Re-Assessing Elite-Public Gaps in Political Behavior
Supplementary Appendix

August 1, 2020

Contents

1 Experimental data 2
  1.1 Meta-analysis study inclusion criteria 2
  1.2 Robustness checks and additional results 4
    Figure 1: Joint distribution of cell means 5
    Figure 2: Average treatment effects 6
    Figure 3: Distribution of sample sizes by elite status 7
    Figure 4: Sensitivity analyses: restricting the minimum acceptable elite sample size 7
    Figure 5: Compression tests 9
    Figure 6: Difference-in-differences between elites and masses across paired treatments 10
    Figure 7: Comparing observed effect sizes with a null distribution 11
  1.3 Tests for publication bias 12
    Figure 8: Funnel plot 12
    1.3.1 Additional metaregression tests 12
    Table 2: Metaregression results: elite-public interaction 15
  1.4 List of studies included in the meta-analysis 15

2 Observational data 18
  2.1 Demographics 18
    Figure 9: Age distributions in elite and mass samples 19
    Figure 10: Gender distributions in elite and mass samples 20
    Figure 11: Education distributions in elite and mass samples 21
  2.2 Alternative empirical strategies 22
    2.2.1 Subsetting method 22
    Figure 12: Elite-public gaps, before and after adjusting for compositional differences (simulation method) 23
    Table 3: Decomposing elite-public gaps using subsetting method 24
    Figure 13: Partisan divides between elite and mass samples over time 24
    2.2.2 Blinder-Oaxaca decomposition analysis 25
    Table 4: Partisan divides between elite and mass samples over time 26
    Figure 14: Two-way decomposition analysis 27
  2.3 Recalculating elite-public gaps using a more restrictive set of foreign policy elites 27
    Figure 15: Government elite and mass foreign policy preferences, after adjusting for compositional differences 29
  2.4 Are the results the artifact of elite cues? 30
    Table 5: Decomposing elite-public gaps using news engagement rather than education 31
1 Experimental data

1.1 Meta-analysis study inclusion criteria

As noted in the main text, a study meets the inclusion criteria for the meta-analysis of paired experiments on political elites and mass samples if it contains:

1. an experiment where the treatments are randomly assigned by an experimenter, and

2. the same experiment is fielded both on a sample of political elites (current or former politicians, civil servants, military officers, etc.) and a mass public or convenience sample.

Thus, studies outside the bounds of the inclusion criteria include:

- Studies fielded just on elites (e.g. Carnevale, Inbar and Lerner, 2011; Hafner-Burton, LeVeck and Victor, 2017; Arceneaux, Dunaway and Soroka, 2018; Hardt, 2018; Walgrave et al., 2018; Banuri, Dercon and Gauri, 2019)

- Observational studies comparing elites and masses. This is particularly common in experimental economics, where studies are often considered experimental because they involve a stylized game, rather than because they are formally estimating the effects of treatments randomly assigned by an experimenter (e.g. Alatas et al., 2009; LeVeck et al., 2014). It is also common in observational studies where the purpose is to compare distributions of personality traits or characteristics between the two samples (e.g. Rathbun, 2007; Best, 2011; Sherman et al., 2012; Dal Bó et al., 2017; Heß et al., 2018; Dynes, Hassell and Miles, 2019; Clifford, Kirkland and Simas, 2019; Dynes et al., 2019), as well as wargames without random assignment (Pauly, 2018).

- Paired experiments that feature conceptual rather than direct replications (e.g. Bayram, 2017; Kertzer, Renshon and Yarhi-Milo, 2019)

- Studies where the elite sample does not consist of political elites, as is sometimes the case in studies in economics studying the role of expertise, usually in non-political contexts (e.g. samples of stock traders, managers, chess players — Cooper et al., 1999; Fehr and List, 2004; Abbink and Rockenbach, 2006; Palacios-Huerta and Volij, 2009)

- Paired experiments where the elite and mass samples are given non-overlapping sets of treatments and/or outcome variables (e.g. Butler and Dynes, 2016; Friedman and Zeckhauser, Forthcoming; Grossman and Michelitch, 2018; Tomz, Weeks and Yarhi-Milo, 2020) — as in some studies interested

---

1 On the bounds of experiments in political science, see McDermott (2002)
in how elite behavior changes when given information about the public, and how the public’s behavior changes when given information about political elites.

For studies that contained partially overlapping sets of treatments, only those treatments that were fielded on both elite and mass public samples were included.

Importantly, the purpose of the meta-analysis is positive rather than normative: it focuses on the question of how differently elites and masses behave in paired experiments — rather than assessing which group makes “better” or less biased decisions, a judgment most of the studies in the meta-analysis were not equipped to address.

To be as comprehensive as possible, studies were tracked down using Google Scholar and library databases, following up with authors to inquire about work that might have been overlooked, and sending out calls on email listservs, up until March 2019. In order to calculate standardized and comparable quantities of interest across the studies, the analysis required access to the authors’ data; while sharing replication data for published articles is now common in political science (King, 1995), sharing replication data for unpublished work in progress is not. After some efforts to write code for authors to calculate the necessary quantities of interest, which had only mixed success, I instead adopted a protocol in which all datasets shared would remain confidential, and authors of works in progress had the option of sharing their data anonymously if they were concerned their study might change. Fortunately, almost all of the authors of unpublished work were willing to share their data.

It is worth noting the implications of the above inclusion criteria. In particular, there are a number of studies in economics interested in whether experts play the same non-equilibrium strategies as novices in games where formal theories make sharp predictions about how players should behave (e.g. Camerer, Ho and Chong, 2003; List and Mason, 2011). However, although these studies are experimental in the tradition of experimental economics, the data they produce are observational because they do not analyze treatments randomly assigned by the experimenter, thereby rendering them outside the scope of the meta-analysis.

Indeed, because theoretical assumptions vary based on the formal model being tested, it would be difficult to construct a meta-analysis aggregating across all of these studies. Nonetheless, one implication of this restriction is that the meta-analysis is unable to speak to the question of how much more “rational” elites are compared to masses – although individual studies in the meta-analysis, like the prospect theory experiments of Linde and Vis (2017) and Sheffer et al. (2018), or the motivated reasoning experiments of Baekgaard et al. (2019) suggest both samples tend to behave similarly in the particular domains studied.

---

2 Authors were also informed that the request wasn’t an audit experiment designed to see who replies to requests to share data.

3 In contrast, studies testing propositions from formal models that include random assignment, like Butler and Kousser (2015), are included.

4 I am grateful to an anonymous reviewer for this point.
1.2 Robustness checks and additional results

Although the analysis in the main text focuses on our main quantity of interest — the difference-in-difference, testing whether the differences between treatment and control differs between elites and masses — I include two additional quantities of interest below. First, as an initial cut at the results, our first quantities of interest are the cell means \( E[Y_{i00}], E[Y_{i10}], E[Y_{i01}], E[Y_{i11}] \), which tell us the average responses of elites and masses in each experimental condition. Figure 1 presents the joint distribution of cell means between elites and masses in each experimental condition of each paired treatment. The dashed diagonal line depicts what the results would look like if the cell means within each experimental condition were identical between respondents from mass public samples, versus respondents from elite samples. The plot shows that in general, the correlation in responses between elite and mass respondents within a given experimental condition is fairly high \( r = 0.82 \), but that we also see a number of conditions where the responses differ more substantially. In the control conditions, the median difference (expressed as an absolute value) between elites and masses in a given control condition is 5 percentage points; in the treatment conditions, the median difference is 8 percentage points. However, we also see some more substantial differences between elites and masses in each condition, some by as large as up to 30-45 percentage points.

Second, Figure 2 presents the average treatment effects \( E[Y_{i10} - Y_{i00}], \) and \( E[Y_{i11} - Y_{i01}] \), which despite overplotting, point to similar conclusions as those gleaned from the cell means: most of the ATEs are fairly similar between the elite and mass samples, and are distributed along the main diagonal. A regression line from a Deming regression model also hews quite close to the main diagonal: a 10 percentage point increase in the average treatment effect reported in the mass sample (along the x axis) is associated with an 12 percentage point increase in the average treatment effect reported in the elite sample (along the y axis). At the same time, there are some ATEs that differ substantially between the two groups.

Figure 3 presents the distribution of sample size across the paired experiments. It shows that the elite samples in the data are generally smaller than the mass samples, as one might expect — both by necessity given the relative difficulty of convincing elites to participate in academic studies, and also because the population of elites is definitionally smaller than that of the mass public as a whole; for studies comparing legislators to nationally representative public samples, for example, even an elite experiment that captured the entire population of legislators would still have a smaller sample size than a nationally representative public sample typically would. The median sample size for the calculation of each treatment effect is \( N = 236 \) in the elite samples (implying 118 respondents in each treatment condition), and \( N = 976 \) (implying 488 respondents in each treatment condition) in the mass samples; a quarter of all the treatment effects in the elite samples are estimated on sample sizes under \( N = 75 \).
Figure 1 shows that the cell means for elites and mass responses within each paired treatment condition are fairly similar to one another (r = 0.81), suggesting relatively similar preferences or beliefs between elites and masses. However, there are a number of experimental conditions where we see substantial deviations between elite and mass beliefs and preferences. The blue line and grey bands depict a linear regression line with 95% confidence intervals, while the dashed grey line depicts what a perfect correlation would look like between elite and mass beliefs and preferences within each experimental condition.
Figure 2 shows that the average treatment effects for elites and masses; the ATEs tend to be clustered along the dashed diagonal line (as does the regression line from a Deming regression model, in red, suggesting that a 10 percentage point increase in an ATE in the mass sample is associated with an 12 percentage point increase in an ATE in the elite sample). At the same time, however, there are some paired treatment effects that differ substantially between the two groups. The plot also shows that the 95% confidence intervals are often quite wide, raising questions about statistical power the manuscript explores below.
Figure 3: Distribution of sample sizes by elite status

Figure 3 shows that most experiments on political elites are fielded on relatively small samples; the median number of respondents used for the calculation of each of the average treatment effects among the elite samples in the data is $N = 236$; in the mass samples, the median number of respondents used for the calculation of each of the average treatment is $N = 976$.

Figure 4: Sensitivity analysis: restricting the minimum acceptable elite sample size

Figure 4 varies the minimum acceptable elite sample size, recording the proportion of difference-in-difference estimates that are statistically significant as we raise the elite sample size threshold. The results are relatively stable as we raise the inclusion threshold from the 10th to the 75th percentile; it is only when we raise the threshold to above the 85th percentile that the proportion of significant differences noticeably increases.
The relatively small elite sample sizes lead to a question about interpreting the main results: to what extent are the relatively small elite-public gaps in the data an artifact of underpowered studies? Rather than dichotomizing the studies at an arbitrary threshold, I adopt a sensitivity analysis approach in Figure 4, in which I vary the minimum acceptable elite sample size from the smallest recorded in the dataset, to the largest, recording the proportion of difference-in-difference estimates that are statistically significant as we raise the elite sample size threshold. Thus, for example, if the minimum sample size threshold is set to the 10th percentile (corresponding to studies with at least \( N = 37 \) elite subjects), 12% of the difference-in-differences are statistically significant; as the relatively flat line indicates, we obtain similar results as we raise the threshold up to around the 75th percentile (corresponding to studies with at least \( N = 873 \) elite subjects); it is only when we raise the threshold to above the 85th percentile that the proportion of significant differences noticeably increases. This suggests that the relatively small elite public gaps reported in the main text are less likely to be an artifact of small elite sample sizes; in fact, the elite-public gaps obtained in these studies are slightly larger the smaller the elite sample is (\( r = -0.20 \)).

Because our dependent variable scales are bounded, we might be worried our findings about minimal elite-public gaps are downwardly biased because of compression. If elite respondents in the control group are significantly more (less) supportive of a policy than mass respondents in the control, the ATE for elite respondents may face ceiling (floor) effects and be artificially smaller than the ATE for mass respondents, simply because the treatment has less room to exert an impact. Figure 5 shows it is indeed the case that difference-in-differences decrease in value when elite baseline support is higher, but an examination of the data finds very few cases where compression concerns apply.

For example, the largest negative difference in cell means between elites and masses in the control condition occurs in Busby et al. (2020). In their mass public sample, they find that support for the use of force to address the Iranian nuclear program unilaterally is \( E[Y_{i00}] = 0.67 \) on a scale of 0-1, which barely increases when going multilaterally through the United Nations, to \( E[Y_{i10}] = 0.69 \). Among foreign policy elites, they find unilateral support is \( E[Y_{i01}] = 0.35 \), which jumps in the UN condition to \( E[Y_{i11}] = 0.55 \). The treatment effect is thus significantly larger for elites than for masses, and the difference-in-difference is significant. Nonetheless, it is not clear why a ceiling effect should apply at only 0.67 on a 0-1 point scale, and if anything, ceiling effects here would cause us to be overestimating the magnitude of elite-mass differences. Similarly, the largest positive difference in cell means between elites and masses in the control condition occurs in Mintz, Redd and Vedlitz (2006), where 58.3% of the undergraduate sample in the control condition wants to do something (rather than do nothing) in the foreign policy scenario, while 92.0% of the military sample in the control condition wants to do something. In this particular example, neither treatment exerts a significant effect.
Figure 5 shows that the difference-in-differences decrease in value when elite support in the baseline condition is higher, but an examination of the data suggests there is little reason to be concerned about compression effects.

Figure 1 in the main text analyses elite-public gaps in decision-making by presenting difference-in-difference estimates along with 95% confidence intervals. Since our interest in elite-public gaps is about magnitude rather than direction, the figure transforms the differences-in-difference point estimates by taking their absolute value. While this quantity of interest is likely intuitive to many experimentalists (to what extent does the difference between treatment and control in each experiment vary between elites and masses?), the absolute value transformation applies to the point estimates rather than their confidence intervals (which remain symmetric), suggesting the value of alternative visualizations of the data. Figure 7 below thus adopts a different tack, calculating the absolute value of the difference-in-difference estimates, and plotting the density distributions of these observed effect sizes in red. It then generates a simulated null distribution, sampling from a normal distribution with a mean of zero and the observed sample variance $N$ times, taking the absolute value of the samples, and plotting the simulated null in blue. The plot illustrates the same pattern in the main results, reiterated by a Kolmogorov-Smirnov test ($D = 0.23$, $p < 0.01$): we see a relatively high proportion of small effect sizes indicating relatively minor differences between elites and masses in these paired experiments, but also a small number of relatively large effects – suggesting the important of modeling this heterogeneity explicitly with a meta-regression model to theoretically explain the conditions in which elite-public gaps in experiments are larger or smaller.

Finally, Figure 6 replicates the difference-in-difference analysis from the main text, but without taking absolute values, letting us study the direction of the gap rather than just the distance. The plot shows
that 94 (58%) of the 162 difference-in-differences are positive; of the 19 difference-in-differences that are statistically significant once the false discovery rate is controlled for (shown in white), 15 (79%) are positive in a sign. On the whole, then, it appears that elites tend to be more, rather than less, responsive to the treatments. It is therefore not the case that elites are less likely to respond to the treatments because they possess stronger prior attitudes, for example.

Figure 6: Difference-in-differences between elites and masses across paired treatments

Figure 6 replicates the analysis on the main text, but without taking the absolute value of the difference-in-difference, letting us test whether experimental results on the two samples differ by direction rather than just by distance. Estimates statistically significant after controlling for multiple comparisons are shown in white; estimates statistically significant only as long as multiple comparisons are not controlled for are shown in grey. The plot shows that elites generally appear more responsive to the study treatments than mass samples are; of the 19 statistically significant differences depicted in white, 15 (79%) are positive in sign.
Figure 7 compares the observed density distribution of our quantity of interest (the absolute value of the difference-in-differences between treatment and control, between elites and masses, for each of our paired treatment effects) in red, with a simulated null distribution in blue (sampling from a normal distribution with a mean of zero and the observed sample variance, and taking the absolute value to produce a half-normal distribution). The pattern confirms the results from the main analysis: we see a relatively high proportion of small effect sizes indicating relatively small differences between elites and masses in these paired experiments, but also a small number of relatively large effects – suggesting the importance of modeling this heterogeneity explicitly with a meta-regression model to theoretically explain the conditions in which elite-public gaps in experiments are larger or smaller.
1.3 Tests for publication bias

To test for publication bias, I estimate a pair of random-effects models: one for the average treatment effects from the experiments reviewed here, and another for the difference-in-difference estimates. They suggest little evidence of publication bias. This may reflect practices in this area of research, or it could also reflect the fact that the quantity of interest I employ here — the difference-in-difference between treatment and control between elites and masses — differs from that used by the authors themselves.\footnote{Li and Owen (2020) show that publication bias is usually most severe when the quantity under investigation is the central focus of the paper.} Because all of the studies included here were analyzed the same way, it also meant that some of the studies were analyzed slightly differently than as conducted by the original authors: for example, some of the original studies were less interested in the average treatment effect than in conditional effects, which I omit here to facilitate comparability across studies.

Figure 8: Funnel plots

(a) Average treatment effects

(b) Difference-in-differences

Figure 8 suggests little concern for publication bias among the studies included here.

1.3.1 Additional metaregression tests

The paired experiments presented in the main text vary from one another in multiple ways, from the types of samples used, to the types of questions studied. The meta-regression in the main text seeks to model this heterogeneity theoretically, using experiment-level covariates to explore the conditions in which elite-public gaps in decision-making are larger or smaller. The analysis in the main text uses GDP per capita data from the International Monetary Fund as a contextual covariate, to test whether elite-public gaps are starker
Table 1: Metaregression results: additional contextual moderators

<table>
<thead>
<tr>
<th>Contextual moderator</th>
<th>GDP per capita</th>
<th>GINI</th>
<th>log(Population)</th>
<th>Individualism</th>
<th>Number of experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>p</td>
<td>β</td>
<td>p</td>
<td>β</td>
</tr>
<tr>
<td>Military elite</td>
<td>-0.024</td>
<td>0.470</td>
<td>-0.024</td>
<td>0.472</td>
<td>-0.022</td>
</tr>
<tr>
<td>Political elite</td>
<td>0.079</td>
<td>0.049</td>
<td>0.065</td>
<td>0.055</td>
<td>0.070</td>
</tr>
<tr>
<td>Student sample</td>
<td>0.125</td>
<td>0.011</td>
<td>0.133</td>
<td>0.008</td>
<td>0.134</td>
</tr>
<tr>
<td>Domestic politics</td>
<td>0.050</td>
<td>0.118</td>
<td>0.052</td>
<td>0.098</td>
<td>0.051</td>
</tr>
<tr>
<td>IPE</td>
<td>0.026</td>
<td>0.738</td>
<td>0.008</td>
<td>0.837</td>
<td>-0.001</td>
</tr>
<tr>
<td>Representation</td>
<td>0.217</td>
<td>0.042</td>
<td>0.187</td>
<td>0.005</td>
<td>0.181</td>
</tr>
<tr>
<td>International security</td>
<td>0.106</td>
<td>0.030</td>
<td>0.107</td>
<td>0.023</td>
<td>0.103</td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.088</td>
<td>0.400</td>
<td>-0.019</td>
<td>0.840</td>
<td>-0.050</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Embeddedness</th>
<th>Obedience</th>
<th>General trust</th>
<th>UN Parking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contextual moderator</td>
<td>β</td>
<td>p</td>
<td>β</td>
<td>p</td>
</tr>
<tr>
<td>Military elite</td>
<td>-0.008</td>
<td>0.695</td>
<td>0</td>
<td>0.960</td>
</tr>
<tr>
<td>Political elite</td>
<td>-0.040</td>
<td>0.197</td>
<td>-0.037</td>
<td>0.224</td>
</tr>
<tr>
<td>Student sample</td>
<td>0.041</td>
<td>0.101</td>
<td>0.044</td>
<td>0.072</td>
</tr>
<tr>
<td>Domestic politics</td>
<td>0.129</td>
<td>0.017</td>
<td>0.129</td>
<td>0.016</td>
</tr>
<tr>
<td>IPE</td>
<td>0.016</td>
<td>0.814</td>
<td>0</td>
<td>0.996</td>
</tr>
<tr>
<td>Representation</td>
<td>0.143</td>
<td>0.043</td>
<td>0.130</td>
<td>0.078</td>
</tr>
<tr>
<td>International security</td>
<td>0.096</td>
<td>0.033</td>
<td>0.088</td>
<td>0.044</td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.031</td>
<td>0.373</td>
<td>-0.021</td>
<td>0.610</td>
</tr>
</tbody>
</table>

Random effects meta-regression with clustered SEs at the experiment level. Each model includes a different contextual moderator, listed above.
in developing countries than industrialized ones. The analysis finds little support for this hypothesis. Yet there are a wide variety of other potential contextual factors one could test for. In Table 1, I replicate the analysis from the main text in model 1, and then turn to seven other country-level characteristics. Model 2 uses GINI coefficient data from the World Bank to test whether elite-public gaps are starker in countries with greater income inequality. Model 3 uses logged population data to test whether elite-public gaps are larger in bigger countries rather than smaller ones. Model 4 uses data from Hofstede’s individualism scale, frequently used in cross-cultural psychology and anthropology (Hofstede, 2001). Its use here thus lets us test whether collectivist societies feature larger or smaller elite-public gaps than individualist ones. Model 5 uses the Schwartz value scheme (as administered on World Values Survey) to test whether experiments fielded in cultures scoring high in embeddedness (an emphasis on traditional order, in-group solidarity, respect for tradition, and so on) display larger elite-public gaps (Schwartz, 1992). Model 6 uses child-rearing measures from World Values Survey data to test whether the size of elite-public gaps is larger in cultures emphasizing obedience; similar child-rearing measures are often used at the individual-level to study right-wing authoritarianism in US contexts (Hetherington and Weiler, 2009). Model 7 uses a measure of generalized trust from the World Values Survey, to test whether the magnitude of elite-public gaps differs from highly trusting societies (Uslaner, 2002). Finally, model 8 uses data from Fisman and Miguel (2007), who obtain the number of unpaid New York City parking violations by diplomats at the United Nations, whose diplomatic plates granted them immunity until 2002. Fisman and Miguel use this data (calculated as the number of violations per diplomat) as a measure of country corruption norms, showing it is strongly correlated with survey-based country-level corruption measures; I use it here to test whether political elites from countries whose diplomats are more likely to feel above the law tend to display larger elite-public gaps. As before, all meta-regression models feature standard errors clustered at the experiment level.

Table 1 suggests two key patterns. First, these country-level variables never approach statistical significance. It is possible that other contextual factors would significantly affect the size of elite-public gaps, or that these country-level characteristics would significantly predict the size of elite-public gaps if the studies were fielded in a wider set of countries: for example, due to the difficulties gaining access to political elites in non-democratic societies, the studies in the meta-analysis tended to be conducted in democratic contexts. Nonetheless, these eight contextual factors don’t appear to explain variation in elite-public gaps in the 12 countries in which these experiments were fielded.

Second, the effects of the other variables are fairly consistent across the models, even though missing data for the country-level covariates in models 4-8 means that the analysis is also being executed on different

---

6 Data for the latter five of the seven country-level covariates come from Schulz et al. (2019).

7 Uganda, classified as a closed anocracy in the Polity IV data, is a potential exception.
subsamples of countries. Across all models, the substantively largest effects come from the topic of the experiment: representation experiments, in which elites and the public are asked to predict how the public will react, or elites and the public are asked to assess how elites will behave, display particularly large differences.

Finally, one of the findings from Table 1 in the main text was that elite-mass studies that use student samples obtain significantly larger elite-public gaps than those that use more diverse adult samples. However, Table 2 suggests this result is being driven by a significant interaction effect between the type of elite and public sample: elite-public gaps are particularly pronounced when the elite sample consists of military officials, and the public sample consists of undergraduates.

Table 2: Metaregression results: elite-public interaction

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>SE</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.172</td>
<td>0.132</td>
<td>0.214</td>
</tr>
<tr>
<td>GDP per capita</td>
<td>0.002</td>
<td>0.002</td>
<td>0.315</td>
</tr>
<tr>
<td>Military elite</td>
<td>-0.043</td>
<td>0.026</td>
<td>0.160</td>
</tr>
<tr>
<td>Political elite</td>
<td>0.087</td>
<td>0.035</td>
<td>0.040</td>
</tr>
<tr>
<td>Student sample</td>
<td>-0.029</td>
<td>0.097</td>
<td>0.772</td>
</tr>
<tr>
<td>Military elite $\times$ student sample</td>
<td>0.276</td>
<td>0.098</td>
<td>0.012</td>
</tr>
<tr>
<td>Political elite $\times$ student sample</td>
<td>0.139</td>
<td>0.108</td>
<td>0.222</td>
</tr>
<tr>
<td>Domestic politics</td>
<td>0.048</td>
<td>0.023</td>
<td>0.149</td>
</tr>
<tr>
<td>IPE</td>
<td>0.112</td>
<td>0.102</td>
<td>0.290</td>
</tr>
<tr>
<td>Representation</td>
<td>0.294</td>
<td>0.121</td>
<td>0.029</td>
</tr>
<tr>
<td>International security</td>
<td>0.110</td>
<td>0.034</td>
<td>0.026</td>
</tr>
</tbody>
</table>

Random-effects meta-regression with clustered SEs at the experiment-level. Reference categories for elite sample characteristics: bureaucrats or heterogeneous elite samples; for experimental domain: apolitical tasks (e.g. stylized experimental economic games).

1.4 List of studies included in the meta-analysis

The meta-analysis in the main text is conducted on 162 paired treatment effects from 48 paired experiments across the following 26 studies comparing elites and masses. See the main text for a discussion of the inclusion criteria.


---

*For example, model 8 drops all experiments conducted in the United States from the analysis, since US diplomats to the UN did not have diplomatic immunity.*

Christensen, Julian and Donald P. Moynihan. 2020. “Motivated reasoning and policy information: Politicians are more resistant to debiasing interventions than the general public.” Working paper.


Martin, Lucy and Pia Raffler. 2020. “Fault Lines: The Effects of Bureaucratic Power on Electoral Account-
ability.” American Journal of Political Science Forthcoming.


2 Observational data

2.1 Demographics

See Figures 9-11 for mosaic plots of the age, gender, and education distributions across the elite and mass samples; the coverage of demographic variables across survey waves has important implications for the empirical strategy, as noted in Appendix §2.2 below. The figures show important changes over time: for example, in 1975 and 1998, the elite sample mostly consisted of 40-64 year olds, thereby skewing older than the public as a whole. By the most recent surveys, the age composition of the elite sample shifted, perhaps reflecting survey mode differences as the elite surveys went online.
Figure 9: Age distributions in elite and mass samples

(a) 1975

(b) 1998

(c) 2002

(d) 2004

(e) 2014

(f) 2016

(g) 2018
Figure 10: Gender distributions in elite and mass samples

(a) 1975

(b) 1986

(c) 1998

(d) 2002

(e) 2004

(f) 2014

(g) 2016

(h) 2018
Figure 11: Education distributions in elite and mass samples

(a) 1998

(b) 2014

(c) 2016

(d) 2018
2.2 Alternative empirical strategies

The analysis of the observational data in the main text uses a simple simulation method, in which, for each wave of the survey, I bootstrap the public data 5000 times; in each bootstrap, I estimate a series of linear regression models, in which each of the foreign policy questions asked in a given year is regressed on a series of dummy variables denoting whether the respondent has a college degree, is male, has an income at above the median-level for the public sample in that year, and is aged 40 or above. I then save the predicted values, thereby obtaining an estimate of what a counterfactual public sample composed of 40+ year-old males with college degrees and above-median incomes would think about each foreign policy issue.\(^9\) The virtue of this approach is that it follows the call of Manski (2013), Glynn and Ichino (2015), and others to leverage qualitative knowledge to motivate quantitative analyses.

There are, of course, a range of other methods one could use to study the effects of compositional differences on elite-public gaps, each of which comes with a different set of tradeoffs. One such possibility is matching (Iacus, King and Porro, 2012), in which elites and non-elites would be matched on their demographic characteristics in order to identify the effect of “eliteness”, all while avoiding functional form assumptions. The challenge is that many of the years of Chicago Council data feature spotty coverage of demographic characteristics. 4 of the 12 waves are missing information about gender; 5 of the 12 waves are missing age; 8 of the 12 waves are missing education; 12 of the 12 waves are missing income. This high level of missingness is problematic given matching’s selection on observables identification strategy (Sekhon, 2009, 496). The fact that income is missing for all observations, then, means that the effect of eliteness is always confounded with that of income, rendering the identification strategy implausible.

Below I turn to two alternative approaches. First, I use a subsetting method, which similarly constructs a counterfactual public, but without making the linear functional form assumptions from the bootstrapping approach. Then I use decomposition analysis, a method frequently used in labor economics for the types of problems being studied here.

2.2.1 Subsetting method

Since this simulation approach relies on linear functional form assumptions, in Table 3 below I use a subsetting method instead, in which I simply save the subset of respondents in a given year who are male, aged 40 or above, with a college degree, and with income at above the median-level. The results below show that we obtain similar results to the bootstrapping method, with adjusted elite-mass correlations ranging from 0.59 (in 2016) to 0.91 (in 1986).

---

\(^9\)See Figure 12 for an alternative visualization of the simulation results, focusing on the magnitude of elite-public gaps before and after adjustment.
Figure 12: Elite-public gaps, before and after adjusting for compositional differences (simulation method)
Table 3: Decomposing elite-public gaps using subsetting method

<table>
<thead>
<tr>
<th>Year</th>
<th>$\gamma$</th>
<th>$Z$</th>
<th>$r$</th>
<th>$\Delta r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1975</td>
<td>0.321</td>
<td>0.892</td>
<td>0.127</td>
</tr>
<tr>
<td>2</td>
<td>1978</td>
<td>0.248</td>
<td>0.815</td>
<td>0.175</td>
</tr>
<tr>
<td>3</td>
<td>1982</td>
<td>0.312</td>
<td>0.86</td>
<td>0.175</td>
</tr>
<tr>
<td>4</td>
<td>1986</td>
<td>0.295</td>
<td>0.905</td>
<td>0.11</td>
</tr>
<tr>
<td>5</td>
<td>1990</td>
<td>0.293</td>
<td>0.868</td>
<td>0.129</td>
</tr>
<tr>
<td>6</td>
<td>1994</td>
<td>0.347</td>
<td>0.877</td>
<td>0.179</td>
</tr>
<tr>
<td>7</td>
<td>1998</td>
<td>0.343</td>
<td>0.793</td>
<td>0.252</td>
</tr>
<tr>
<td>8</td>
<td>2002</td>
<td>0.31</td>
<td>0.874</td>
<td>0.142</td>
</tr>
<tr>
<td>9</td>
<td>2004</td>
<td>0.256</td>
<td>0.804</td>
<td>0.158</td>
</tr>
<tr>
<td>10</td>
<td>2014</td>
<td>0.261</td>
<td>0.864</td>
<td>0.124</td>
</tr>
<tr>
<td>11</td>
<td>2016</td>
<td>0.152</td>
<td>0.588</td>
<td>0.208</td>
</tr>
<tr>
<td>12</td>
<td>2018</td>
<td>0.144</td>
<td>0.762</td>
<td>0.055</td>
</tr>
</tbody>
</table>

Figure 13: Partisan divides between elite and mass samples over time
The subsetting method is also useful because we can also use it to adjust for partisanship. As Figure 13 shows, the partisan composition of the elite sample departs dramatically from that of the mass public sample beginning in the mid-2000s, with the elite sample eventually becoming disproportionately Democratic. This skew is especially noticeable in 2016 and 2018, when there are 3.1 and 5.1 times as many Democrats as Republicans in the elite sample, respectively, despite there being only 1.1 and 1.3 times as many Democrats as Republicans in the public sample during this same time period. This imbalance presumably isn’t due to foreign policy leaders becoming less Republican over time, but rather, due to changes in the survey’s sampling strategy and response rates. It therefore raises the possibility that some of the elite-public gap during these years are due to divergent partisan compositions of the two samples. Table 4 therefore uses the same subsetting method as in Table 3, but this time also subsetting on partisanship.\footnote{Partisanship information is not available for the public sample in 1994, so that year’s observations are dropped from both Figure 13 and Table 4.}

The left-hand panel compares the policy preferences of the elite sample in a given year with those from the subset of respondents in the public sample who are male, aged 40 or above, with a college degree, income at above the median-level, and who identify as Democrats. Since this is a comparison between a Democratic public and elites of all partisan affiliations, the adjustment performs poorly for most years, but performs well for 2016 and 2018, the two years where the partisan imbalance in the elite sample is the most skewed: 40-49% of the elite-public gap during these years can be accounted for by adjusting for these four demographic variables, raising the elite-public correlations from $r = 0.59$ and $r = 0.76$ in Table 3 to $r = 0.79$ and $r = 0.95$, respectively.

The next two panels make within-party comparisons, comparing Democratic members of the public with Democratic elites (in the middle panel), and Republican members of the public with Republican elites (in the right-hand panel); as before, the public subsamples also subset on gender, age, education, and income. Here, the correlations don’t markedly improve, apart from 2016 and 2018 (the two years where the partisan mismatch between elites and masses is the most stark), and are usually higher between Democrats than between Republicans — raising important questions for future work.

### 2.2.2 Blinder-Oaxaca decomposition analysis

An alternative approach well suited to studying elite-public gaps, and less clunky than the subsetting analysis above, is Blinder-Oaxaca decomposition analysis (Blinder, 1973; Oaxaca, 1973). Although less common in political science (for an exception, see Chiba, Machain and Reed (2014)), it is frequently used in labor economics to determine how much of a gap between two groups is due to observable characteristics rather than unobservable ones. The canonical example involves decomposing the gender gap in wages, testing the extent to which the gender gap is attributable to men and women differing from one another in observable
characteristics (such as education, or the number of hours worked), versus unobservable ones (such as discrimination).

Given spotty coverage of demographic information across survey waves, I restrict the analysis to the four waves of survey data that include information about elites’ ages, gender, and education: 1998, 2014, 2016, and 2018. These waves also include other covariates, such that where possible, I also control for partisanship (available all four waves), race (available 2014-18), and political ideology (available 1998, 2014 and 2016). I then estimate a twofold Blinder-Oaxaca decomposition analysis, in which the gap between masses \( (\bar{Y}_A - \bar{Y}_B) \) is expressed as \( \Delta \bar{Y} = (\bar{X}_A - \bar{X}_B)' \beta_B + \bar{X}_A (\beta_A - \beta_B) \), where the first term captures the portion of the elite-public gap attributable to compositional differences, and the second captures the unexplained portion.

I estimate the proportion of the gap attributable to compositional differences for each foreign policy question in each year, and plot the density distribution of estimates in Figure 14 below, in which the red vertical line indicates the median proportion attributable to demographic differences in a given year – analogous to \( \frac{\gamma}{2} \) from the theoretical framework in the main text. In 1998, the median proportion of the elite-public gap attributable to compositional differences is 0.49 in 1998, 0.71 in 2014, 0.56 in 2016, and 0.70 in 2018. A direct comparison to the simulations from the main text is somewhat difficult by the fact that the decomposition analysis uses a more extensive set of covariates than the simulation method does, but

---

11 The cost of this restriction is that we exclude roughly 73% of the questions in the foreign policy dataset, which is why the analysis in the main text uses a coarser simulation approach instead.

12 I use the Neumark (1988) decomposition, estimating results using the \texttt{Oaxaca} package in \texttt{R} (Hlavac, 2018).
the key takeaway is similar: a non-trivial proportion of the elite-public gaps can be attributed to relatively mundane compositional effects — and if anything, the findings in the main text appear to be conservative.

Figure 14: Two-way decomposition analysis

The vertical line presents the median proportion of elite-public gaps in a given year attributable to compositional differences, analogous to $\frac{\gamma}{2}$ from the main text. The plot truncates the x axis for purposes of legibility.

2.3 Recalculating elite-public gaps using a more restrictive set of foreign policy elites

The analysis in the main text calculates elite public gaps using the full sample of elites from each wave of the CCGA polls. As noted in the main text, these heterogeneous elite samples contain a variety of different types of foreign policy opinion leaders. The codebook to the 1986 sample, for example, reports a sampling frame of leaders drawn from the administration (e.g. the White House Office, National Security Council, State Department), and members of the House of Representatives and Senate drawn from specific committees and subcommittees germane to foreign affairs (e.g. the Committee on Armed Services, Committee on Foreign Affairs, the Subcommittee on Trade from the Committee on Ways and Means), but also “vice presidents in charge of international affairs” from the top 100 Fortune 500 companies, presidents of the largest 50 labor unions, members of the media (e.g. news directors and network newscasters from ABC, NBC, and CBS, editors of Foreign Policy, Foreign Affairs, and so on), academics (presidents and faculty who teach foreign affairs at top universities), religious leaders, special interest groups, and presidents from foreign policy
organizations/think tanks. The notion of who counts as “elite” in foreign policy is somewhat ill-defined in the existing literature (see Hafner-Burton, Hughes and Victor 2013; Bahador, Entman and Knüpfer 2019 for definitions of political elites more broadly), but the heterogeneous elite sample used here captures the breadth IR scholars often have in mind when they invoke the existence of a “foreign policy establishment” stretching from Capitol Hill to Foggy Bottom, the Ivory Tower to Wall Street (Busby and Monten, 2008).\textsuperscript{13}

On the one hand, one might argue that the diversity of this sample might cause us to underestimate the magnitude of the elite-public gap, under the assumption that senators are more “elite” than think tank fellows, say, such that compositional effects would account for less of the elite-public gap amongst a narrower set of elites. On the other hand, one might argue that the heterogeneity of this sample might cause us to overestimate the magnitude of the elite-public gap, under the assumption that individuals who are exposed to electoral pressures have a greater incentive to think more similarly to the public, due to the constraints of accountability (Sheffer et al., 2018).

As a robustness check, I use the same bootstrapping procedure as before to estimate elite-public gaps adjusting for basic demographic differences, but this time restricting the elite sample to elites from either the executive or legislative branches of government at the time the survey is fielded, for a total of 1069 elites. This is a conservative measure, since it excludes individuals who were previously in government and have since migrated to think tanks, interest groups, or who were now serving as professors of practice in academia. In the 2014 survey, for example, when elite respondents were asked to indicate whether they had served in government, 51 of 52 think tank respondents had prior government service, as did 17 of 36 interest group respondents, and 98 of 192 academics.\textsuperscript{14} In this sense, some caution should be taken interpreting the results below given the significant reduction in the effective sample size.

Figure 15 presents a scatterplot similar to Figure 3 in the main text. As before, it presents both the unadjusted (in light red) and adjusted (in blue) public preferences. Unlike in the main text, however, the preferences of this narrower subsample of political elites in government positions is depicted on the y axis, rather than that of the full sample of political elites. The red line depicts a loess smoother using the unadjusted data, while the blue line depicts a loess smoother using the adjusted data. As before, however, it’s evident that adjusting for basic demographic differences makes the public responses more closely resemble the elite sample, and the statistics in each panel show that once we adjust for basic compositional differences

\textsuperscript{13}In this sense, both the experimental and observational analyses feature heterogeneity, but in different ways: the meta-analysis on experimental data features more geographic heterogeneity (including experiments fielded in 12 different countries) but less temporal heterogeneity (unlike the observational analysis, which includes survey questions fielded in a single country, but over a 43 year period). Topically, the experimental analysis is broader than the observational analysis (featuring experiments about domestic politics, IPE, international security, political representation, and relatively apolitical studies), but the observational analysis also goes into greater depth (with 1504 polling questions encompassing a wide range of foreign policy questions, from the Vietnam War to the Transpacific Partnership, such that it covers far more than just a single case).

\textsuperscript{14}This also excludes respondents from the “Military” group in the 2014 survey, which consisted of alumni from the National War College (NWC), 250 of 278 of whom indicated they had served in government.
Figure 15: Government elite and mass foreign policy preferences, after adjusting for compositional differences

<table>
<thead>
<tr>
<th>Year</th>
<th>Elites</th>
<th>Mass</th>
<th>Gap</th>
<th>Adjusted</th>
<th>Unadjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1975</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
<td><img src="image4" alt="Graph" /></td>
<td><img src="image5" alt="Graph" /></td>
</tr>
<tr>
<td>1978</td>
<td><img src="image6" alt="Graph" /></td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
<td><img src="image9" alt="Graph" /></td>
<td><img src="image10" alt="Graph" /></td>
</tr>
<tr>
<td>1982</td>
<td><img src="image11" alt="Graph" /></td>
<td><img src="image12" alt="Graph" /></td>
<td><img src="image13" alt="Graph" /></td>
<td><img src="image14" alt="Graph" /></td>
<td><img src="image15" alt="Graph" /></td>
</tr>
<tr>
<td>1986</td>
<td><img src="image16" alt="Graph" /></td>
<td><img src="image17" alt="Graph" /></td>
<td><img src="image18" alt="Graph" /></td>
<td><img src="image19" alt="Graph" /></td>
<td><img src="image20" alt="Graph" /></td>
</tr>
<tr>
<td>1990</td>
<td><img src="image21" alt="Graph" /></td>
<td><img src="image22" alt="Graph" /></td>
<td><img src="image23" alt="Graph" /></td>
<td><img src="image24" alt="Graph" /></td>
<td><img src="image25" alt="Graph" /></td>
</tr>
<tr>
<td>1994</td>
<td><img src="image26" alt="Graph" /></td>
<td><img src="image27" alt="Graph" /></td>
<td><img src="image28" alt="Graph" /></td>
<td><img src="image29" alt="Graph" /></td>
<td><img src="image30" alt="Graph" /></td>
</tr>
<tr>
<td>1998</td>
<td><img src="image31" alt="Graph" /></td>
<td><img src="image32" alt="Graph" /></td>
<td><img src="image33" alt="Graph" /></td>
<td><img src="image34" alt="Graph" /></td>
<td><img src="image35" alt="Graph" /></td>
</tr>
<tr>
<td>2002</td>
<td><img src="image36" alt="Graph" /></td>
<td><img src="image37" alt="Graph" /></td>
<td><img src="image38" alt="Graph" /></td>
<td><img src="image39" alt="Graph" /></td>
<td><img src="image40" alt="Graph" /></td>
</tr>
<tr>
<td>2004</td>
<td><img src="image41" alt="Graph" /></td>
<td><img src="image42" alt="Graph" /></td>
<td><img src="image43" alt="Graph" /></td>
<td><img src="image44" alt="Graph" /></td>
<td><img src="image45" alt="Graph" /></td>
</tr>
<tr>
<td>2014</td>
<td><img src="image46" alt="Graph" /></td>
<td><img src="image47" alt="Graph" /></td>
<td><img src="image48" alt="Graph" /></td>
<td><img src="image49" alt="Graph" /></td>
<td><img src="image50" alt="Graph" /></td>
</tr>
<tr>
<td>2016</td>
<td><img src="image51" alt="Graph" /></td>
<td><img src="image52" alt="Graph" /></td>
<td><img src="image53" alt="Graph" /></td>
<td><img src="image54" alt="Graph" /></td>
<td><img src="image55" alt="Graph" /></td>
</tr>
<tr>
<td>2018</td>
<td><img src="image56" alt="Graph" /></td>
<td><img src="image57" alt="Graph" /></td>
<td><img src="image58" alt="Graph" /></td>
<td><img src="image59" alt="Graph" /></td>
<td><img src="image60" alt="Graph" /></td>
</tr>
</tbody>
</table>
between the two samples, the magnitude of the gaps shrinks. While the proportion of the elite-public gap accounted by these demographic measures is slightly lower (by an average of 0.05 units), as is the adjusted correlation coefficient (by an average of 0.046 units), the overall pattern of results is the same as before. If we use the subsetting method instead, as in the previous section, also controlling for partisanship, the performance in 2016 and 2018 improves further, with the adjusted correlation increasing to $r = 0.67$ in 2016, and $r = 0.92$ in 2018. In this sense, the results find little support for assumptions that higher-level elites show substantially larger (or substantially smaller) elite-public gaps.

### 2.4 Are the results the artifact of elite cues?

One potential critique of using observational data to study elite-public gaps might be that elite partisan cues may suppress the magnitude of the gap recovered: if the public’s foreign policy preferences are heavily dependent upon those of trusted political elites, as top-down models of public opinion about foreign policy would suggest (Zaller, 1992; Berinsky, 2009; Baum and Groeling, 2009), the public may already have been pre-treated with information that should cause the elite-public gap to appear smaller than it actually is.

There are several reasons why we might be less concerned about the effects of elite cues here. First, the existence of elite-public gaps poses challenges for elite-driven theories of public opinion in foreign policy: if the public is as receptive to elite messages as top-down models predict, these foreign policy disconnects presumably shouldn’t exist in the first place (Kertzer and Zeitzoff, 2017). Second is the diversity of questions in the dataset, which include specific policy attitudes (which might be more susceptible to elite messages) but also include more general foreign policy orientations, which are usually thought of as being more stable (Rathbun et al., 2016). Moreover, elite cues should be the most powerful for questions that are highly salient and loom large in political discourse, whereas the questions in the dataset vary considerably along this dimension. Third, for elite cues to suppress the magnitude of the elite-public gap, the public would need to be in a one-sided information environment, rather than a polarizing one (Zaller, 1992): this argument should therefore only hold for issues where Republicans and Democrats in Washington are on the same page, rather than for issues where they disagree; if we imagine there has been a secular increase in polarization in foreign policy issues over time, the implication is that the magnitude of the elite-public gap should increase over the course of these surveys, but the analysis in the main text finds little evidence to suggest this is the case.

Where elite cues pose the largest threat to our empirical strategy is in the use of education to decompose the elite-public gap. Education is a variable implicating multiple mechanisms, from political knowledge, to cosmopolitan values, to economic interests (Delli Carpini and Keeter, 1996; Hainmueller and Hiscox,
but of particular relevance here is that education is sometimes used as a proxy for exposure to elite
cues (Zaller, 1992). One potential interpretation of these results, then, is that controlling for demographic
differences between elites and masses also inadvertently proxies for the effects of exposure to elite cues. I
therefore replicate the main analysis, but condition on engagement with the news, in lieu of education —
the former being a more direct measure of exposure to elite cues than the latter, since the news serves as
the primary vector through which political information is transmitted (e.g. Gadarian, 2010). As Table 5
below shows, when interest in the news is used in lieu of education, the proportion of the elite-gap accounted
for by basic demographics substantively shrinks in size compared to the standard models in the main text,
by an average of 7.7 percentage points per survey wave. This suggests that the analysis in the main text
isn’t simply capturing the effect of elite cues. Since disentangling the effects of education from exposure to
elite cues is inevitably difficult with secondary observational data, future work should extend this analysis
by obtaining more direct measures of exposures to elite cues.

Table 5: Decomposing elite-public gaps using news engagement rather than education

<table>
<thead>
<tr>
<th>Year</th>
<th>γ</th>
<th>Z</th>
<th>r</th>
<th>Δr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1975</td>
<td>0.165</td>
<td>0.845</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>1978</td>
<td>0.186</td>
<td>0.766</td>
<td>0.127</td>
</tr>
<tr>
<td>3</td>
<td>1982</td>
<td>0.212</td>
<td>0.806</td>
<td>0.121</td>
</tr>
<tr>
<td>4</td>
<td>1986</td>
<td>0.157</td>
<td>0.869</td>
<td>0.073</td>
</tr>
<tr>
<td>5</td>
<td>1990</td>
<td>0.168</td>
<td>0.81</td>
<td>0.071</td>
</tr>
<tr>
<td>6</td>
<td>1994</td>
<td>0.219</td>
<td>0.814</td>
<td>0.115</td>
</tr>
<tr>
<td>7</td>
<td>1998</td>
<td>0.182</td>
<td>0.688</td>
<td>0.147</td>
</tr>
<tr>
<td>8</td>
<td>2002</td>
<td>0.081</td>
<td>0.714</td>
<td>0.068</td>
</tr>
<tr>
<td>9</td>
<td>2004</td>
<td>0.149</td>
<td>0.741</td>
<td>0.095</td>
</tr>
<tr>
<td>10</td>
<td>2016</td>
<td>0.088</td>
<td>0.489</td>
<td>0.11</td>
</tr>
</tbody>
</table>


Mintz, Alex, Steven B. Redd and Arnold Vedlitz. 2006. “Can We Generalize from Student Experiments to the Real World in Political Science, Military Affairs, and International Relations?” *Journal of Conflict


