of this month / *long time horizon condition*: early next year (about 8 months from now)]. Military decision makers need to implement a policy that will provide a secure environment for the visit [*short time horizon condition*: as soon as possible / *long time horizon condition*: that is planned about 8 months from now]. Officials emphasize that immediate solution is [*short time horizon condition*: critical and you should be less concerned about future prospects of the situation (i.e. continued insurgency in later time) / *long time horizon condition*: not that critical and thus you should be less concerned with current insurgency attacks. You must be more concerned about future prospects of the situation and the planned visit next year.]

Below are 7 potential policy options to address the situation. For each option, experts estimated the chances of success in the [*short time horizon condition*: short term, as well as the projected long term outcomes / *long time horizon condition*: long term, as well as the projected short term outcomes]. Lastly, military officials provided an estimate to the expected number of casualties of these policy options (relatively [*Casualties condition high*: high / *Casualties condition low*: low)]. Based on the scenario and the information below we ask that you choose

which of the options should be accepted for further consideration by senior decision makers and which to reject. Only accept the options that you consider to be worthy for further deliberation."

After the scenario, respondents face the first decision task - screening of policy options. The 7 options are described using a table format.

	Chances of success in (Short / Long) term	(Long / Short) term outcomes	Expected Casualties	Choose one option
Policy 1	85%	Good / Bad	70-90 / 10-15	Accept / Reject
Policy 2	80%	Bad / Good	70-90 / 10-15	Accept / Reject
Policy 3	75%	Bad / Good	70-90 / 10-15	Accept / Reject
Policy 4	70%	Good / Bad	70-90 / 10-15	Accept / Reject
Policy 5	65%	Bad / Good	70-90 / 10-15	Accept / Reject
Policy 6	60%	Good / Bad	70-90 / 10-15	Accept / Reject
Policy 7	55%	Bad / Good	70-90 / 10-15	Accept / Reject

Selection Phase
Following this stage, in the next page, participants are asked to consider the alternatives they accepted for further evaluation and select their preferred option to address the situation described in the scenario.

1. Below are the list of policies you chose to evaluate further. Which one do you favor the most to address the conflict situation?

Option 1
Option 2
.
Option n

2. Below are the list of policies you chose to evaluate further. For each, we ask you to rate your level of support (1- low support to 7 - high support).

Policy option 1...n:

Low support 1 2 3 4 5 6 7 High support

A.4: Embedded Experiment (new Choice-sets)

In the next page, we ask participants to select policy options from different combinations that we provide. We add the scenario in order to ensure participants are still aware of the setting of decision. The scenario can be either ST or LT, depending on the initial treatment participants are assigned into. The wording of ST/LT in the questions below also depends on the initial treatment. In the example below, the ST condition is used.

"Below are different sets of policy options to address the situation in Afghanistan (attached below in case you need a refresher). In each question, we ask you to select your preferred option."

1. Which of the two options below do you prefer? (policy options are described using three elements: the chances of success in the short term, the projected outcomes in the long term, and the expected number of casualties from each policy).

	Success in (Short / Long) term	(Long / Short) outcomes	Expected Casualties
(1)	80%	Bad	10-15
(2)	70%	Good	10-15

2. Now, a new option is introduced. Which of the policy options do you prefer?

	Success in (Short / Long) term	(Long / Short) outcomes	Expected Casualties
(1)	80%	Bad	10-15
(2)	75%	Good	70-90
(3)	70%	Good	10-15

3. Consider a new set of options to address the situation. Which do you prefer?

	Success in (Short / Long) term	(Long / Short) outcomes	Expected Casualties
(1)	80%	Good	70-90
(2)	70%	Bad	70-90

4. Now, a new option is introduced to this set. Which of the policy options do you prefer?

	Success in (Short / Long) term	(Long / Short) outcomes	Expected Casualties
(1)	80%	Good	70-90
(2)	75%	Bad	10-15
(3)	70%	Bad	70-90

A.5: General issues questionnaire

1. How Important is the situation described in the text to US national security?

Not Important 1 2 3 4 5 6 7 Very Important

2. What is your assessment of the costs associated with your selected policy to improve the security of US headquarters?

Low Costs 1 2 3 4 5 6 7 High Costs

3. Based on the scenario you read, when is it more critical to address the insurgents' attacks on Bagram air base?

- In the short term (this month)

- In the long run (next year).

4. What is the target of the attack described in the scenario?
   - An American military base.
   - An American embassy building.
5. In which location did the incident took place?
   - Iran.
   - Kenya.
   - Afghanistan.

A.6: Demographic characteristics questionnaire

1. State your gender
   - Male.
   - Female.
2. State your age:
3. Rank your political orientation on a 1-7 scale (1 being Democrat and 7 is Republican):

   Democrat 1 2 3 4 5 6 7 Republican

4. How informed do you consider yourself about US foreign and security policy?
   - Not well informed.
   - Somewhat informed.
   - Well informed.
   - Very informed.
5. What is your highest level of education?
   - Less than high school.
   - High school.
   - Some college.
   - Bachelor's degree.
   - master's degree or higher.

Appendix B

In this section, I discuss issues related to the experiment itself - tests for internal validity, followed by analysis of the sample including descriptive characteristics, power analysis for sample size and the external validity debate.

B.1: Internal Validity

One of the main advantages of experiments is their ability to measure direct causal effects in complex situations (Hyde 2015; Morton and Williams 2008). However, it is essential to ensure the internal validity of the design. Therefore, I provide evidence regarding the internal validity of the experimental design. The findings show that for the most part, the design properly captures the different concepts and respondents are aware of the information in both the scenario and experimental treatments.

First, a central concern when using survey experiments is that respondents pay attention to the details described in the text (Berinsky, Margolis and Sances 2014; Clifford and Jerit 2015) and complete their tasks seriously (Anson 2018; Bayram 2018). To address these concerns, I utilize two items as factual manipulation checks (FMCs) (Kane and Barabas 2018). For both items, a high proportion of respondents recognize details about the location of attack (91%, a military base), and the country where the conflict takes place (90%, Afghanistan).

The other concern for internal validity is to ensure that the experimental treatments operate as expected. I utilize several items for these checks. The first item asks whether respondents are aware of the main temporal treatment ("When is it more critical to address the attack?"). About 85% of respondents in the short-term condition answer this item correctly, for the long-term condition, the number is almost 80%. Then, the results from probit and ANOVA models show that the time horizon treatment is significant and in the expected direction. The second item asks respondents to rate the level of human costs in their preferred policy. The results of an ANOVA model show that the high costs condition led respondents to evaluate the expected costs as higher than the low cost condition.

The final item is intended to assess the degree of importance that respondents assign to this scenario. Measured on a 1-7 scale, the mean response is 5.73 which is relatively high. In

addition, I run an OLS regression model with the importance item as a dependent variable and the experimental treatments as independent variables. The results show that the time horizon treatment is negative and significant suggesting that respondents perceive a short term objective as more important ($m = 5.82$ for short-term vs. $m = 5.65$ for long-term, $p < .05$).

Taken together, the results of these tests indicate that the design offers a relatively strong internal validity. Most of the findings show that participants are aware of the differences between conditions and understand the context of the conflict situation.

B.2: Power Analysis

When using experiments as a data collection method, one concern is with regard to statistical power. *Power* is the probability of correctly rejecting the null hypothesis when the alternative hypothesis is correct (Cohen 1977). When computing power, one element that must be part of the computation is participants' sample size (Judd, Westfall and Kenny 2017). An accepted notion is that increasing the size of the sample improves the power of the test, i.e. identifying the effects we want to estimate. Below, I present findings from a power analysis of the experiment.

My experimental design includes eight experimental and two baseline (control) conditions, for a total of 10 conditions. The sample I recruited for this research was 1100 participants, after removing respondents that failed to complete the necessary items, the final sample consists of 1036 participants.

I employ a one-way ANOVA power analysis which is best fit for comparing multiple groups. For my purposes, the test calculates the necessary sample size in each **group/condition** to obtain power in the conventional levels of 0.8-0.9, as well as accepted effect sizes of 0.1, 0.25, and 0.4 (Cohen 1977).

The table below presents the results. For the more common effect sizes (0.25 and 0.4), and power of 0.8, the test recommends a sample size of 250-260 ($d = 0.25$) and 105-110 ($d = 0.4$) participants. For a higher power (0.9), recommended sample sizes are bigger, but still well short of my recruited sample.

Table 1: Power Analysis: One-way ANOVA

Power	Effect Size (Cohen's d)	n
0.8	0.1	1573.45
	0.25	258.96
	0.4	106.55
0.9	0.1	1991.38
	0.25	325.82
	0.4	132.63

In addition, figure 2 presents the expected effect sizes based on different sample sizes for achieving test power of 0.8 and 0.9 (based on One-way ANOVA model). Both plots show that to achieve the conventional effect sizes of 0.2-0.4, the number of respondents **per cell** does not exceed 50. For my sample, the number of participants per cell is 100-110.

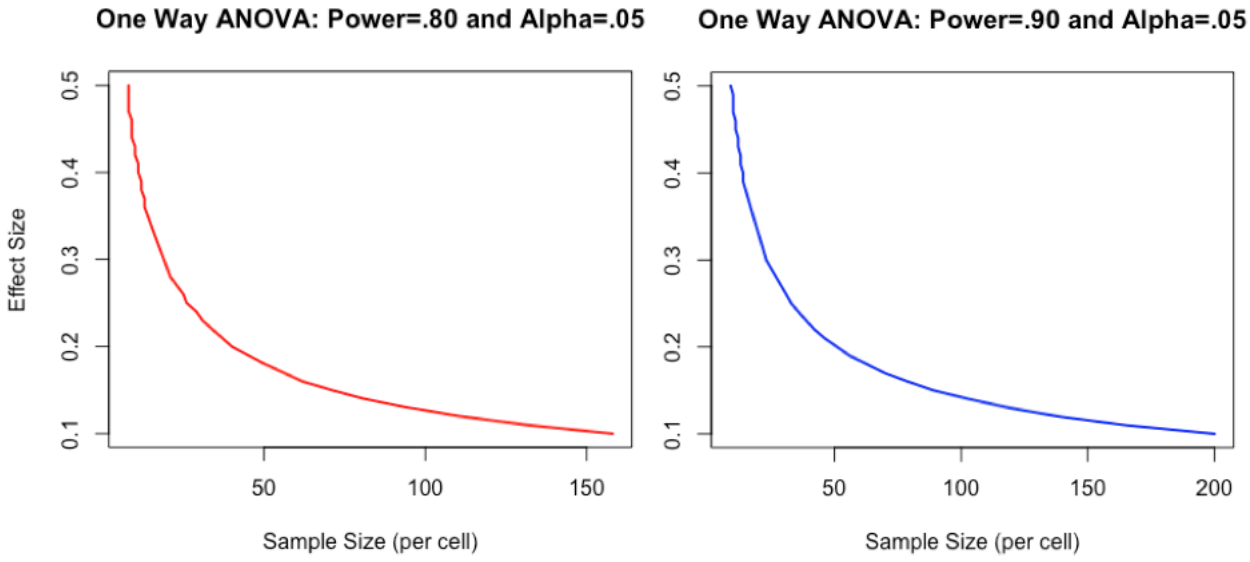

Figure 2: Sample and effect size: Power analysis

B.3: Sample

I recruit a sample of 1100 participants. The final sample size is 1036 after removing those participants that had multiple non-responses. Similar to other MTurk samples, most respondents have an undergraduate or graduate degrees. At the same time, the proportion of individuals with these degrees in the general population is lower (around 50% according to 2019 census

Table 2: Summary Statistics - MTurk Sample

		Statistics:			
Characteristic	Levels	n	Proportion	Mean	Median
Gender	Women	518	50%		
	Men	518	50%		
Age		1036		38	35
Partisanship	Democratic	215	21%		
	2	143	14%		
	3	109	11%		
	4	201	19%		
	5	141	14%		
	6	103	9.9%		
	Republican	124	12%		
	Total	1036		3.69	4
Education	Less than High School	3	0.3%		
	High School	95	9.2%		
	Some College	296	29%		
	Bachelor's Degree	484	47%		
	Master's Degree of Higher	156	15%		
	Total	1034		3.67	4
Foreign Pol. Knowledge	Not Well Informed	118	11%		
	Somewhat Informed	483	47%		
	Well Informed	309	30%		
	Very Informed	126	12%		
	Total	1036		2.42	2

Total Sample Size: 1036

data),[1] suggesting the my sample is highly-educated.

Nevertheless, for the relevant quantities of interest in the study (for instance, the screening phase), I find variation amongst those highly educated individuals, and in the expected direction. For example, the average choice-set size for respondents with high school education is lower than those with an undergraduate or graduate degrees. Similar results are evident for those with some college education compared to those with higher degrees. Therefore, I recognize that the sample is skewed towards highly educated respondents (as most MTurk samples). Yet, I am still able to find variation among participants, and in the direction I expected based on my theory, further amplifying my argument that knowledge matters.

B.3.1 External Validity

One of the central challenges of experimental work is questions about its external validity - does a sample of approximately 1000 individuals properly represents the decision-making process in a foreign policy scenario? This issue has been discussed in many previous studies (Hyde 2015; McDermott 2002, 2011).

There are two levels to this issue. First, in general, past work has shown that MTurk samples can be used to properly replicate work that used national representative sample (Coppock, Leeper and Mullinix 2018; Kam and Simas 2010). Also, over the last decade, MTurk samples became widely used in political science research and are published in a wide variety of journals (Healy and Lenz 2014; Huff and Kertzer 2018; Kertzer, Renshon and Yarhi-Milo 2019; Tomz and Weeks 2013).

Second, the main challenge when using a public sample relates to the decision task I explore - choices in foreign policy situations. Previous work argue that elites possess different decision-making skills and thus using the public for such decision tasks may introduce biases into the outcomes (Renshon 2015; Hafner-Burton, Hughes and Victor 2013). While I recognize this issue, I argue that using the public for such empirical tests is more than a convenience choice - politicians' decision calculus is at least partially influenced by public opinion (Baum and Potter 2015). In other words, foreign policy choices can be a 'bottom-up' process (Kertzer and Zeitzoff

[1] See, 2019 Census

2017) based primarily on accountability and politicians' desire to remain in office (De Mesquita et al. 2003).

An extensive amount of research provide evidence for these arguments, both earlier work on public and elite opinion in the context of using force (Wittkopf and Maggiotto 1983; Page and Shapiro 1983), as well as recent studies that compare public and elite samples in a variety of foreign policy (especially conflict) situations. In all these studies, a substantial amount of congruence exists between the views of elites and the general public (Friedman, Lerner and Zeckhauser 2017; Renshon, Yarhi-Milo and Kertzer 2016; Sheffer et al. 2018; Sheffer and Loewen 2018; Tomz, Weeks and Yarhi-Milo 2019).[2]

Lastly, from a methodological standpoint, experiments are designed mainly to test hypotheses that are deduced from a given theory. Further echoing sentiments expressed by Mook (1983, p. 386) that "Ultimately what makes research findings of interest is that it help us understand everyday life. That understanding, however, comes from theory or the analysis of mechanism; it is not a matter of 'generalizing' the finding themselves". He suggests, then, that the results simply support the logic of the theory, given that the experiment is an appropriate representation of the theory. Further supporting this logic, in a recent review, Hafner-Burton et al. (2017) argue that for theories which are expected to apply to *any decision-maker* (emphasis mine), using convenient samples for empirical analysis is appropriate as long as researchers assure that the investigated treatment is randomly assigned, thus allowing an appropriate identification of the causal effects.[3]

[2] A recent meta-analysis of studies that measure both elites and public opinion finds that both populations respond to treatments in very similar ways, see Kertzer (2020)

[3] In addition to this elaborate discussion, I offer, in the introduction of this study, a brief historical example (Bush, Clinton and the crisis in Bosnia 1992-1993) of my causal mechanism at work.

Appendix C

C.1: Choice-set size (OLS Regression models)

In the first empirical analysis, I assess how the experimental conditions affect the size of the choice-set. In the main text, I focus on estimating differences by using ANOVA models. Below, I complement the analysis with two OLS regression models that estimate the effects of the experimental treatments and individual covariates on the size of the choice-set.

Model 1 estimates the experimental conditions only. In model 2, I add individual covariates. Overall, the results are similar to the ANOVA models. For the main time horizon treatment in model 2, the coefficient is on the verge of statistical significance ($p = 0.101$).

Table 1 also includes Model 3 which estimates, using univariate regression, how time horizons affect the choice-set size in the reduced sample (only baseline conditions). The results of the model are described in the main text and are the basis for the density plots in figure 4.

Table 3: Time Horizons and Choice Set Size: OLS Regression Models

	Choice Set Size		
	Model 1	Model 2	Model 3
Time Horizon	0.154*	0.147	0.342*
	(0.092)	(0.089)	(0.204)
Reciprocal Outcomes	0.115	0.145	
	(0.103)	(0.036)	
Casualties_Low	−1.044***	−1.022***	
	(0.198)	(0.192)	
Casualties_High	−1.038***	−1.042***	
	(0.199)	(0.194)	
Gender		0.034	0.137
		(0.092)	(0.205)
Age		−0.012***	−0.008
		(0.003)	(0.008)
Partisanship		0.105***	0.085*
		(0.022)	(0.049)
Education		0.109**	0.014
		(0.054)	(0.115)
FP Knowledge		0.219***	0.259**
		(0.055)	(0.118)
Constant	3.899***	3.002***	3.186***
	(0.111)	(0.296)	(0.628)
Observations	1036	1034	202
R^2	0.053	0.104	0.035

Notes: *p<0.1; **p<0.05; ***p<0.01
Standard errors in parenthesis

C.2: Choice-set composition (Probit conditional models)

The table below details two probit interaction models used to estimate the effect of time horizons on the likelihood of accepting a policy into the choice set. The conditional effects are depicted in the main text in figure 5.

Table 4: Time Horizons and Choice Set Composition: Probit Regression Models

	Accepting Policy Option	
	Policy 1	Policy 2
Time Horizon	0.123	0.227*
	(0.222)	(0.125)
Reciprocal Outcomes	−1.977***	1.481***
	(0.177)	(0.148)
Casualties	−0.096	−0.186*
	(0.110)	(0.192)
Time Horizon x		
Reciprocal Outcomes	0.429*	−0.077
	(0.256)	(0.214)
Gender	0.115	−0.050
	(0.115)	(0.103)
Age	0.007	−0.002
	(0.004)	(0.004)
Partisanship	0.086***	−0.033
	(0.027)	(0.025)
Education	0.044	0.122**
	(0.068)	(0.061)
FP Knowledge	−0.055	−0.021
	(0.071)	(0.063)
Constant	1.059***	−0.306
	(0.373)	(0.312)
Observations	821	821
Pseudo R^2	0.296	0.208

Notes: *p<0.1; **p<0.05; ***p<0.01
Standard errors in parenthesis
Models based on reduced sample - no baseline conditions

Appendix D

D.1: Policy selection (Multinomial regression model)

The table below replicates the analysis of the multinomial regression model of policy selection (table 1). Including the choice-set size as a control variable raises concerns of post-estimation bias (Montgomery, Nyhan and Torres 2018). Recent work devote much attention to this issue and offer a variety of methods to account for this problem (Acharya, Blackwell and Sen 2016; Klar, Leeper and Robison 2020; Mummolo 2016). However, most of these tools are irrelevant for my analysis, mostly since the theory I propose suggests that the choice-set size is affected by the experimental treatments.

Despite the relative irrelevance of this bias to my theory, I address it below. I replicate the model in table 1 but omit the choice-set size variable from the analysis. The results are displayed in table 5 below. Similar to the main text, selecting policy 1 is the reference category. For all three alternatives policies, the results are similar to the main text, suggesting that in this context, we should be less concerned about this bias.

Table 5: Policy Selection - Multinomial Logit Model

	Policy Selected		
	Policy 2	Policy 3	Policy 4
Time Horizon	−0.471**	−0.806**	−0.924*
	(0.200)	(0.402)	(0.560)
Reciprocal Outcomes	3.652***	2.493***	1.030**
	(0.214)	(0.381)	(0.524)
Casualties	−0.815***	−0.578	−0.209
	(0.196)	(0.367)	(0.471)
Age	−0.024***	−0.093***	−0.016
	(0.008)	(0.025)	(0.022)
Gender	0.225	−0.392	−0.349
	(0.205)	(0.426)	(0.547)
Partisanship	0.011	0.219**	0.137
	(0.050)	(0.100)	(0.134)
Education	0.233*	0.300	−0.024
	(0.123)	(0.267)	(0.319)
FP knowledge	0.239*	0.634**	0.190
	(0.128)	(0.250)	(0.329)
Constant	−5.1051***	−5.148***	−4.192**
	(0.736)	(1.431)	(1.664)

Notes: N = 1,020; Pseudo R^2 = 0.325; Base category is select Policy 1

*p<0.1; **p<0.05; ***p<0.01; Standard errors in parenthesis

Appendix E

E.1: Contextual preference reversal - cross tabs

To estimate the emergence of contextual preference reversal (Howes et al. 2016; Tsetsos, Usher and Chater 2010) in the embedded experiment, I conduct an initial analysis that uses cross-tabulations of the preferred policy options between both combinations of choice-sets. I use a χ^2 test to check for differences.

This preliminary analysis finds that the introduction of a new option into the original set (as described in either set B or D) changes the distribution of preferences for all options. In the table below, I present cross tabulations (frequencies and percentages) of respondents' preferred policies with either two- or three-options choice sets. In both panels, the number of respondents who experience contextual preference reversal is quite substantial (most obvious in set D). These differences are significant based on a χ^2 test.

Table 6: Preference Reversal

	Set B			
Set A	Pol. 1	Pol. 2	Pol. 3	**Total**
Pol. 1	73.4% (332)	19.03% (86)	7.5% (34)	452
Pol. 2	5.65% (33)	78.9% (461)	15.4% (90)	584
	Set D			
Set C	Pol. 1	Pol. 2	Pol. 3	**Total**
Pol. 1	47.5% (459)	3.01% (29)	49.3% (477)	965
Pol. 2	29.5% (21)	56.3% (40)	14.08% (10)	71

Notes: Panel 1 (Set A): $\chi^2 = 518.914$, $p = 0.00$
Panel 2 (Set C): $\chi^2 = 304.634$, $p = 0.00$

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining causal findings without bias: Detecting and assessing direct effects." *American Political Science Review* 110(3):512–529.

Anson, Ian G. 2018. "Taking the time? Explaining effortful participation among low-cost online survey participants." *Research & Politics* 5(3):1–8.

Baum, Matthew A and Philip BK Potter. 2015. *War and democratic constraint: How the public influences foreign policy*. Princeton University Press.

Bayram, A Burcu. 2018. "Serious subjects: A test of the seriousness technique to increase participant motivation in political science experiments." *Research & Politics* 5(2):1–6.

Berinsky, Adam J, Michele F Margolis and Michael W Sances. 2014. "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys." *American Journal of Political Science* 58(3):739–753.

Clifford, Scott and Jennifer Jerit. 2015. "Do attempts to improve respondent attention increase social desirability bias?" *Public Opinion Quarterly* 79(3):790–802.

Cohen, Jacob. 1977. *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Academia Press, Inc.

Coppock, Alexander, Thomas J. Leeper and Kevin J. Mullinix. 2018. "Generalizability of heterogeneous treatment effect estimates across samples." *Proceedings of the National Academy of Sciences* 115(49):12441–12446.

De Mesquita, Bruce Bueno, Alastair Smith, R. M. Siverson and J. D. Morrow. 2003. *The logic of political survival*. MIT press.

Friedman, Jeffrey A, Jennifer S Lerner and Richard Zeckhauser. 2017. "Behavioral consequences of probabilistic precision: experimental evidence from national security professionals." *International organization* 71(4):803–826.

Hafner-Burton, Emilie M, D Alex Hughes and David G Victor. 2013. "The cognitive revolution and the political psychology of elite decision making." *Perspectives on Politics* 11(2):368–386.

Hafner-Burton, Emilie M, Stephan Haggard, David A Lake and David G Victor. 2017. "The Behavioral Revolution and International Relations." *International Organization* 71(S1):S1–S31.

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.

Howes, Andrew, Paul A Warren, George Farmer, Wael El-Deredy and Richard L Lewis. 2016. "Why contextual preference reversals maximize expected value." *Psychological review* 123(4):368.

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

Hyde, Susan D. 2015. "Experiments in international relations: lab, survey, and field." *Annual Review of Political Science* 18:403–424.

Judd, Charles M, Jacob Westfall and David A Kenny. 2017. "Experiments with more than one random factor: Designs, analytic models, and statistical power." *Annual Review of Psychology* 68:601–625.

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

Kane, John V. and Jason Barabas. 2018. "No Harm in Checking: Using Factual Manipulation Checks to Assess Attentiveness in Experiments." *American Journal of Political Science* 63(1):234–249.

Kertzer, Joshua. 2020. "Re-Assessing Elite-Public Gaps in Political Behavior." Forthcoming in the American Journal of Political Science.

Kertzer, Joshua D, Jonathan Renshon and Keren Yarhi-Milo. 2019. "How Do Observers Assess Resolve?" *British Journal of Political Science* . First View.

Kertzer, Joshua D and Thomas Zeitzoff. 2017. "A Bottom-Up Theory of Public Opinion about Foreign Policy." *American Journal of Political Science* 61(3):543–558.

Klar, Samara, Thomas Leeper and Joshua Robison. 2020. "Studying Identities with Experiments: Weighing the Risk of Posttreatment Bias Against Priming Effects." *Journal of Experimental Political Science* 7(1):56–60.

McDermott, Rose. 2002. "Experimental methods in political science." *Annual Review of Political Science* 5(1):31–61.

McDermott, Rose. 2011. "New directions for experimental work in international relations." *International Studies Quarterly* 55(2):503–520.

Montgomery, Jacob M, Brendan Nyhan and Michelle Torres. 2018. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." *American Journal of Political Science* 62(3):760–775.

Mook, Douglas G. 1983. "In defense of external invalidity." *American psychologist* 38(4):379.

Morton, Rebecca B and Kenneth C Williams. 2008. Experimentation in political science. In *The Oxford Handbook of Political Methodology*, ed. Henry E. Brady Janet M. Box-Steffensmeier and David Collier. New York: Oxford University Press pp. 339–356.

Mummolo, Jonathan. 2016. "News from the other side: How topic relevance limits the prevalence of partisan selective exposure." *The Journal of Politics* 78(3):763–773.

Page, Benjamin I and Robert Y Shapiro. 1983. "Effects of public opinion on policy." *American political science review* 77(01):175–190.

Renshon, Jonathan. 2015. "Losing face and sinking costs: Experimental evidence on the judgment of political and military leaders." *International Organization* 69(3):659–695.

Renshon, Jonathan, Keren Yarhi-Milo and Joshua Kertzer. 2016. "Democratic leaders, crises and war: Paired experiments on the Israeli Knesset and public." Working Paper.

Sheffer, Lior and Peter John Loewen. 2018. "Accountability, Framing Effects, and Risk-Seeking by Elected Representatives: An Experimental Study with American Local Politicians." *Political Research Quarterly* 72(1):49–62.

Sheffer, Lior, Peter John Loewen, Stuart Soroka, Stefaan Walgrave and Tamir Sheafer. 2018. "Nonrepresentative Representatives: An Experimental Study of the Decision Making of Elected Politicians." *American Political Science Review* 112(2):302–321.

Tomz, Michael, Jessica Weeks and Keren Yarhi-Milo. 2019. "Public Opinion and Decisions about Military Force in Democracies." *International Organization* pp. 1–25. First View.

Tomz, Michael R and Jessica LP Weeks. 2013. "Public opinion and the democratic peace." *American Political Science Review* 107(04):849–865.

Tsetsos, Konstantinos, Marius Usher and Nick Chater. 2010. "Preference reversal in multiattribute choice." *Psychological review* 117(4):1275–1293.

Wittkopf, Eugene R and Michael A Maggiotto. 1983. "Elites and masses: A comparative analysis of attitudes toward America's world role." *The Journal of Politics* 45(2):303–334.