Online appendix to Jeffrey A. Friedman, "The World Is More Uncertain Than You Think: Assessing and Combating Overconfidence Among 2,000 National Security Officials"

Contents

1. Cohort information
2. Informational treatment
3. Debiasing estimates
4. Heterogeneity in national security officials' overconfidence, Brier Scores, and receptivity to feedback
5. Consistency of overconfidence across questions

1. Cohort information

This study's data were gathered via nineteen training exercises conducted at the Canadian Forces College, the NATO Defense College, the Norwegian Defense Intelligence School, and the U.S. National War College. In order to maintain anonymity, respondents were asked to provide limited demographic information: gender, active-duty military versus civilian status, and U.S. citizenship. Some cohorts were small enough that asking participants to reveal gender and/or military status would have compromised anonymity, and so those questions were dropped from those surveys. Table S1 describes the composition of each cohort and notes which experimental treatments each cohort received.

Figure S1 shows that judgmental calibration was consistent for every institution that participated in the survey. Figure S2 shows that judgmental calibration was also consistent across gender, military/civilian status, and U.S./non-U.S. citizenship.

Cohort	*N*	*% Female*	*% Active-duty military*	*% Civilian*	*% U.S. citizens*	*Qualitative vs. quantitative judgments*	*Randomly assigned training*	*Randomly assigned question variants*	*Survey includes forecasts*
U.S. National War College, 2015	208	15	75	25	87	✓			✓
U.S. National War College, 2016	224	14	68	32	88	✓	✓		
U.S. National War College, 2017	193	16	68	32	90	✓	✓		
Norwegian Def. Intell. School, 2017	13	.	.	.	0				
Canadian Forces College, 2018	31	.	.	.	0		✓		
Norwegian Def. Intell. School, 2018	13	.	.	.	0				
U.S. National War College, 2018	184	19	70	30	88	✓	✓		
Canadian Forces College, 2019	27	15	67	33	0		✓		
NATO Defense College, 2019	60	43	45	55	7		✓		
Norwegian Def. Intell. School, 2019	12	.	.	.	0				
U.S. National War College, 2019	205	20	74	26	84		✓		
Canadian Forces College, 2020	20	25	70	30	5		✓	✓	
NATO Defense College, 2020	139	9	84	16	7		✓	✓	
NATO Defense College, 2021a	62	3	85	15	5		✓	✓	
NATO Defense College, 2021b	78	12	85	15	10		✓	✓	
NATO Defense College, 2022a	68	7	85	15	4		✓	✓	
U.S. National War College, 2022a	211	17	68	32	86		✓	✓	✓
NATO Defense College, 2022b	59	12	80	20	7		✓	✓	✓
NATO Defense College, 2023	87	11	83	17	3		✓	✓	✓
Total	**1,894**	**16**	**73**	**27**	**59**				

Table S1. *Cohort composition and treatment assignment*.

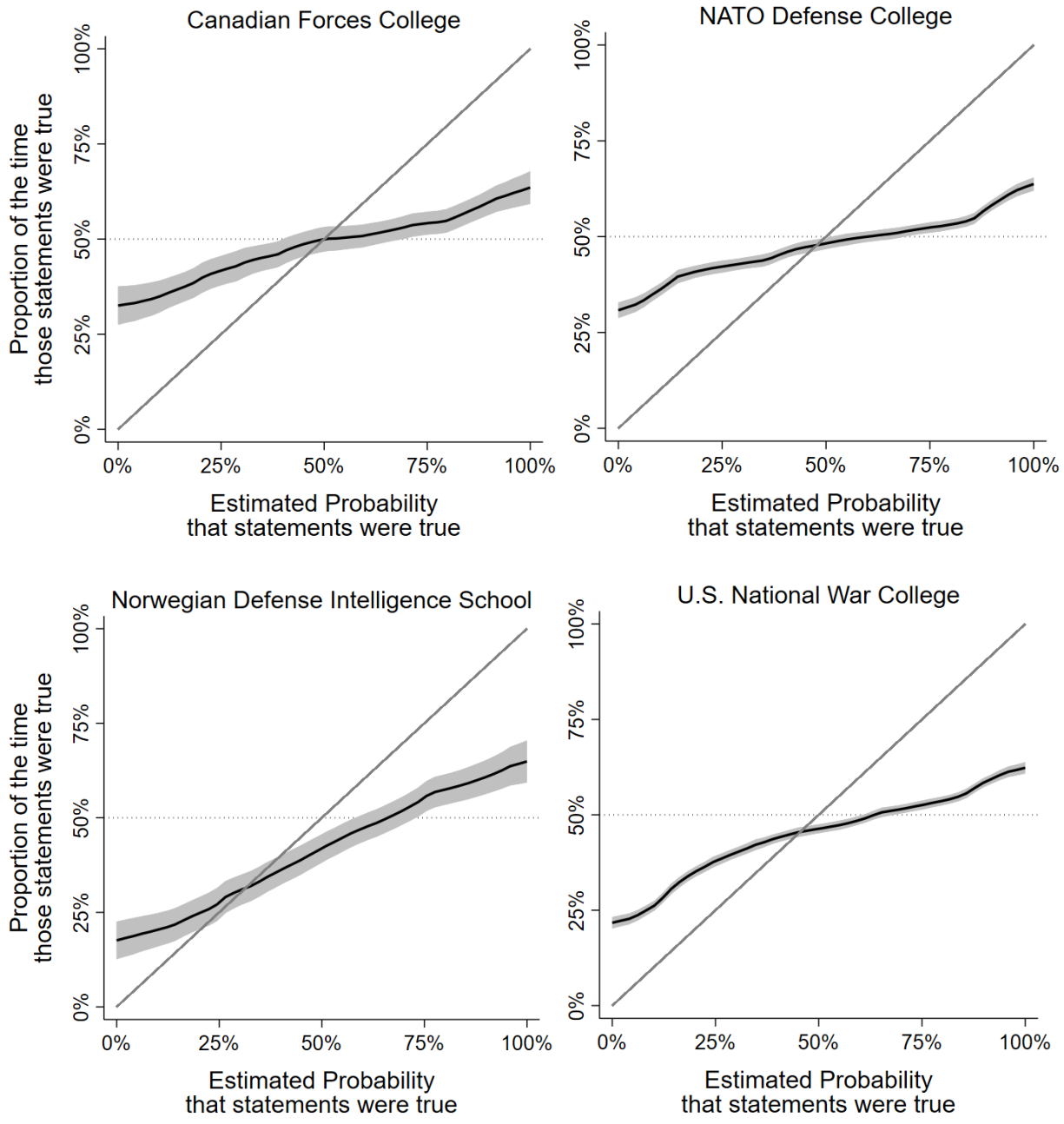


Figure S1. *Performance across institutions*.

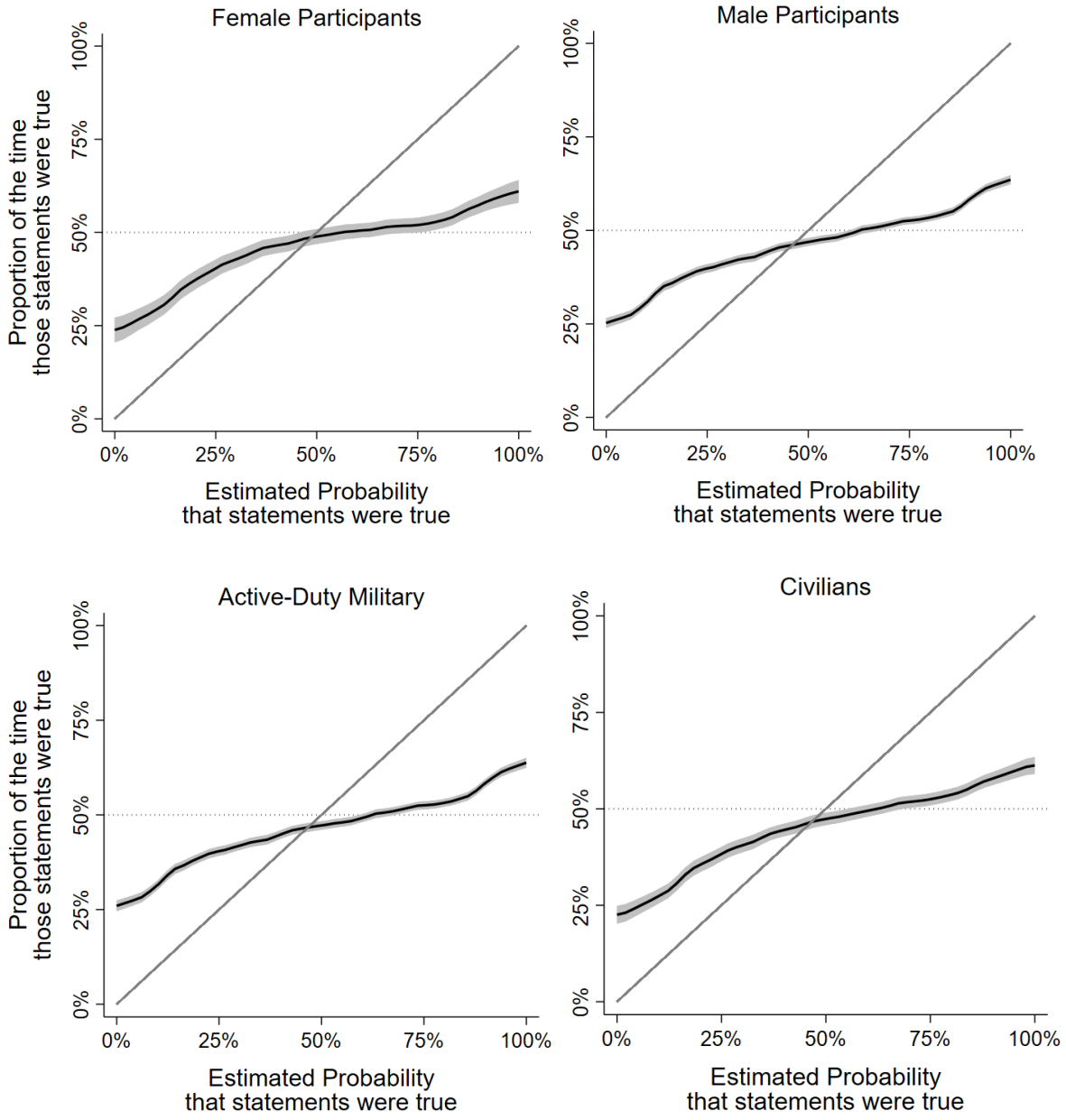


Figure S2. *Performance across demographics*.

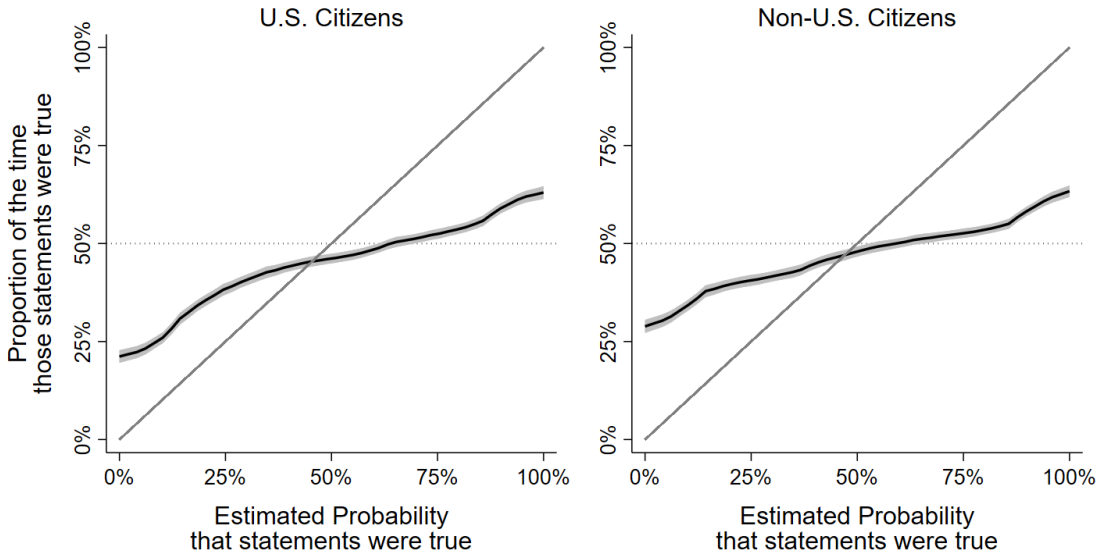


Figure S2 (continued). *Performance across demographics*.

Calibrating assessments of uncertainty

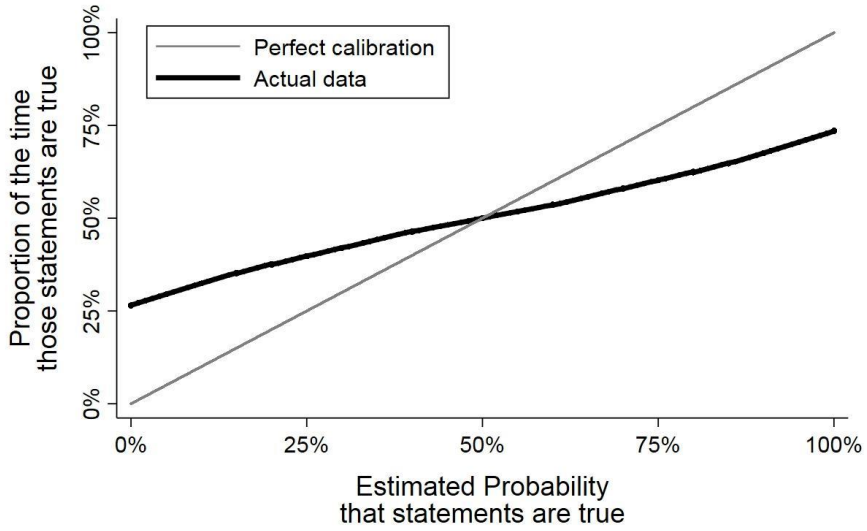
Most people are overconfident when they assess uncertainty. This means that they assign probabilities that are too *high* when they think that statements are likely to be true, and that they assign probabilities that are too *low* when they think that statements are unlikely to be true.

For example, when past participants in surveys like this one have said that they think a statement is 90% likely to be true, we have found that those statements are only true about 65% of the time. When participants have said that they think a statement has a 0% chance of being true, those statements have actually been true about 25% of the time.

<Page Break>

The graph shown below is based on data drawn from previous versions of the survey --it represents roughly 30,000 probability assessments made by nearly 1,000 national security professionals in the United States, Canada, and Europe. The graph's horizontal axis captures the probability that participants assigned to their judgments. The graph's vertical axis captures the proportion of those statements that were actually true.

If participants' judgments were "perfectly calibrated," then the data would fall along a 45-degree line, so that when people said that statements have a 25% chance of being true, those statements would actually be true 25% of the time. You can see that the actual data are significantly miscalibrated as a result of participants' overconfidence.



<Page Break>

Most people therefore perform better on these surveys when they make judgments with lower degrees of certainty. In past surveys, 94 out of every 100 participants would have received better scores if they had shifted all of their estimates closer to 50%. In lecture, we will see whether that same pattern holds with responses to this year's survey as well.

Figure S3. *Informational treatment*.

2. Informational treatment

Figure S3 presents the informational treatment described in the article. This information was generated by analyzing assessments of uncertainty provided by prior survey cohorts.[1] Respondents who were asked to estimate verbal probabilities were provided with similar data based on the analysis shown in Figure 7.

3. Debiasing estimates

Overconfidence was nearly universal among study participants. For these purposes, an individual is considered to be overconfident if we can improve their performance by making every one of their judgments less certain. To determine this fact, we can define the *certitude* of a probability estimate, p, as the absolute value of the difference between p and 0.5. We can then remove a fraction of the certitude associated with each participant's judgments and re-estimate their Brier Score.[2] Continually iterating this process reveals the "optimal" amount of certitude to extract from each participant's judgments in order to minimize their Brier Score. Figure S4 shows that 96 percent of participants would have benefited from making every one of their judgments less certain.[3] On average, participants would have optimized their performance by extracting 62 percent of the certitude from their judgments. (For example, by reducing a probability estimate of

[1] The only difference between these data and those presented in the study's main text are that the calibration curve given to participants was designed to be symmetrical: for example, it treated every probability estimates of 70 percent as also reflecting an estimate of 30 percent, and vice versa. This was intended to make it easier for participants to interpret the graph.
[2] Each iteration removed one percent of the certitude associated with a participant's estimates: for example, compressing a probability estimate of 0.95 to 0.945, or changing an estimate of 0.30 to 0.302.
[3] Data fit with a local polynomial. The shaded area reflects a 95% confidence interval.

80 percent to 61 percent, or by changing a probability estimate of 40 percent to 46 percent.) In other words, most of the certainty that participants expressed in this study was unjustified.

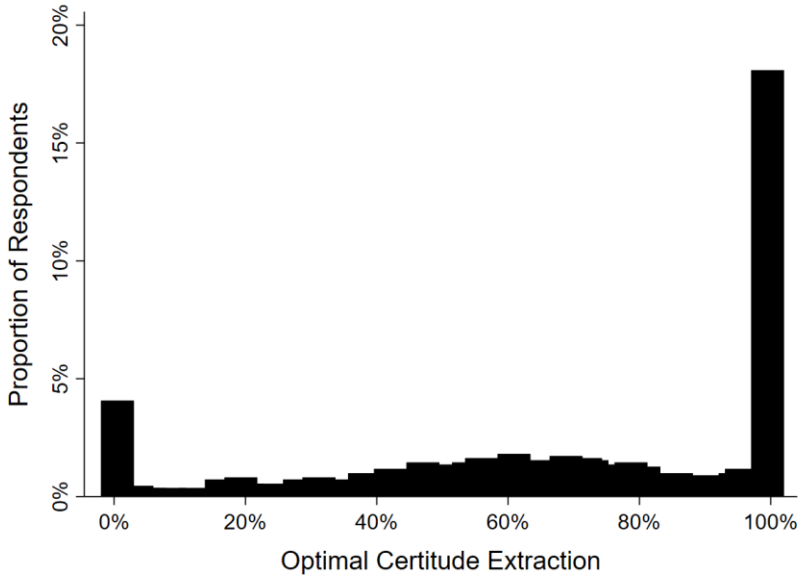


Figure S4. *Extracting certitude from national security officials' judgments systematically improved those judgments' accuracy*.

4. Heterogeneity in national security officials' overconfidence, Brier Scores, and receptivity to feedback

The measure of "optimal certitude extraction" described above provides a useful way of quantifying each national security official's overconfidence. This, in turn, facilitates examining the degree to which some groups of national security officials are more (or less) overconfident than others. Table S2 provides this information by running t-tests comparing different groups of national security officials to each other. Each row in the table represents a separate t-test.

	Group 1	Group 2	Difference [95% confidence interval]	Statistical significance
Gender	Female: 0.66 (N=214)	Male: 0.62 (N=1,187)	-0.03 [-0.08, 0.01]	p=0.13
Active-duty military?	Yes: 0.64 (N=1,044)	No: 0.61 (N=357)	-0.03 [-0.06, 0.01]	p=0.13
U.S. citizen?	Yes: 0.60 (N=720)	No: 0.65 (N=750)	0.06 [0.03, 0.09]	p<0.001

Table S2. *Variations across individuals in overconfidence as measured by "optimal certitude extraction" (i.e. the proportion of certainty in each respondent's estimates that was unjustified).* The overall mean for optimal certitude extraction was 62 percent, with a standard deviation of 29 percent.

Table S2 reveals no statistically-significant differences in respondents' overconfidence with respect to gender or civilian/military status. U.S. citizens tended to be less overconfident than non-U.S. citizens, which may reflect the fact that the survey was administered in English, and would thus have been more legible to native English speakers.[4] This effect is also substantively small, accounting for just one-fifth of a standard deviation in overconfidence across the full sample.

[4] Overconfidence varied across organizations, with respondents from the National War College and the Norwegian Defense Intelligence School displaying less overconfidence than the mean and respondents from the NATO Defense College displaying more overconfidence than the mean (62 percent). Differences across organizations are difficult to interpret since these organizations received different questions tailored to each cohort. This makes it impossible to know whether variations in overconfidence across organizations reflects answering questions that had different degrees of difficulty (especially as respondents at each organization came from different backgrounds and thus possessed different kinds of knowledge).

	Model 1: Gender	Model 2: Mil/Civ Status	Model 3: US/non-US
Received feedback	-0.018 (.004)***	-0.016 (.007)*	-0.012 (.005)*
Female	-0.004 (.007)		
Female * feedback	0.003 (0.010)		
Active-duty military		0.010 (0.006)	
Active-duty military * feedback		-0.002 (.008)	
U.S. citizen			-0.011 (.005)*
U.S. citizen * feedback			-0.011 (.007)
Intercept	0.292 (.003)***	0.284 (.006)***	0.296 (.003)***
N	1,301	1,301	1,332
R^2	0.018	0.021	0.035

Table S3. *Variations across individuals in performance and receptivity to feedback.* The dependent variable in each model is individuals' average Brier Scores. Models estimated via ordinarily least squares (OLS) regression. Sample limited to national security officials who assessed probabilities numerically.

Table S3 presents a similar analysis, examining variation in performance as measured via national security officials' average Brier Scores (mean 0.280, standard deviation 0.063).[5] The table contains three models which explore heterogeneity as a function of gender, military/civilian status and US/non-US citizenship, respectively. Each model includes a binary variable indicating whether the national security official received feedback on other participants' probability estimates and an interaction term that captures whether subsets of the sample were more (or less) receptive to that feedback than others.

These models present similar findings to those in Table S2. They show that neither gender nor military/civilian status were correlated with national security officials' performance in assessing uncertainty. U.S. citizens performed better on the exercise than non-U.S. citizens, but that difference is substantively small (roughly one-sixth of a standard deviation). None of these demographic variations predicted national security officials' receptivity to feedback, as measured by the interaction terms in Table S3.

5. Consistency of overconfidence across questions

To what extent does the overconfidence shown in Figure 1 reflect a generalizable cognitive bias, as opposed to national security officials making particularly poor estimates in response to a small number of questions? In order to resolve this ambiguity, we would ideally calibrate the accuracy of national security officials' judgments in response to every question in the data set, and then examine heterogeneity in performance across questions. It is, however, impossible to do that

[5] The sample is limited to national security officials who assessed numeric probabilities, given the inherent subjectivity of quantifying the accuracy of verbal estimates of uncertainty.

– as the article describes, we can only calibrate the quality of probabilistic judgments when combining data across multiple questions simultaneously.

A different approach to examining variations in overconfidence across questions in the data set is thus to employ a "bootstrapping" procedure in which we (i) limit the analysis to a small subset of questions, (ii) estimate the degree to which national security officials were overconfident in answering those questions, and (iii) repeat this process a large number of times. If measures of overconfidence remain stable across samples, that would indicate this is a generalizable bias rather than a finding that only pertains to small subset of questions that the study posed.

Replication materials contain code for running a bootstrapping procedure that randomly samples 5 questions from every survey cohort. This is a relatively small fraction of the study's data (each cohort answered 30-40 questions). This procedure quantifies "overconfidence" as the difference between the average probability that national security officials assigned to statements being true and the proportion of those statements that were actually true. This measure requires "inverting" probabilities below 50 percent (i.e. replacing p with $1 - p$) as well as reversing the outcomes of statements that respondents thought were unlikely to be true. In this formulation, a 70 percent probability assigned to a true statement is equivalent to a 30 percent probability assigned to a false statement. Without performing this inversion, overconfidence in the form of overestimating probabilities above 50 percent would cancel out overconfidence in the form of underestimating probabilities below 50 percent.[6]

[6] This procedures follows the method used in many classic calibration studies, such as Baruch Fischhoff, and Paul Slovic, "Calibration of Probabilities" in Daniel Kahneman, Paul Slovic, and Amos Tversky eds., *Judgment Under Uncertainty* (New York: Cambridge University Press, 1982), pp. 294-305.

Repeating this process 10,000 times reveals that national security officials' average overconfidence was 0.18. (This is equivalent to assigning an 88 percent probability to statements that were true 70 percent of the time.) The standard deviation in overconfidence measured across 10,000 trials was just 0.02, the 95 percent confidence interval for national security officials' overconfidence was [0.15, 0.22], and all 10,000 trials resulted in overconfident estimates. These statistics confirm that the overconfidence national security officials displayed in this study was generalizable, as virtually any subset of the data we examine reveals a similar tendency for respondents to attach excessive certainty to their judgments.