Underreporting in Political Science Survey Experiments:

Comparing Questionnaires to Published Results

Online Appendix

This online appendix provides details about the sample and data collection, and provides an example of how the coding procedures were implemented. The dataset was also used in Franco et al (2014) and some of the descriptions are taken from the supplementary materials of the earlier paper. Replication data available at Franco et al. (2015).

1. Overview of the TESS Program

Our analysis leverages TESS (Time-sharing Experiments in the Social Sciences), an NSF-sponsored program, which runs survey-based experiments on representative samples of the U.S. adult population at no cost to researchers. Researchers submit proposals to TESS, which then distributes grants through a competitive peer review process. Accepted studies are administered over the Internet to a panel of survey respondents assembled by GfK Custom Research (formerly known as Knowledge Networks), a high-quality market research firm.

One possible concern is that TESS studies may not be representative of political science research in general, especially research based on observational data. While TESS experiments are clearly not a random sample of all research conducted in political science, it is unlikely that underreporting is less severe than what is described here. Many empirical studies appearing in political science journals are based on analyses of "off-the-shelf" survey datasets (such as the American National Election Study (ANES), the World Values Survey, and the Cooperative Congressional Election Survey (CCES)) or cross-country datasets (such as the Correlates of War and Militarized Interstate Disputes datasets). It is likely to be much easier to find significant relationships and run multiple unreported tests in such datasets because they contain a much

larger number of variables. For instance, the ANES has hundreds of possible variables to analyze and the CCES common content section has had over 50 questions in recent years. In comparison, the TESS studies in our sample asked an average of 15 survey items. The potential analyses to be run from other publicly available datasets are therefore much greater in number than a typical TESS survey experiment. Further, TESS studies may be unrepresentative given that authors are aware that their questionnaires and data will eventually be made public. Again, this should produce less underreporting than we might see in typical empirical research where the complete data are not public.

2. Sample and Data Collection

The first step of the data collection was to determine whether the results from each TESS experiment appeared in a peer-reviewed political science journal. In this study, we define "publication" as an article appearing in a peer-reviewed journal. Accordingly, we do not examine books, book chapters, working papers, conference papers, and dissertations because pre-analysis plans have been primarily suggested for academic journals. Further, the peer review process is considered to be most stringent for journal articles and therefore specification search might be most pronounced for these types of publications.

To determine whether a study was published, we first performed various searches on Google Scholar and ISI Web of Science for: (1) the name of the study (as well as key words from the study title); (2) the authors' names; (3) the words "TESS" or "Time-sharing Experiments in the Social Sciences." We also examined the vitaes of scholars who received TESS grants and reviewed their published papers to see if the TESS experiments had appeared in print. After identifying articles that potentially included the results of each study, we read each one to verify that the results relied on data collected through TESS. We then contacted the

authors of the studies for which we were unable to find any trace online and asked authors to send us any articles that they had published using the data. These contact efforts allowed us to identify additional published papers that escaped our searches, usually because the title had changed from the TESS project or because the paper was forthcoming. Some TESS studies have yielded publications in papers published outside of political science, but we focus on the political science articles in this paper.

3. Example of Coding Procedure

To illustrate the coding procedure, we show how we coded the design features of Malhotra and Popp (2012) based on information from the questionnaire and the published article. Malhotra and Popp (2012) explore how information about the threat of terrorist attacks affects public policy attitudes and how these effects are conditioned by partisanship and fear of terrorism.

Experimental Conditions

Questionnaire. The questionnaire indicates that respondents were assigned to one of four experimental conditions: (1) a control group where respondents received no information about terrorist threat; (2) an experimental condition where respondents read a paragraph where they were told that the threat of a damaging terrorist attack in the United States in the next five years was between 5%-95% (the specific percentage was randomly assigned in multiples of five); (3) an experimental condition where respondents were shown the Department of Homeland Security (DHS) advisory system and told that the threat level in certain sectors of the country was "elevated" (yellow); and (4) an experimental condition where respondents were shown the DHS advisory system and told that the threat level in certain sectors of the country was "severe" (red). Therefore, we coded this study as having **4 experimental conditions.**

Published Article. The published article mentioned experimental groups (1) and (2) but makes no mention of the experimental conditions involving the DHS advisory system (see p. 38 of the paper). Therefore, we concluded that only one of the experimental conditions and the control group was reported. Consequently, we coded this study as reporting **2 experimental conditions**.

Outcome Variables

Questionnaire. Following the presentation of the terrorist threat treatment information, the TESS questionnaire included 11 post-treatment survey items: (1) "How should the United States government inform the American public about the current level of terrorist threat?" (not asked of control group); (2) "In your opinion, how likely is it that the United States will experience a damaging terrorist attack in the next five years?"; (3) "How concerned are you that the United States will suffer a damaging terrorist attack in the next five years?"; (4) "How concerned are you that you or your family will be victims of a damaging terrorist attack in the next five years?"; (5) "If the United States suffers a damaging terrorist attack in the next five years, how many people do you think will die in the attack?"; (6) "Do you support or oppose the U.S. government using wiretaps to listen in on citizens' phone conversations in terrorism investigations?"; (7) "Do you support or oppose a law requiring libraries to turn over to terrorism investigators records of what books people have checked out?"; (8) "Do you support or oppose limits on airline passengers carrying liquids or gels (e.g., beverages, toothpaste, shampoo) onto airplanes?"; (9) "Do you support or oppose the United States launching a military attack against Iran, which American government officials have accused of supporting terrorist organizations such as al Qaeda?"; (10) "Do you support or oppose the proposal to build a fence along the United States-Mexico border to prevent illegal immigrants from entering the U.S.?"; (11) "What

is the current threat level for the U.S. national government?" (asked of control group only). Hence, we coded this study as having **11 outcome variables**.

Published Article. The published article only makes mention of outcome variables (6), (7), (8), and (9) (see p. 38 of the paper). Therefore, we coded the published article as reporting **4 outcome variables**.

Other Items

Questionnaire. Prior to the presentation of the terrorist threat treatment information, the TESS questionnaire included 3 pre-treatment survey items: (1) "People often have to take risks when making financial, career, or other life decisions. Generally speaking, how comfortable are you taking risks?"; (2) "What would you estimate to be the percentage chance that the United States will suffer a damaging terrorist attack in the next five years? Please only type in whole numbers between 0 and 100."; and (3) "How much of the time do you think you can trust the government in Washington to do what is right?" Hence, we coded this study as having **3 other items**.

Published Article. The published article only makes mention of survey item (2) (see p. 38 of the paper). Therefore, we coded the published article as reporting **1 other item**.

4. Results for Other Items

Description. Survey items asked before the presentation of the treatment or otherwise not plausibly affected by the experimental manipulations fall into this category. Some of these items could be used to assess heterogeneous treatment effects (also known as treatment effect moderators). If researchers have *a priori* theoretical expectations that responses to multiple, pre-treatment survey items moderate the treatment effect but only report the heterogeneous treatment effects that emerge as statistically significant, then the probability of making a type I error is

larger than what is reflected by the reported *p*-values. Similarly, these additional items could also be used for covariate adjustment, where the intent is to explain variation in the outcome variables and increase the precision of the estimates. However, specification search is possible in covariate-adjusted models as well. Because we do not know why the researchers chose to include these additional variables, we are much more tentative in our conclusions with respect to underreporting of heterogeneous treatment effects, compared to the more clear cases of unreported experimental conditions or outcome variables.

In enumerating these items, we exclude "profile variables" asked among panelists well before the TESS survey experiment was conducted (such as basic demographics, partisan identification, and political ideology) because TESS provides them to researchers free of charge. As pointed out in the main text, our analysis relies on the assumption that since researchers face a tradeoff between survey length and sample size they only include items that they intend to use in statistical analysis. However, it is quite possible that specification search also occurs with respect to these profile variables. In some cases researchers asked these profile variables as part of their TESS modules (for instance, to obtain a fresher measure of partisanship). Because it was not always clear from the questionnaires if such variables were measured again because researchers requested them or simply because the survey company wanted to refresh these data, we did not classify these variables as "Other Items." Therefore, it is possible that we measure underreporting in this category with error, biasing against us detecting underreporting. As a caveat, if for some reason the authors included pre-treatment survey items without intending to use them for moderation analyses, but then test for moderation *post hoc*, complete reporting may actually not be preferred.

Results. The results for the "other items" category are presented in Table O1 and Figure

O1. On average, questionnaires included 4.9 items that could not be classified as experimental treatments or outcome variables. The published papers reported only 2.1 of these items on average (95% confidence interval around the mean difference of 2.8 is 1.8 to 3.9). Our conclusions with respect to the third category of survey items are more tentative because it is unclear how the researchers intended to use the items. Nonetheless, given that survey time on TESS is a scarce resource, we assume that these items did indeed serve a legitimate research purpose that warranted reporting.

References

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science*. 345(6203): 1502-1505.

Franco, Annie, Neil Malhotra and Gabor Simonovits. 2015. "Replication data for: Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results", http://dx.doi.org/10.7910/DVN/28766 Dataverse [Distributor] V1 [Version], January 21, 2015.

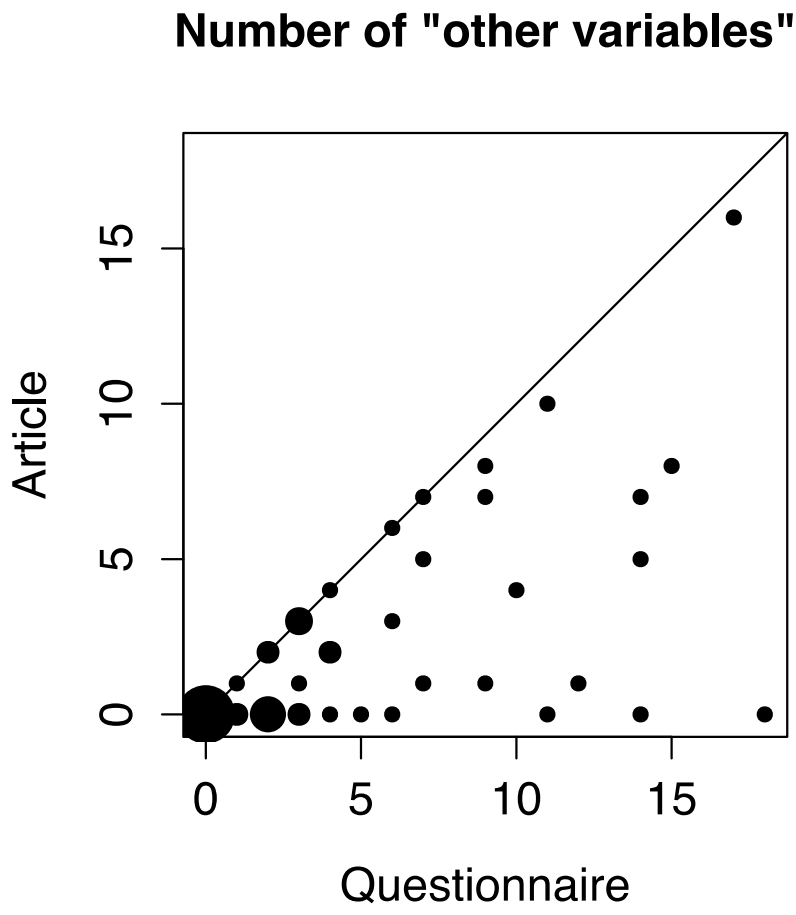
Malhotra, Neil, and Elizabeth Popp. 2012. "Bridging Partisan Divisions over Antiterrorism Policies: The Role of Threat Perceptions." *Political Research Quarterly*. 65(1): 34-47.

Table O1: Underreporting of "Other Items" in Published TESS Studies

	Mean	S.E.	95% C.I.
Q	4.9	0.7	[3.6, 6.2]
A	2.1	0.5	[1.3, 3.1]
Q-A	2.8	0.5	[1.8, 3.9]

Note: *Q* refers to the number of items not classified as experimental conditions or outcome variables included in the survey questionnaire. *A* refers to the number of these variables reported in the published paper. *Q – A* represents the degree of underreporting. Standard errors and confidence intervals calculated by drawing 100,000 bootstrap replications.

Figure O1: Comparing Number of "Other Items" in Questionnaires and Published Results in TESS Studies



Note: Point size is proportional to the number of studies with a particular questionnaire-article value pair.