A Note on Dropping Experimental Subjects who Fail a Manipulation Check

Peter M. Aronow, Jonathon Baron, and Lauren Pinson*

January 9, 2016

Abstract

Dropping subjects after a post-treatment manipulation check is common practice across the social sciences, presumably to restrict estimates to a subpopulation of subjects who understand the experimental prompt. We show that this practice can lead to serious bias and argue for a focus on what is revealed without discarding subjects. Generalizing results developed in Lee (2009) and Zhang and Rubin (2003) to the case of multiple treatments, we provide sharp bounds for potential outcomes among those who would pass a manipulation check regardless of treatment assignment. These bounds may have large or infinite width, implying that this inferential target is often out of reach. As an application, we replicate Press, Sagan and Valentino (2013) with a design that does not drop subjects that failed the manipulation check and show that the findings are likely stronger than originally reported. We conclude with suggestions for practice, namely corrections to the experimental design.

---

*Peter M. Aronow is Assistant Professor, Departments of Political Science and Biostatistics, Yale University, 77 Prospect St., New Haven, CT 06520 (Email: peter.aronow@yale.edu). Jonathon Baron is Doctoral Student, Department of Political Science, Yale University, 115 Prospect St., New Haven, CT 06511. Lauren Pinson is Doctoral Student, Department of Political Science, Yale University, 115 Prospect St., New Haven, CT 06511. Author names are in alphabetical order and do not reflect relative contributions, which the authors consider to be equal. We thank Allan Dafoe, Don Green, Daniel Masterson, Ben Miller, and Betsy Levy Paluck for helpful comments and conversations. Special thanks to Daryl Press, Scott Sagan, and Ben Valentino for generous assistance and materials in replication. We also thank the Yale Institution for Social and Policy Studies Summer Research Lunch group for valuable feedback.

1 Introduction

Manipulation checks are a valuable means of checking the robustness of experimental results in studies based on subjects' attention to treatments, for instance, treatment frames presented in survey experiments. In some studies, researchers may be inclined to exclude those subjects who fail the manipulation check from further analysis. Nominally, the goal of removing subjects is to make sure that we restrict our estimates to a population of subjects who understand the experimental prompt (Wilson, Aronson and Carlsmith, 2010, p. 66). However, this practice may lead to serious bias in estimation, as dropping subjects may induce an asymmetry across treatment arms.

In this note, we show that the practice of dropping subjects based on a manipulation check should generally be avoided. We provide a number of statistical results establishing that doing so can bias estimates or undermine identification of causal effects. We also show that this practice is equivalent to inducing differential attrition across treatment arms, which may induce bias of unknown sign and magnitude.[1] We do not claim that our statistical formulations are particularly novel—they follow from well-known results about conditioning on post-treatment variables and attrition—but, given the prevalence of this practice, we believe that the relationship between these findings and practice in experimentation is under-appreciated.[2]

Our contribution is not solely negative—we provide a number of positive results. First, we reiterate the well-known result that the *intent-to-treat* effect is point identified: if subjects are not discarded, a well-defined causal effect can be estimated consistently. Furthermore, we show that when the result of the manipulation check does not depend on the treatment, an alternative causal quantity—the average treatment effect among those who would pass the manipulation check under all conditions—may be estimated dropping subjects. This condition can be ensured in the design of an experiment, by conditioning solely on checks that are delivered before the experimental treatment is administered. When this condition fails, we provide sharp bounds for the average treatment effect among those who would pass the manipulation check under all conditions. Taken together, our results suggest extreme caution in dropping subjects who fail a post-treatment manipulation check.

In elaborating the potential pitfalls of dropping subjects who fail a manipulation check, we consider Press, Sagan and Valentino (2013)'s (henceforth PSV) survey experiment on public opinion about nuclear weapons. We provide a number of results from an augmented replication of PSV that does not drop subjects that failed the manipulation check. Our findings do not contradict the primary substantive findings of PSV, but instead reinforce their claims: we find that study's exclusion of subjects who failed the manipulation check produced weaker findings than would likely have been returned by a full sample. We then conclude with recommendations for applied practice,

---

[1]The point has been made before, but has not to our knowledge been formalized. E.g., Gerber and Green (2012, 212) notes that attrition may be induced when "[r]esearchers deliberately discard observations. Perhaps ill-advisedly, laboratory researchers sometimes exclude from their analysis subjects who seem not to understand the instructions or who fail to take the experimental situation seriously" but do not provide further discussion of this point.

[2]For recent examples in political science, see Hoffman et al. (2013), Maoz (2006), Turner (2007), Crawford et al. (2013), De Oliveira, Guimond and Dambrun (2012), and Small, Lerner and Fischhoff (2006)

namely a focus on what is revealed without discarding subjects.

2 Identification of causal effects

Suppose we have an i.i.d. sample from (Y,S,Z), where Y denotes the subject's response, S denotes the result of a manipulation check (1 if the subject passed, 0 if the subject failed), and Z denotes the subject's treatment assignment (1 for treatment 1, 2 for treatment 2, ...). Without loss of generality, assume that the support of Z is {1,...,K}, where K is finite.

We make three assumptions to proceed. First, we assume that both potential responses and potential results from the manipulation check are stable, by invoking SUTVA (Rubin, 1980), which implies both no interference between units and no multiple unobserved versions of the treatment.

Assumption 1 (SUTVA)**.** $Y = \sum_{z=1}^{K} Y(z)\mathbf{I}(Z=z)$ *and* $S = \sum_{z=1}^{K} S(z)\mathbf{I}(Z=z)$.

Second, we assume that the treatment is not systematically related to potential outcomes or potential manipulation check results, as would be ensured by random assignment of the treatment.

Assumption 2 (Ignorability)**.** *For all* $z \in \{1,...,K\}$, $(Y(z),S(z)) \perp\!\!\!\perp Z$, *with* $\Pr(Z=z) > 0$.

Assumption 2 can be ensured at the design stage by randomizing treatment assignment across subjects.

Finally, we require that at least some subjects pass the manipulation check in both treatment and control.

Assumption 3 (Non-zero passing rates)**.** *For all* $z \in \{1,...,K\}$, $\Pr[S|Z=z] > 0$.

Note that Assumption 3 is verifiable from the experimental data.

Without discarding subjects, all mean potential outcomes are well-identified, and their differences are also point identified. These differences are sometimes referred to as intent-to-treat effects (Gerber and Green, 2012). As we proceed, plug-in estimators will be consistent given suitable regularity conditions (e.g., finite third moments), with the bootstrap providing a basis for asymptotic inference.

Lemma 1. $\mathrm{E}[Y|Z=z] - \mathrm{E}[Y|Z=z'] = \mathrm{E}[Y(z) - Y(z')]$.

Randomization ensures intent-to-treat effects are point identified, and can be estimated simply by examining differences in means.

In order to assess the operating characteristics of dropping subjects, we must formalize the presumed inferential target of a researcher who chooses to drop subjects based on a manipulation check. Here we consider one possible target that seems natural: $\mathrm{E}[Y(z) - Y(z')|S(1) = S(2) = \ldots = S(K) = 1]$, or the average treatment effect among subjects who would pass under all treatment conditions. There is a condition under which dropping subjects who fail a manipulation check recovers this quantity, namely that in which treatment is assigned to a subject is entirely unrelated to whether or not that subject passes or fails the manipulation.

Corollary 1. *If* $\Pr[S(1) = S(2) = \ldots = S(K)] = 1$*, then* $\mathrm{E}[Y|S = 1, Z = z] - \mathrm{E}[Y|S = 1, Z = z'] = \mathrm{E}[Y(z) - Y(z')]|S(1) = S(2) = \ldots = S(K) = 1]$.

Corollary 1 implies sufficient conditions for discarding subjects to be unproblematic if the inferential target is the average potential outcomes among those who pass the manipulation check. In short, the treatment cannot affect whether or not a subject passes the manipulation check (e.g. if the treatment impacts subjects' ability to pass the manipulation check itself, for instance by inducing variable degrees of stress, or even if subjects receive treatments for different lengths of time). Thus we can find cases where dropping subjects is acceptable: if a manipulation check precedes treatment (i.e., a pre-treatment attention check), then discarding subjects is not a problem, at least for characterizing effects for a well-defined subpopulation of units.

In order to justify discarding subjects, it is not, however, sufficient to have show there is no average effect of Z on S.

Corollary 2. $\mathrm{E}[S|Z = 1] = \mathrm{E}[S|Z = 2] = \ldots = \mathrm{E}[S|Z = K]$ *does not imply that* $\mathrm{E}[Y|S = 1, Z = z] - \mathrm{E}[Y|S = 1, Z = z'] = \mathrm{E}[Y(z) - Y(z')|S(1) = S(2) = \ldots = S(K) = 1]$.

Corollary 2 reinforces an important point: even if the failure rates are identical across treatments, conditioning on a manipulation check may still induce bias. This is because the types of subjects who fail the manipulation check under one treatment may not be the same as those who fail under a different treatment.

To this end, potential outcomes among those who would pass regardless of condition are not generally point identified. In a generalization of Lee (2009) (which impose a monotonicity assumption) and Zhang and Rubin (2003), we derive sharp bounds on potential outcome means $\mathrm{E}[Y(z)|S(1) = 1, S(2) = 1, \ldots, S(K) = 1]$.[3]

Proposition 1. *Suppose that* $\Pr[S(1) = S(2) = \ldots = S(K) = 1] > 0$ *and that* Y *is continuous with unbounded support. Let* $Q_{Y|Z=z,S=1}(.)$ *denote the conditional quantile function of* Y *given* $Z = z$ *and* $S = 1$*. Then, sharp bounds for* $\mathrm{E}[Y(z)|S(1) = S(2) = \ldots = S(K) = 1]$ *are given by*

$$\mathrm{E}\left[Y|Y \leq Q_{Y|Z=z,S=1}\left(1 - \sum_{z' \in \{1,\ldots,K\}: z' \neq z} \frac{\Pr[S = 0|Z = z']}{\Pr[S = 1|Z = z]}\right), Z = z\right]$$
$$\leq \mathrm{E}[Y(z)|S(1) = S(2) = \ldots = S(K) = 1] \leq$$
$$\mathrm{E}\left[Y|Y \geq Q_{Y|Z=z,S=1}\left(\sum_{z' \in \{1,\ldots,K\}: z' \neq z} \frac{\Pr[S = 0|Z = z']}{\Pr[S = 1|Z = z]}\right), Z = z\right]$$

when $\sum_{z' \in \{1,\ldots,K\}: z' \neq z} \frac{\Pr[S=0|Z=z']}{\Pr[S=1|Z=z]} < 1$*, else these bounds are given by* $-\infty \leq \mathrm{E}[Y(z)|S(1) = S(2) = \ldots = S(K) = 1] \leq \infty$.

Assuming that $\Pr[S(0) = 1, S(1) = 1] > 0$ ensures that $\mathrm{E}[Y(z)|S(0) = 1, S(1) = 1]$ exists and continuity of Y ensures that the quantile function is well-defined. As with Lee (2009)'s bounds,

[3]We thank Ben Miller for helpful discussions regarding the formulation of Proposition 2.

even when Y is discrete, then bounds can be constructed simply by trimming the observations associated with the upper or lower $\sum_{z' \in \{1,\ldots,K\}: z' \neq z} \frac{\Pr[S=0|Z=z']}{\Pr[S=1|Z=z]}$th proportions of the empirical distributions of subjects who pass the manipulation check under treatment and control. With weighted data, this entails using the weighted empirical distribution function.

In the Appendix, we report a set of simulations to evaluate the properties of the difference-in-means estimator after dropping subjects and of the proposed bounds. In short, we show, all else equal, that bias tends to increase as the average potential outcomes of subjects who would pass the control manipulation check diverge from those who would pass the treatment manipulation check and as failure rates increase. We further show that the width of the bounds increases as failure rates increase, but also as the variance of potential responses increases.

Taken together, our results establish the following: (i) Intent-to-treat effects are point identified. (ii) Potential outcomes among those who would pass a manipulation check under all conditions are not generally point identified. (iii) Sharp bounds for potential outcomes among those who would pass a manipulation check under all conditions may not have finite width. (iv) Showing that equal proportions of subjects failed the manipulation check across all conditions is not sufficient to justify dropping subjects, because the types of subjects that comprise those groups may differ between treatments. (v) If the manipulation check precedes (or its result is otherwise unrelated to) treatment assignment, dropping subjects who fail a manipulation check does not lead to bias in estimation of outcomes for those who would pass the manipulation check under all treatment conditions.

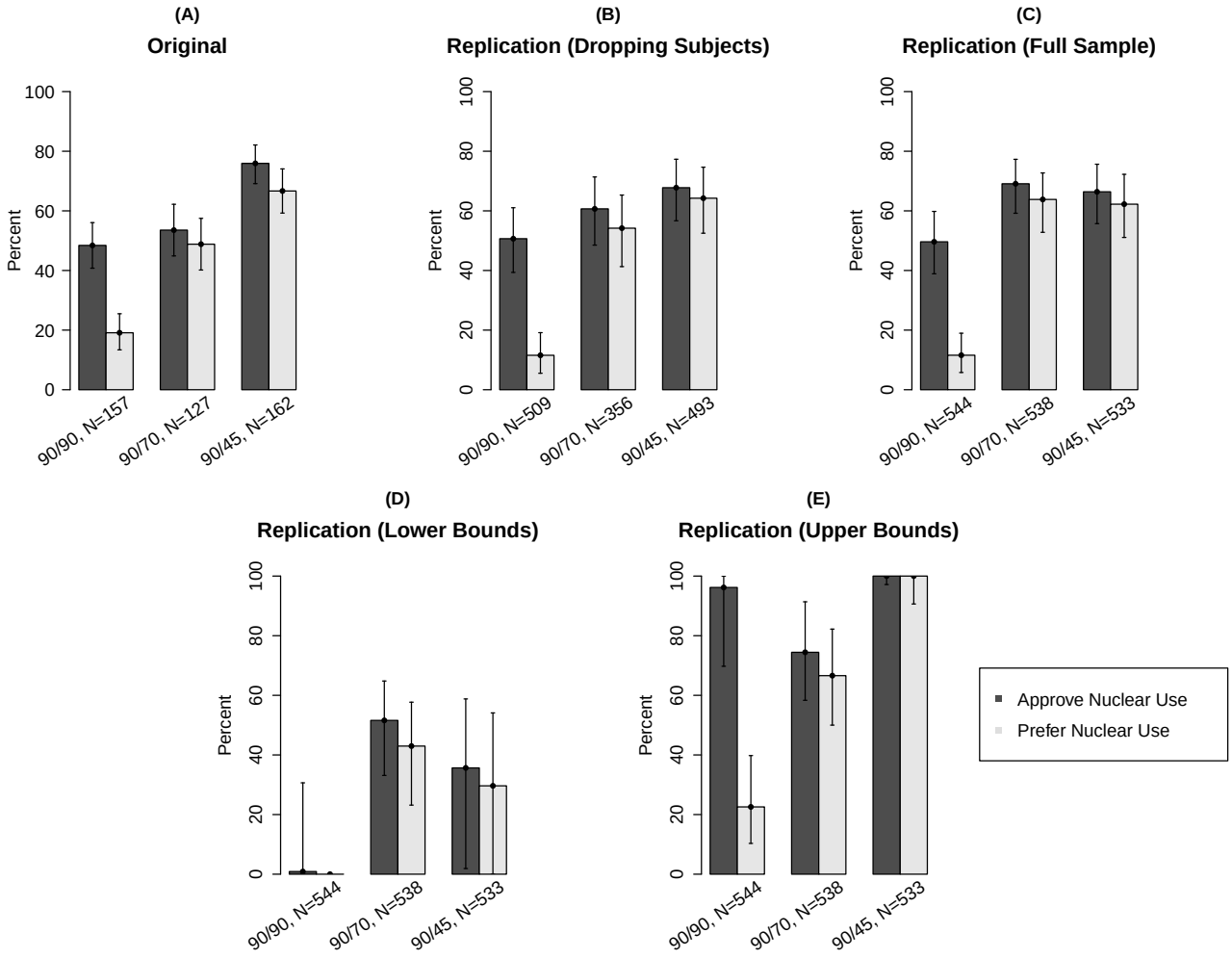
3 Application

We now discuss these findings in the context of PSV, which presents a survey conducted using a representative sample of voting-age American citizens. PSV reports on whether public opinion regarding nuclear-weapons use is shaped primarily by ethical or strategic considerations. PSV's five treatment conditions detail a scenario in which an Al Qaeda nuclear-weapons lab in Syria obtains weapons-grade uranium intended for offensive use against the United States. We focus on the first three treatments, which manipulate the relative expected effectiveness of a nuclear or conventional strike on the Al Qaeda facility. The treatments describe the effectiveness ratios of nuclear/conventional weapons at 90-percent/90-percent, 90-percent/70-percent, and 90-percent/45-percent, respectively (these conditions are henceforth referred to as 90-90, 90-70, and 90-45). As demonstrated in Panel A of Figure 1 above, PSV reports a strongly monotonic increase in subjects' approval of and preference for nuclear-weapons use as nuclear weapons' effectiveness increases relative to that of conventional weapons.

However, we will show that this finding—namely a strong monotonic increase—is partly attributable to dropping subjects, and that, without dropping subjects, even stronger results are obtained. PSV utilized a post-treatment manipulation check as a means of gauging subjects' attention to the treatment articles; subjects who fail the manipulation check are dropped from the analysis. This check asked subjects to choose from five options whether the treatment they had received had concluded that nuclear weapons would be equally effective, moderately effective, or much more effective than conventional weapons; these answers were intended to correspond to the 90-90, 90-

70, and 90-45 comparisons, respectively. Subjects who failed the manipulation check were given an opportunity to read the article a second time, but their responses were dropped due to failure.

Figure 1: Results from Press, Sagan, and Valentino (2013) and Replication



Comparisons of original and weighted replication data. Panel A presents results from PSV with subjects dropped; Panel B presents results from the replication with subjects dropped; Panel C presents results from the replication using the full sample; Panel D presents results imputing the lower bounds for all treatment conditions; Panel E presents results imputing the upper bounds for all treatment conditions. Vertical bars represent 95% confidence intervals on point estimates calculated using the bootstrap.

We conducted a successful replication of PSV, showing that the primary findings of the study are robust, including robustness to treatment variations not considered in the original article. (Details are provided in Appendix B.) One notable variation in our replication was that subjects were not informed of whether they had passed or failed the manipulation check, and data were collected

regardless, allowing us to assess the consequences dropping subjects. Our study used 2,733 subjects recruited from Amazon.com Mechanical Turk.[4] Note that there is a good body of evidence that subjects on Mechanical Turk tend to be more attentive than other (representative) samples (Hauser and Schwarz, 2015), and thus it is possible that our passing rates are higher than those of PSV. If so, then all else equal, the bias in our estimates after dropping subjects would be smaller, and the width of our bounds would also be smaller than the bounds that would be associated with the original data.

Subjects were compensated at $.50 each, with the added chance of winning a $100 bonus if they passed the manipulation check. We used inverse probability weighting to adjust the replication sample to match the covariate distribution of the sample used in PSV, and computed estimates after weighting.[5] We then performed our analysis both including and excluding subjects who failed the manipulation check.

The results of our replication are presented in Panels B and C of Figure 1. PSV argues that a clear majority of subjects both approve of and prefer a prospective nuclear strike in only 90-45. However, we show that these findings are actually attenuated as a consequence of dropping the subjects who failed from the analysis. Panel C demonstrates that including data from all subjects, regardless of performance on the manipulation check, alters results substantially, rendering estimates in 90-70 and 90-45 practically indistinguishable. The discrepancy is notable in its substantive importance for the results of PSV: subjects dropped from the analysis actually perceived a "moderate" decline in conventional weapons' relative effectiveness to be a "significant" decrease; including such subjects in the analysis provides even stronger evidence against a nuclear-nonuse norm than PSV depicts.

Importantly, the replication also illustrates the importance of Corollary 2—showing the equivalence between failure rates under two conditions does not imply that the types of individuals who fail are equivalent. Treatments 1 and 3 have statistically and substantively indistinguishable failure rates (6.4% vs. 7.5%), and yet the covariate profiles of those who failed the manipulation check are strikingly different, as can be seen in Table 1 in Appendix C. For example, 73% of subjects who fail the manipulation check under Treatment 1 are male, but only 30% of the subjects who fail the manipulation check under Treatment 3 are male. Similar differences are found in political party, religion, region, and racial composition.

We also report sharp bounds for the average potential outcomes among subjects who would pass the manipulation check regardless of treatment assignment. We are unable to calculate analogous bounds using the PSV data, because no information was available on the proportion of

[4]Three subjects were omitted from analysis because of technical difficulties that prevented us from verifying that they completed the survey; 2,730 subjects are included in the analysis below.

[5]Let $R_i = 1$ if an observation is in the replication study, else let $R_i = 0$ if an observation was in the original data. We performed a logistic regression of R_i on the following covariates $\mathbf{X}_i$: Education, Party, Religion, Political Interest, Income, Gender, News Interest, Voter Registration, Birth Year, Region, Race, and Ideology (with mean-imputation for missingness). Using the output of this logistic regression, we computed a predicted value $p_i = \Pr[R_i = 1|\mathbf{X}_i]$ for each observation i. To reweight the replication study to the original study's covariate profile, we weighted each observation in the replication sample by $\frac{p_i}{1-p_i}$.

subjects who were dropped. Panel D depicts our results imputing the lower bound, whereas Panel E shows our estimates imputing the upper bound. We observe that the bounds for all treatment conditions have overlapping regions, in both outcome variables, making it impossible to fully differentiate results across the treatments. The bounds are largely uninformative, suggesting the fundamental uncertainty induced in attempting to estimate effects among the subpopulation of subjects who would always pass the manipulation check. If these effects are indeed the inferential target of the researcher, very little is revealed by the experimental data, and dropping subjects may introduce serious bias of unknown sign and magnitude (as any values within the range of the bounds are compatible with the experimental data).

4 Conclusion

We reiterate that our critiques do not apply to research designs that use pre-treatment attention checks to screen subjects, as established by Corollary 2. Attention checks placed before treatment can be used to prune subjects from final analysis in a principled manner. Although screening changes the inferential target from the whole population to a subpopulation of subjects who are paying sufficient attention and have the ability to pass the attention check, it does not compromise internal validity. The use of manipulation checks in the pilot stage can provide information for improving the interpretability of treatments and manipulation checks to maximize passing rates in the manipulation. The use of pre-treatment manipulation checks and piloting may also improve estimates by focusing only on the subpopulation of subjects who are able to pass and do so. Of course, there is significant debate about best practices here. In this vein, we recommend Berinsky, Margolis and Sances (2014)'s analysis of the benefits and drawbacks of screening, as well as their recommendations for practice. Berinsky, Margolis and Sances (2014)'s suggested use of screens to assess subjects' attention on a continuum represents a transparent approach to presenting findings by showing the results conditional on each value of attention. However, doing so with a post-treatment manipulation check would continue to introduce the issues discussed here.

In general, we stress the importance of manipulations that are sufficiently clear so as to minimize the necessity to remove subjects based on a lack of comprehension. Although we have proposed bounds for causal effects among the subjects who would always pass the manipulation check, these bounds will be uninformative in practice when failure rates are high. Pilot studies may help to ensure that the treatments are understood by subjects as the researchers intended, with the caveat that the pilot population may be unrepresentative of the final test population. We underline that best practice should maintain a focus on intent-to-treat effects, which are generally point identified and have a clear substantive interpretation. The credibility of the experiment ultimately rests on the quality of the manipulation, rather than posthoc statistical adjustments. As a means of validation, manipulation checks can help researchers to understand whether or not this criterion has been met. But dropping subjects who fail a manipulation check presented after the intervention may introduce biases of unknown sign and magnitude.

A Proofs

Proof of Lemma 1. The result follows from linearity of expectations. □

Proof of Corollary 1. $\Pr[S(1) = S(2) = \ldots = S(K)] = 1$ implies $S = S(1) = S(2) = \ldots = S(K)$, thus ensuring $(Y(1), Y(2), ..., Y(K), S) \perp\!\!\!\perp Z$. Joint independence implies that $\mathrm{E}[Y(z)|S = 1, Z = z] - \mathrm{E}[Y(z')|S = 1, Z = z'] = \mathrm{E}[Y(z)|S(1) = 1, S(2) = 1, \ldots, S(K) = 1] - \mathrm{E}[Y(z')|S = 1]$. □

Proof of Corollary 2. We prove the claim via a simple counterexample. Suppose $\mathrm{Supp}(Z) = \{1, 2\}$ and

$$f(Y(1), Y(2), S(1), S(2)) = \begin{cases} 1/3 & : Y(1) = 1, Y(2) = 1, S(1) = 0, S(2) = 1 \\ 1/3 & : Y(1) = 0, Y(2) = 0, S(1) = 1, S(2) = 0 \\ 1/3 & : Y(1) = 0, Y(2) = 0, S(1) = 1, S(2) = 1 \\ 0 & : otherwise. \end{cases}$$

Note that $\mathrm{E}[S(2) - S(1)] = 0$ and $\mathrm{E}[\tau|S(1) = S(2) = 1] = \mathrm{E}[\tau|S(1) = 1, S(2) = 0] = \mathrm{E}[\tau|S(1) = 0, S(2) = 1] = 0$. $\mathrm{E}[Y|S = 1, Z = 2] - \mathrm{E}[Y|S = 1, Z = 1] = 0 - 1/2 = -1/2$. □

Proof of Proposition 1. We will follow the general logic of Lee (2009), and technical details carry through from the proof of Lee's Proposition 1a. Without loss of generality, we consider the upper bound for $\mathrm{E}[Y(1)|S(1) = S(2) = \ldots = S(K) = 1]$.

Define $U = 1$ if $S(2) = \ldots = S(K) = 1$, else let $U = 0$. Then $\mathrm{E}[Y(1)|S(1) = S(2) = \ldots = S(K) = 1] = \mathrm{E}[Y(1)|U = 1, S(1) = 1]$. We do not observe the joint distribution of $(Y(1), U)|S(1) = 1$, as we never jointly observe $Y(z)$ and $S(z')$, for $z \neq z'$. Let $p_{U^0} = \Pr[U = 0|S(1) = 1]$. Given continuity of $Y(1)$, then among all possible joint distributions $(Y(1), U)|S(1) = 1$, $\mathrm{E}[Y(1)|U = 1, S(1) = 1]$ is maximized when $U = 1$ for all $Y(1) \geq Q_{Y(1)}(p_{U^0})$. By weak monotonicity of the quantile function, it suffices to maximize p_{U^0} to find a maximum for $\mathrm{E}[Y(1)|U = 1, S(1) = 1]$.

We again do not observe the joint distribution $(U, S(1))$. By σ-additivity, a sharp upper bound is obtained for $\Pr[U = 0]$ is obtained when the regions where $S(2), S(3), ..., S(K)$ each equal zero are disjoint, with $\Pr[U = 0] \leq \begin{cases} \sum_{k=2}^{K} \Pr[S(k) = 0] & : \sum_{k=2}^{K} \Pr[S(k) = 0] < 1 \\ 1 & : \sum_{k=2}^{K} \Pr[S(k) = 0] \geq 1 \end{cases}$. Thus, among all possible joint distributions $(U, S(1))$, $\Pr[U = 0|S(1) = 1] = p_{U^0}$ is maximized when $p_{U^0} = \begin{cases} \frac{\sum_{k=2}^{K} \Pr[S(k)=0]}{\Pr[S(1)=1]} & : \frac{\sum_{k=2}^{K} \Pr[S(k)=0]}{\Pr[S(1)=1]} < 1 \\ 1 & : \frac{\sum_{k=2}^{K} \Pr[S(k)=0]}{\Pr[S(1)=1]} \geq 1 \end{cases}$. Thus if $\frac{\sum_{k=2}^{K} \Pr[S(k)=0]}{\Pr[S(1)=1]} < 1$, a sharp upper bound is given by $\mathrm{E}[Y(1)|U = 1, S(1) = 1] \leq \mathrm{E}\left[Y(1)|Y(1) \leq Q_{Y(1)|S(1)=1}\left(1 - \sum_{k=2}^{K} \frac{\Pr[S(k)=0]}{\Pr[S(1)=1]}\right)\right]$, else the upper bound is infinite.

By random assignment and SUTVA, the conditional distribution of $Y(1)|S(1) = 1$ is equivalent to the conditional distribution of $Y|S = 1, Z = 1$, and the marginal distributions of $S(k)$ are each equivalent to $S|Z = k$. Thus a sharp upper bound is given by $\mathrm{E}[Y(1)|U = 1, S(1) = 1] \leq \mathrm{E}\left[Y|Y \geq Q_{Y|Z=1,S=1}\left(\sum_{k=2}^{K} \frac{\Pr[S=0|Z=k]}{\Pr[S=1|Z=k]}\right), Z = 1\right]$ when $\sum_{k=2}^{K} \frac{\Pr[S=0|Z=k]}{\Pr[S=1|Z=k]} < 1$, else the upper bound is infinite. The bounds are invariant to indexing of treatments Z, thus yielding the general upper bound in Proposition 1. Analogous calculations yield lower bounds.

□

B Details of Replication of Press, Sagan, and Valentino (2013)

Our replication and pre-analysis plan are hosted at EGAP (ID: 20150131AA). Our replication included three major variations, the analysis of which underlines the robustness of PSV. We list these analyses in turn below.

First, because the original experiment was performed prior to the onset of the Syrian civil war, we sought to assess whether the results were invariant to shifts in time and context (i.e. whether the results might differ in our replication, given the political changes that have occurred in Syria). We thus randomized whether treatment frames presented the scenario in Syria or Lebanon, which was used as an analog to pre-civil-war Syria; treatments were assigned through a 2×5 factorial design. We found no statistically or substantively significant difference between Syria and Lebanon treatment frames, demonstrating that the results presented in PSV are robust to these temporal and contextual changes.

Second, we analyzed whether the PSV study's use of post-treatment covariates introduced bias.[6] We added another treatment (rendering our augmented replication a $2 \times 2 \times 5$ factorial design) that randomized whether subjects answered these questions before or after treatment. This analysis failed to reveal any statistically or substantively significant results.

Third, as noted above, we performed weighting on our survey sample to approximate the experimental population used by PSV. Our subjects were recruited from Mechanical Turk, and likely constituted an unrepresentative sample. As noted in the main text, we used logistic regression and inverse probability weighting (IPW) to assign treatment probabilities and corresponding weights for each subject. We did observe differences between the weighted and unweighted analyses, but neither undermined the substantive findings of PSV.

[6] See, e.g., Angrist and Pischke (2008).

C Simulations

We assume a binary treatment Z with $\Pr(Z=1)=1/2$. We generated potential outcomes $Y(0)=Y(1)=\lambda[S(1)-S(0)]+N(0,\sigma)$, and vary λ, σ, and the joint distribution of $(S(0),S(1))$. Note that in the simulation, we have assumed that there is no effect of the treatment whatsoever, and the results would be invariant to the introduction of any constant treatment effect. λ represents the divergence in potential outcomes between those who would pass and those who would fail the manipulation check and σ represents the unexplained variability of potential outcomes. To put our results in asymptopia, we assume $N=1000$, and perform $100,000$ simulations.

Table 1 presents the results of our simulations. We first discuss the bias of the difference-in-means estimator after dropping subjects. We show that bias tends to increase as λ – the divergence between the average potential outcomes of subjects who would pass the control manipulation check and that of those who would pass the treatment manipulation check – increases. See, e.g., row 1 vs. 2. As failure rates increase, not necessarily differentially across treatment arms, we also see that bias increases; compare rows 1-4 to 5-8 to 9-12. Furthermore, as $\rho(S(0),S(1))$ – the correlation between potential responses to the manipulation check – decreases, bias also increases, as evidenced by, e.g., row 4 vs. row 1.

The width of the bounds also depends on multiple factors. As the variability of potential outcomes increases (characterized by σ, and to a lesser extent λ), the width of the bounds increases, as evidenced by comparing, e.g., row 1 vs. 2 vs. 3. The width of the bounds also depends on failure rates; again compare rows 1-4 to 5-8 to 9-12. The bounds do not depend on any unobservable features of the joint distributions of potential outcomes and responses to the manipulation check. To wit, the width of the bounds does not change as $\rho(S(0),S(1))$ is varied; compare, e.g., row 1 to row 4.

Table 1: Simulations demonstrating the effects of dropping.

	λ	$\Pr[S(0)=1]$	$\Pr[S(1)=1]$	$\rho(S(0),S(1))$	σ	Bias	Bound Width
1	1	0.8	0.8	0.25	1	0.301	1.795
2	10	0.8	0.8	0.25	1	3.005	5.229
3	1	0.8	0.8	0.25	10	0.301	16.921
4	1	0.8	0.8	0.5	1	0.201	1.764
5	1	0.6	0.8	0.25	1	0.456	2.909
6	10	0.6	0.8	0.25	1	4.559	10.285
7	1	0.6	0.8	0.25	10	0.457	26.867
8	1	0.6	0.8	0.5	1	0.315	2.856
9	1	0.6	0.6	0.25	1	0.598	4.815
10	10	0.6	0.6	0.25	1	5.973	19.867
11	1	0.6	0.6	0.25	10	0.601	43.742
12	1	0.6	0.6	0.5	1	0.398	4.708

Simulations performed with $N=1,000$ and $100,000$ simulations; bound widths are presented as averages over all simulations.

D Additional Summary Statistics

Below, we present distributions of the reweighted covariate profiles of subjects in our replication study, disaggregated by treatment condition and performance on the manipulation checks.

Table 2: Covariate distributions among subjects who failed the manipulation check.

	Treatment 1		**Treatment 2**		**Treatment 3**	
	Mean	(SE)	Mean	(SE)	Mean	(SE)
Education	2.969	(0.184)	3.179	(0.274)	2.759	(0.224)
Democrat	0.325	(0.138)	0.567	(0.112)	0.339	(0.148)
Republican	0.370	(0.188)	0.182	(0.075)	0.158	(0.081)
Independent	0.221	(0.112)	0.234	(0.091)	0.180	(0.115)
Other Party	0.011	(0.009)	0.002	(0.001)	0.007	(0.005)
Religion	4.563	(0.523)	4.708	(0.326)	3.616	(0.500)
Political Interest	3.477	(0.254)	3.793	(0.152)	2.850	(0.314)
Income	7.737	(0.793)	5.923	(0.572)	7.467	(0.997)
Gender	0.729	(0.125)	0.296	(0.088)	0.248	(0.103)
News Interest	2.109	(0.316)	1.598	(0.143)	2.657	(0.369)
Registered to Vote	0.771	(0.117)	0.923	(0.051)	0.846	(0.115)
Unregistered to Vote	0.229	(0.117)	0.076	(0.050)	0.153	(0.115)
Don't Know if Registered	0.000	(0.000)	0.001	(0.001)	0.001	(0.001)
Birth Year	1973.392	(5.358)	1956.628	(4.060)	1971.237	(3.667)
Northeast	0.038	(0.030)	0.100	(0.039)	0.181	(0.122)
Midwest	0.025	(0.017)	0.450	(0.125)	0.167	(0.116)
South	0.382	(0.147)	0.313	(0.098)	0.612	(0.147)
West	0.555	(0.158)	0.137	(0.047)	0.040	(0.025)
White	0.623	(0.147)	0.898	(0.034)	0.796	(0.119)
Black	0.300	(0.134)	0.035	(0.015)	0.012	(0.011)
Latino/Hispanic	0.065	(0.057)	0.023	(0.015)	0.168	(0.118)
Asian	0.012	(0.012)	0.009	(0.004)	0.019	(0.015)
Native American	0.000	(0.000)	0.008	(0.007)	0.000	(0.000)
Middle Eastern	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)
Mixed Race	0.000	(0.000)	0.026	(0.017)	0.004	(0.005)
Other Race	0.000	(0.000)	0.000	(0.000)	0.001	(0.001)
Ideology	2.778	(0.199)	3.319	(0.265)	3.215	(0.422)
Observatons	$N = 35$		$N = 182$		$N = 40$	

Table 3: Covariate distributions among subjects who passed the manipulation check.

	Treatment 1		**Treatment 2**		**Treatment 3**	
	Mean	(SE)	Mean	(SE)	Mean	(SE)
Education	3.624	(0.149)	3.554	(0.153)	3.542	(0.159)
Democrat	0.534	(0.056)	0.407	(0.061)	0.423	(0.061)
Republican	0.173	(0.037)	0.239	(0.056)	0.237	(0.052)
Independent	0.260	(0.041)	0.327	(0.068)	0.306	(0.057)
Other Party	0.019	(0.014)	0.013	(0.007)	0.004	(0.002)
Religion	3.534	(0.187)	3.629	(0.220)	4.129	(0.242)
Political Interest	3.696	(0.087)	3.743	(0.106)	3.726	(0.120)
Income	8.006	(0.340)	7.443	(0.445)	7.910	(0.370)
Gender	0.386	(0.051)	0.322	(0.054)	0.385	(0.056)
News Interest	1.734	(0.081)	1.794	(0.083)	1.752	(0.117)
Registered to Vote	0.935	(0.020)	0.914	(0.025)	0.923	(0.021)
Unregistered to Vote	0.064	(0.020)	0.083	(0.025)	0.070	(0.020)
Don't Know if Registered	0.001	(0.001)	0.002	(0.002)	0.007	(0.005)
Birth Year	1962.94	(1.750)	1960.419	(1.902)	1961.317	(1.545)
Northeast	0.188	(0.045)	0.199	(0.055)	0.119	(0.027)
Midwest	0.249	(0.057)	0.230	(0.055)	0.179	(0.040)
South	0.316	(0.049)	0.275	(0.053)	0.509	(0.061)
West	0.248	(0.046)	0.296	(0.061)	0.193	(0.045)
White	0.786	(0.038)	0.821	(0.048)	0.736	(0.063)
Black	0.079	(0.022)	0.072	(0.034)	0.143	(0.055)
Latino/Hispanic	0.052	(0.016)	0.075	(0.036)	0.080	(0.043)
Asian	0.028	(0.007)	0.006	(0.002)	0.013	(0.005)
Native American	0.002	(0.001)	0.001	(0.001)	0.006	(0.003)
Middle Eastern	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)
Mixed Race	0.014	(0.006)	0.017	(0.007)	0.018	(0.011)
Other Race	0.039	(0.024)	0.008	(0.007)	0.005	(0.003)
Ideology	2.908	(0.174)	3.146	(0.183)	3.208	(0.159)
Observations	$N = 509$		$N = 356$		$N = 493$	

Table 4: Covariate distributions for all subjects.

	Treatment 1		**Treatment 2**		**Treatment 3**	
	Mean	(SE)	Mean	(SE)	Mean	(SE)
Education	3.592	(0.142)	3.403	(0.152)	3.493	(0.149)
Democrat	0.524	(0.053)	0.472	(0.061)	0.417	(0.057)
Republican	0.183	(0.037)	0.216	(0.045)	0.232	(0.049)
Independent	0.258	(0.039)	0.289	(0.055)	0.298	(0.054)
Other Party	0.018	(0.013)	0.008	(0.005)	0.004	(0.002)
Religion	3.584	(0.184)	4.064	(0.218)	4.097	(0.227)
Political Interest	3.685	(0.083)	3.763	(0.087)	3.671	(0.117)
Income	7.994	(0.330)	6.844	(0.386)	7.883	(0.347)
Gender	0.402	(0.049)	0.312	(0.048)	0.376	(0.052)
News Interest	1.753	(0.079)	1.715	(0.08)	1.806	(0.114)
Registered to Vote	0.927	(0.020)	0.918	(0.024)	0.918	(0.022)
Unregistered to Vote	0.072	(0.020)	0.080	(0.024)	0.075	(0.021)
Don't Know if Registered	0.001	(0.001)	0.002	(0.001)	0.007	(0.004)
Birth Year	1963.445	(1.678)	1958.889	(2.087)	1961.942	(1.506)
Northeast	0.181	(0.042)	0.159	(0.037)	0.123	(0.027)
Midwest	0.238	(0.055)	0.319	(0.067)	0.179	(0.038)
South	0.319	(0.048)	0.290	(0.051)	0.516	(0.060)
West	0.262	(0.044)	0.232	(0.045)	0.183	(0.042)
White	0.778	(0.036)	0.852	(0.033)	0.739	(0.059)
Black	0.090	(0.022)	0.057	(0.022)	0.135	(0.052)
Latino/Hispanic	0.053	(0.016)	0.054	(0.023)	0.085	(0.041)
Asian	0.027	(0.006)	0.008	(0.002)	0.014	(0.005)
Native American	0.002	(0.001)	0.004	(0.003)	0.006	(0.003)
Middle Eastern	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)
Mixed Race	0.013	(0.005)	0.021	(0.008)	0.017	(0.010)
Other Race	0.037	(0.023)	0.005	(0.004)	0.004	(0.003)
Ideology	2.904	(0.166)	3.223	(0.150)	3.209	(0.150)
Observations	$N = 544$		$N = 538$		$N = 533$	

References

Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

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.

Crawford, Jarret T, Jennifer L Brady, Jane M Pilanski and Heather Erny. 2013. "Differential effects of right-wing authoritarianism and social dominance orientation on political candidate support: The moderating role of message framing." *Journal of Social and Political Psychology* 1(1):5–28.

De Oliveira, Pierre, Serge Guimond and Michael Dambrun. 2012. "Power and Legitimizing Ideologies in Hierarchy-Enhancing vs. Hierarchy-Attenuating Environments." *Political Psychology* 33(6):867–885.

Gerber, Alan S and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation*. WW Norton.

Hauser, David J and Norbert Schwarz. 2015. "Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants." *Behavior research methods* pp. 1–8.

Hoffman, Aaron M., Christopher R. Agnew, Laura E. VanderDrift and Robert Kulzick. 2013. "Norms, Diplomatic Alternatives, and the Social Psychology of War Support." *Journal of Conflict Resolution* .

Lee, David S. 2009. "Training, wages, and sample selection: Estimating sharp bounds on treatment effects." *The Review of Economic Studies* 76(3):1071–1102.

Maoz, Ifat. 2006. "The Effect of News Coverage Concerning the Opponents' Reaction to a Concession on Its Evaluation in the Israeli-Palestinian Conflict." *The Harvard International Journal of Press/Politics* 11(4):70–88.

Press, Daryl G, Scott D Sagan and Benjamin A Valentino. 2013. "Atomic aversion: experimental evidence on taboos, traditions, and the non-use of nuclear weapons." *American Political Science Review* 107(01):188–206.

Rubin, Donald B. 1980. "Comment." *Journal of the American Statistical Association* 75(371):591–593.

Small, Deborah A., Jennifer S. Lerner and Baruch Fischhoff. 2006. "Emotion Priming and Attributions for Terrorism: Americans' Reactions in a National Field Experiment." *Political Psychology* 27(2):289–298.

Turner, Joel. 2007. "The messenger overwhelming the message: Ideological cues and perceptions of bias in television news." *Political Behavior* 29(4):441–464.

Wilson, Timothy D., Elliot Aronson and Kevin Carlsmith. 2010. *The Art of Laboratory Experimentation*. John Wiley & Sons, Inc.

Zhang, Junni L and Donald B Rubin. 2003. "Estimation of causal effects via principal stratification when some outcomes are truncated by "death"." *Journal of Educational and Behavioral Statistics* 28(4):353–368.