Placebo Selection in Survey Experiments: An Agnostic Approach

Ethan Porter[1,2,3] and Yamil R. Velez[4*]

[1] School of Media and Public Affairs, The George Washington University, 805 21st Street NW, Washington, DC 20052, USA. Email: evporter@gwu.edu
[2] Institute for Data, Democracy & Politics, The George Washington University, 805 21st Street NW, Washington, DC 20052, USA
[3] Department of Political Science, The George Washington University, 805 21st Street NW, Washington, DC 20052, USA
[4] Department of Political Science, Columbia University, 741 International Affairs Building, New York, NY 10027, USA. Email: yrv2004@columbia.edu

Abstract

Although placebo conditions are ubiquitous in survey experiments, little evidence guides common practices for their use and selection. How should scholars choose and construct placebos? First, we review the role of placebos in published survey experiments, finding that placebos are used inconsistently. Then, drawing on the medical literature, we clarify the role that placebos play in accounting for nonspecific effects (NSEs), or the effects of ancillary features of experiments. We argue that, in the absence of precise knowledge of NSEs that placebos are adjusting for, researchers should average over a corpus of many placebos. We demonstrate this agnostic approach to placebo construction through the use of GPT-2, a generative language model trained on a database of over 1 million internet news pages. Using GPT-2, we devise 5,000 distinct placebos and administer two experiments (N = 2,975). Our results illustrate how researchers can minimize their role in placebo selection through automated processes. We conclude by offering tools for incorporating computer-generated placebo text vignettes into survey experiments and developing recommendations for best practice.

Keywords: survey experiments, placebos

1 Introduction

The rationale supporting placebo usage in medical and pharmacological research is well-justified. Researchers wish to distinguish the effects of the putative treatment from ancillary factors associated with treatment delivery. Differences between treatment and placebo effects have inspired a long literature across multiple disciplines, including medicine and clinical psychology (e.g., Colloca and Benedetti 2005; Geersa and Miller 2014). In medical trials, placebo effects can be traced to specific neurobiological mechanisms (Colloca and Barsky 2020). Within political science, placebos are used in field experiments to account for treatment noncompliance (Nickerson 2005; Gerber *et al.* 2010). In survey experiments, however, justifications for the use of placebos are hard to come by. Instead, many survey experiments that make use of placebos appear to rely on received wisdom more than available evidence. Little research exists concerning the circumstances under which survey experimenters should use placebos and what kinds of placebos they should use.

Addressing this gap is of critical importance, as it has implications for one of the most basic aspects of experimental design: when estimating the treatment effect ($\mathbb{E}[Y_1 - Y_0]$), which Y_0 should we use?[1] While "pure control" conditions directly measure outcomes in the absence of the treatment, placebo conditions tend to hold the treatment mode constant (e.g., text placebos for text treatments) but, at least in theory, provide outcome-irrelevant information or cues before measuring endpoints. Though units assigned to the "pure control" condition may have well-

***Political Analysis* (2022)**
vol. 30: 481–494
DOI: 10.1017/pan.2021.16

Published
14 June 2021

Corresponding author
Yamil R. Velez

Edited by
Sunshine Hillygus

*Authors' names appear in alphabetical order.

[1] Here, Y_1 represents outcome scores for those under treatment and Y_0 represents outcome scores for those under "baseline," however defined.

defined potential outcomes, placebo conditions can become a kind of moving target: outcomes for placebo arms can be higher, lower, or identical to the treatment. One way of reducing this indeterminacy is to assume that the placebo condition has no effect on the outcome. However, if we follow this logic to its most extreme implication, then a control condition alone is sufficient for identification. While "three group" field experiment protocols invoke this assumption to recover effects in the presence of treatment noncompliance (Gerber *et al.* 2010), this often does not apply in the context of survey experiments, for which compliance tends to be high. Because placebo conditions could maximize *or* minimize treatment effects, determining best practices for their selection is crucial for inference.

In this paper, we investigate the use of survey experiments in political science and offer scholars recommendations on best practices for placebo selection. First, we document the frequency of, and justifications for, placebo conditions in published survey experiments. We find that they are used for inconsistent purposes across studies. Our review also reveals that researchers have substantial discretion when choosing placebos. As we argue, this discretion may have worrisome consequences for inference; selecting different placebos may yield different treatment effect estimates for otherwise identical experiments. Then, in formalizing the role of placebos in treatment effect estimates, we clarify the role that placebos play in accounting for nonspecific effects (NSEs). In the medical literature, NSEs are understood as all factors apart from the active ingredients of a drug that might affect patient behavior (Montgomery and Kirsch 1997; Vambheim and Flaten 2017). Rather than relying on intuition or guesswork to craft single placebos, we argue instead that researchers should rely on a large number of placebos from which they can average over.

We demonstrate this agnostic approach to placebo selection via a novel means of constructing placebos: the use of GPT-2, a generative language model trained on a database of over a million internet web pages. We conduct two survey experiments in which we relied on GPT-2 to generate 5,000 distinct placebos that resemble short news vignettes. In order to construct placebos that resembled those used in political science, we seeded GPT-2 with the prompt "today" to generate the placebo news articles. Subjects in the placebo conditions were randomly assigned to distinct autogenerated placebos. Through this process, we eliminated our ability as researchers to select placebos, relying on a far larger number of placebos than is standard in the process. Otherwise, our experiments are replications of Nelson, Clawson, and Oxley (1997) and Mullinix *et al.* (2015). In both experiments ($N = 2{,}975$), while the treatment effects of interest largely replicated, placebo effects proved indistinguishable from zero, producing an estimate of 1% of a control group standard deviation. There was variation in effect sizes due to placebo content, with apolitical automated placebos being more likely to lead to significant effects than political placebos. This finding both illustrates how placebo selection can affect treatment estimates and shows how automated processes can be used to create placebos that account for possible NSEs, such as the topic of the experiment.

Taken together, our evidence leads us to recommend that, at least in text-based survey experiments, researchers who wish to use placebos should minimize their own role in placebo selection. This can be done by turning to automated text generation processes, which can be refined to meet the specific needs of researchers. To facilitate this process, we offer an API for incorporating our design in survey platforms like Qualtrics, and a Python script for creating computer-generated placebos. The automated process we recommend is flexible to researcher needs, allowing for adjustments that account for specific NSEs that a researcher anticipates in advance. Researchers who wish to adopt our approach can and should create corpora specific to the NSEs they intend to address. They can then sample from this corpus to better understand the distribution of placebo effects.

https://doi.org/10.1017/pan.2021.16 Published online by Cambridge University Press

2 The Use of Placebos in Survey Experiments

To assess the role of placebos in survey experiments, we conducted a review of articles published in political science journals between 2009 and 2020 that mentioned the word "placebo." We searched in the *American Journal of Political Science, American Political Science Review, Journal of Politics, Political Psychology, Political Behavior, Public Opinion Quarterly, International Organization*, and *Comparative Political Studies*. We then inspected each article individually, to manually remove those that conducted placebo tests on observational data or, more broadly, used the term for purposes unrelated to experiments. This left us with 22 articles (excluding articles based on field experiments that used placebos). We would like to note that our data collection process may very well undercount the number of placebos, given that scholars do not exclusively use the term "control condition" to refer to pure controls. This underscores the need for precise language to describe baseline conditions in survey experiments, a recommendation we return to in the conclusion.

Our review, summarized in Table 1, indicates that there is little consensus about the use of placebos in survey experiments. Although all placebo conditions in our review hold treatment mode constant (e.g., pairing a text treatment with a text placebo), there is disagreement about whether placebos should, or should not, yield detectable effects. By our count, 45% of the articles tested placebos that could be described as tests of alternative hypotheses. These placebos were expected to generate significant differences from a pure control, but were used to account for a possible experimental confound. For instance, a study on political cues used a placebo that reminded participants about an upcoming election to rule out "election priming" as an alternative explanation. We also encountered placebo conditions that were designed to assess whether two conceptually distinct treatments exerted different effects. These types of placebos strain the conventional definition of placebos.

Fifty-five percent of the papers used placebos on the assumption that they would yield null effects. However, when examining the actual content of placebo conditions, we find that only 27% of studies used placebos with unambiguously apolitical content, with the rest providing theoretically outcome-relevant information to subjects. In sum, researchers agree on mode constancy while sharply disagreeing on whether placebos should (a) yield effects distinguishable from control or (b) convey outcome-relevant information to subjects.

As the previous analysis shows, researchers have significant discretion over placebo selection. To understand the potential consequences of this discretion, consider three possible placebo types: a placebo with effects in the opposite direction of the treatment effect; a placebo with effects in the same direction as the treatment effect; and a placebo with effects close to the pure control. These placebos would have starkly divergent implications for inference, with the first placebo amplifying the effect and increasing the possibility of a Type I error, and the third increasing the possibility of a Type II error.

Unfortunately, in advance of conducting the experiment, researchers do not know what kind of error, if any, they may be inducing. Given several recent scientific developments, this is worrisome.

Table 1. Summary statistics for political science articles using placebo conditions.

	%
Hold treatment mode constant	100
Assume the placebo is inert	55
Test an alternative hypothesis	45
Present apolitical information	27
N	22

The first is the prevalence of treatments that have yielded significant effects in one version of an experiment while proving inert in a replication (e.g., Nosek and Open Science Framework 2015). The difficulties that many scholars have had replicating treatment effects has implications for placebos: in a single experiment, a placebo may have larger effects than expected, precluding the researcher from identifying treatment effects. Even if this placebo effect was not to replicate in subsequent experiments, the presence of placebo effects in the initial experiment may be responsible for a Type II error. Second, several recent, robust survey experiments in political science have detected modest effects (Broockman and Kalla 2020; Coppock, Hill, and Vavreck 2020). Researchers who inadvertently use placebos with effects in the same direction as the treatment may have difficulty discerning small but significant treatment effects. Such researchers would have allocated finite resources to their placebo and also may have made distinguishing between placebo and treatment effects more difficult, again increasing the possibility of Type II error. Finally, in light of evidence about publication bias (Dickersin *et al.* 1987; Franco, Malhotra, and Simonovits 2014), researchers may be tempted to use placebos that shift outcomes in a negative direction relative to a pure control to heighten their probability of detecting significant effects, even at the cost of a Type I error.

3 Placebos and Potential Outcomes

We surmise that the lack of consensus in the literature regarding placebo conditions can be traced to inconsistent definitions of the concept and uncertainty over how placebo conditions should be used. To provide more structure on this problem and illuminate the advantages of our proposed placebo sampling design, we draw on the medical sciences literature and produce a brief formalization of what, specifically, can be learned from a placebo-controlled experiment. Moreover, using this formalization, we show that our agnostic approach makes treatment effect estimates less sensitive to the choice of a single placebo. First, we discuss the standard placebo design, and then move on to a discussion of our method.

We employ the Neyman–Rubin potential outcomes framework (Rubin 2005). Consider an experiment with N respondents. Each respondent is indexed by $i \in \{1, \ldots, N\}$. There are three experimental conditions ($T_i = 0$, $T_i = 1$, $T_i = 2$), with $T_i = 0$ representing a pure control condition, $T_i = 1$ representing the intervention, and $T_i = 2$ representing a placebo condition. We represent potential outcomes under control, treatment, and placebo in Table 2.

As shown in Table 2, there are two key parameters: τ and γ. While τ represents the theoretically relevant portion of the intervention, γ represents the effect of extraneous features of the intervention, such as the treatment delivery mechanism or mode. In a medical trial, τ represents the active ingredients of a pharmaceutical drug, whereas γ represents auxiliary aspects of the treatment environment that might otherwise produce effects via patient expectations or conditioned behaviors (Montgomery and Kirsch 1997). The latter are referred to as NSEs in the medical sciences literature, a naming convention we employ here.

The logic underlying placebo-controlled designs is that assigning respondents to a condition that lacks the intervention, but is otherwise identical to the treatment condition, allows the impact

Table 2. Potential outcomes under control, treatment, and placebo.

i	$Y_i(0)$	$Y_i(1)$	$Y_i(2)$
1	$Y_1(0)$	$Y_1(0) + \tau_1 + \gamma_{11}$	$Y_1(0) + \gamma_{21}$
...	...	...	...
N	$Y_N(0)$	$Y_N(0) + \tau_N + \gamma_{1N}$	$Y_N(0) + \gamma_{2N}$

of NSEs to be subtracted out. However, this relies on the assumption that $\gamma_{1i} = \gamma_{2i}$, or that NSEs are equivalent across the treatment and placebo conditions (Colloca and Benedetti 2005). If $\gamma_{1i} = \gamma_{2i}$, $Y_i(1) - Y_i(2)$ recovers τ_i and $Y_i(2) - Y_i(0)$ recovers γ_i (or the individual placebo effect). Due to the fundamental problem of causal inference, these parameters cannot be estimated for each unit. However, if we assume that potential outcomes are independent of treatment assignment, the placebo-controlled average treatment effect ($PCATE$) is given by $\mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(2)]$ and the APE can be defined as $\mathbb{E}[Y_i(2)] - \mathbb{E}[Y_i(0)]$. This implies that the average effect of the theoretically relevant portion of the intervention can be recovered using a simple difference-of-means comparing subjects in the placebo condition to those in the treatment condition. The first quantity ($PCATE$) can be estimated as:

$$\widehat{PCATE} = \frac{1}{n_1} \sum_{i:T_i=1} Y_i(1) - \frac{1}{n_2} \sum_{i:T_i=2} Y_i(2), \tag{1}$$

where n_1 represents the number of units assigned to the treatment condition and n_2 represents the number of units assigned to the placebo condition.

Similarly, the second quantity, the APE, can be estimated using the following equation:

$$\widehat{APE} = \frac{1}{n_2} \sum_{i:T_i=2} Y_i(2) - \frac{1}{n_0} \sum_{i:T_i=0} Y_i(0), \tag{2}$$

where n_0 represents the number of units assigned to the control condition.

As stated above, equality of NSEs is required to recover the $PCATE$. That is, one must assume that γ_{1i} (or the NSE in the treatment condition) is identical to γ_{2i} (or the NSE in the placebo condition).[2] Otherwise, the $PCATE$ reflects a combined effect of τ and the difference between γ_1 and γ_2. Departures from the equality assumption might be observed when auxiliary factors vary across conditions. For example, in a study estimating the effect of exposure to hostile political discourse on trust in government, where the NSE is "reading an engaging article," equality could be called into question if the placebo article does not match the level of engagement observed in the treatment condition.[3]

Let us now consider potential outcomes under multiple placebos, with each placebo indexed by j ∈ {2, . . ., K+1}, where K is the total number of placebo conditions. Placebo-controlled survey experiments in political science typically set K to 1, and thus, recovery of $\mathbb{E}[\tau_i]$ depends on whether $\mathbb{E}[\gamma_{1i} - \gamma_{2i}] = 0$, where γ_1 represents the NSE in the treatment condition and γ_2 represents the NSE in the placebo condition. If multiple placebo conditions are used, and subjects are randomly assigned to one of K placebos, we can use the following equation to define the placebo sampling-controlled average treatment effect ($PSCATE$):

$$PSCATE = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(T_i) | T_i \geq 2], \tag{3}$$

where $T_i \geq 2$ represents the entire set of K placebos.

Under randomization,[4]

$$\widehat{PSCATE} = \frac{1}{n_1} \sum_{i:T_i=1} Y_i(1) - \frac{1}{n_2} \sum_{i:T_i \geq 2} Y_i(T_i \geq 2), \tag{4}$$

[2] We use the singular form here, but both parameters could also represent a bundle of NSEs.

[3] Another assumption is additivity. Additivity holds that the treatment condition is an additive combination of the intervention and the NSE. This assumption is violated when the intervention (or NSE) modifies the effect of the NSE (or the intervention). However, violations of this assumption are not as well defined as violations of equality.

[4] The APE for a placebo corpus can be estimated via $1/n_2 \sum_{i:T_i \geq 2} Y_i(T_i \geq 2) - 1/n_0 \sum_{i:T_i=0} Y_i(0)$.

https://doi.org/10.1017/pan.2021.16 Published online by Cambridge University Press

which provides an unbiased estimate of the $PSCATE$. The $PSCATE$ can be decomposed in the following way:

$$PSCATE = \mathbb{E}[\tau_i + \gamma_{1i}] - \mathbb{E}[\gamma_{ji} \mid T_i \geq 2]. \tag{5}$$

If we assume equal assignment probabilities and $\mathbb{E}[\gamma_{1i}] = \frac{\sum_{j=2}\mathbb{E}[\gamma_{ji}]}{K}$, $\widehat{PSCATE}$ provides an unbiased estimate of $\mathbb{E}[\tau_i]$.[5] Put another way, one can recover $\mathbb{E}[\tau_i]$ if the NSE in the treatment condition is equivalent to the average NSE estimated for a given placebo corpus (or APE).

If the NSE for a single placebo (when $K = 1$) is representative of the broader pool of placebos, there is no advantage to placebo sampling. However, it is not possible to know whether one is using a placebo condition with an uncharacteristically large or small γ_j without having multiple placebo conditions to compare across. This leads us to advocate for an agnostic approach to placebo construction, where units assigned to the placebo condition are randomly assigned to a single placebo from a larger corpus. The logic underlying this design choice is based on the idea of "stimulus sampling," whereby instead of using a single stimulus to represent a theoretical construct, stimuli are randomly sampled from a distribution (Wells and Windschitl 1999). This design choice reduces researcher degrees of freedom—given that we sample from a larger population of placebo vignettes—and enables us to explore variation in placebo effects. While it is difficult to estimate precise effects for any single placebo using this approach due to the small number of respondents receiving the same placebo, this design choice still ensures that the APE will be representative of the corpus from which it is drawn. It also reduces the sensitivity of estimates to the use of a single placebo j.

What if NSEs between the treatment and placebo corpus are not equivalent when using the placebo sampling design? Under these conditions, we run into the same issues as those in the single placebo case, with $\mathbb{E}[\tau_i]$ being a function of τ and the difference between the γ's. However, we contend that multiple placebos will still provide more theoretically meaningful benchmarks than a single placebo. For instance, suppose the putative intervention is "exposure to an uncivil political discussion" and a vignette describing a competitive horse race is used as a placebo to account for the NSE associated with reading an engaging story. Technically, the treatment effect reflects the mean difference in outcomes between those assigned to read about an uncivil political discussion *versus* those assigned to read about a competitive horse race. While this might be an interesting comparison for some, we expect that most scholars would prefer to estimate the difference between the putative intervention and the average effect of reading an engaging article. Put another way, even if the NSE cannot be subtracted out due to a violation of assumptions, the average effect of reading an engaging article is likely a more useful experimental benchmark than the average effect of a single placebo. With a pure control, one can use this information to compare effect sizes, estimating whether the effect of the intervention is larger, smaller, or equivalent to the effect of alternative interventions that are also theoretically meaningful.

4 An Agnostic Approach to Placebo Construction

As our review of the literature demonstrated, researchers use placebos in inconsistent ways, with potentially troubling implications for inference. We have also shown that the placebo-controlled design, while elegant on its face, depends on strong assumptions. Relying on a single placebo allows for $\mathbb{E}[\tau_i]$ to be recovered under the precise condition that $\mathbb{E}[\gamma_{1i} - \gamma_{2i}] = 0$. We propose an agnostic approach to placebo construction—a placebo sampling design—where individual

[5] In the case of unequal assignment probabilities, this condition becomes $\mathbb{E}[\gamma_{1i}] = \mathbb{E}[\sum_j p_j \gamma_{ji}]/\mathbb{E}[\sum_j p_j]$, where p_j represents the assignment probability in the population.

https://doi.org/10.1017/pan.2021.16 Published online by Cambridge University Press

placebos are sampled from a corpus to better characterize NSEs. In this section, we implement a version of this design by administering replications of two well-known survey experiments.

Each respondent participated in two survey experiments. The experiment order was randomized. We sought to replicate studies with unrelated treatment content and outcome measures for the purpose of assessing variation in placebo effects across different kinds of survey experiments. The first experiment was based on a canonical framing study by Nelson *et al.* (1997), where respondents are randomly assigned to read a news article highlighting free speech or public order as relevant considerations for determining whether the Ku Klux Klan (KKK) should march on a college campus. The second experiment was based on the "student loans" study featured in Mullinix *et al.* (2015), which randomly exposed participants to a frame describing student loan repayment as a matter of individual responsibility and measured preferences for student loan forgiveness. Within each experimental block, participants were randomly assigned to treatment conditions based on the original studies, a placebo condition, and a pure control condition. The pure control condition solely measured outcome variables, whereas those in the placebo condition were randomly assigned to read one of the computer-generated placebo vignettes.[6]

To generate vignettes, we used OpenAI's GPT-2, a generative language model trained on a massive database of over 8 million documents (Radford *et al.* 2019). GPT-2 is a deep neural network involving over 1.5 billion parameters that is capable of answering queries, generating news articles, writing poems, and translating documents.[7] GPT-2 has been shown to produce "convincing" autogenerated text, especially in the context of news. (Indeed, initial release of the full version of GPT-2 was delayed because of its potential for misuse [Zellers *et al.* 2019].)

We leverage GPT-2 for three reasons. First, while GPT-2 can generate text without any text prompts, it can be fine-tuned to produce text from various text genres. For example, seeding GPT-2 with the prompt "import numpy" can produce computer code. Given the use of news articles as placebo vignettes, and the fact that GPT-2 was trained on news texts (Radford *et al.* 2019, pg. 2), the ability for GPT-2 to generate news articles allows us to construct vignettes that are similar to standard placebos. Second, GPT-2 can produce autogenerated output without much user input. As we have argued, researcher degrees-of-freedom are a distinct limitation of the traditional placebo-controlled survey experiment, given that scholars can select placebo vignettes that maximize treatment effects. Our 5,000 placebo vignettes are generated simply by using "Today," as a prompt (see Section 5 for instructions on how to access our placebo database). Our output spans autogenerated news articles about technology, entertainment, domestic politics, and other topics. Third, GPT-2 is publicly available and open source, allowing scholars to improve upon the vignettes presented here. Indeed, an advantage of using GPT-2 to build a corpus when compared to other sources (e.g., LexisNexis) is that it (1) obviates the need to rely on data collection practices like scraping that can violate terms of service agreements, and (2) allows scholars to craft corpus content using different seed words, sentences, or paragraphs. At the end of our paper, we provide links to example code and an API that scholars can incorporate into their work should they decide to include a placebo condition.

It is important to note that the placebo sampling design does not depend on the use of GPT-2. A placebo corpus can be constructed using other sources such as the New York Times corpus (Snowsill *et al.* 2010). However, text generation methods like GPT-2 offer customization of content,

[6] Placebos with errors (e.g., run-on sentences and incomplete sentences) were programmatically removed prior to fielding the experiment, yielding 4,696 total vignettes out of the 5,000 that were generated. Given that some placebo vignettes describe events that never happened, we debriefed all participants by informing them that the vignettes were computer-generated. Though we did not encounter any vignettes with explicitly toxic content, scholars should be aware that generation of such vignettes is possible.

[7] GPT-2's neural network uses a transformer-based architecture. In contrast to recurrent neural networks, which sequentially pass output along a multilayer neural network, transformers do not make assumptions about spatial or temporal interdependencies, can calculate layer outputs in parallel as opposed to sequentially, and allow for long-range dependencies between words.

https://doi.org/10.1017/pan.2021.16 Published online by Cambridge University Press

Table 3. Average placebo effect.

	Estimate	Std. error	95% CI
Student loans frame	.13	.04	[.04, .21]
Free speech frame	.21	.05	[.12, .30]
Public safety frame	.12	.05	[.03, .22]
Placebo	.01	.03	[−.06, .08]
N	5,903		

which is important for ruling out precise NSEs. This possibly comes at the cost of realism, as subjects might be able to detect that some of these placebos are generated by a computer. However, GPT-3, the successor to GPT-2, performs even better than GPT-2 on "Turing test" metrics that gauge if subjects can identify whether a piece of text was created by a computer or human being (Brown *et al.* 2020).[8]

We recruited participants via the online crowdsourcing platform *Amazon Mechanical Turk*, as similar studies of survey experiments have done (White *et al.* 2018; Mummolo and Peterson 2019).[9] Data collection was carried out in two waves spaced approximately 1 month apart (June 18, 2020 and July 15, 2020; Velez and Porter 2021).[10] In total, 2,975 participants completed the study. An advantage of our multiexperiment design is that we can pool responses across both experiments to estimate placebo effects with even more precision. Moreover, given the amount of variation in placebo content generated by GPT-2, maximizing the number of observations exposed to placebo conditions provides an opportunity to explore placebo heterogeneity.

For our key analysis, we model outcomes using the following equation:

$$Y_i = \beta_0 + \beta_1 P_i + \beta_2 S_i + \beta_3 W_i + \textstyle\sum_{k=4}^{7} \beta_k Z_i + \epsilon_i, \tag{6}$$

where Y_i represents the outcome measure in control group standard deviations, P_i represents a placebo condition indicator, experimental block fixed effects are represented as S_i, W_i is a survey wave fixed effect, Z_i is a vector of treatment indicators (three in total), and ϵ is an error term. β_1 recovers $\mathbb{E}[\gamma_j]$, or the APE across placebo conditions. We estimate this equation using linear regression with respondent clustered standard errors, given that each respondent participated in two experimental blocks.

As shown in Table 3, the APE is not only statistically indistinguishable from zero, but also substantively small, representing 1% of a control group standard deviation (SE = .03, p = .73). In contrast, the framing interventions have effect sizes that are orders of magnitude larger, ranging from .12 to .21 standard deviations. Put another way, the effect of "reading a random news vignette" is approximately .01 control group standard deviation units, and accounts for a small share of the persuasive effects observed in the experiments, assuming that placebo assumptions hold. Still, scholars may want to adjust for even more precisely defined NSEs, such as the effect of "reading a positively valenced news vignette" or "reading a political news vignette." We examine these potential NSEs below.

To estimate valence, we rely on the Valence Aware Dictionary and Sentiment Reasoner (VADER), a lexical and rule-based sentiment analysis tool that generates a normalized valence score for text

[8] At the time of writing this article, it is not yet open source or publicly available, but our expectation is that potential issues regarding realism or plausibility will diminish as this technology improves.

[9] We used CloudResearch's Mechanical Turk tool to filter out duplicate IP addresses and suspicious accounts (Litman, Robinson, and Abberbock 2017).

[10] We proceeded with a second wave to increase the precision of our estimates, given that there was a considerable amount of uncertainty in point estimates following first wave data collection (N = 1,382). We include a wave fixed effect in all of our analyses.

https://doi.org/10.1017/pan.2021.16 Published online by Cambridge University Press

Table 4. Average placebo effect by valence.

	Estimate	Std. error	95% CI
Negatively valenced placebo	.00	.05	[−.09, .09]
Positively valenced placebo	.02	.05	[−.06, .09]
N	3,497		

Table 5. Average placebo effect by political content.

	Estimate	Std. error	95% CI
Political placebo	.04	.04	[−.04, .11]
Apolitical placebo	−.02	.04	[−.11, .06]
N	3,497		

(Gilbert and Eric 2014). VADER scores are normalized to range from −1 to 1, with 0 representing a neutral midpoint. Positively valenced vignettes are those with a valence estimate greater than or equal to 0, whereas negatively valenced vignettes are those with a valence estimate less than 0. Examining Table 4, the APE for negatively valenced placebos is approximately zero (or .0005 to be precise; SE = .05, p = .99), whereas the positively valenced placebo APE is .02 (SE = .05, p = .65). Thus, if an intervention is framed using positively valenced words, .02 control group standard deviations can be subtracted from the estimated treatment effect to recover $\mathbb{E}[\tau_i]$, if the requisite assumptions hold. We now move on to our analysis of political and apolitical placebos.

Following the two experimental blocks, participants were asked to categorize the content of their assigned vignettes using the following categories: business, U.S. politics, international politics, sports, entertainment, science, media, crime, weather, personal story, or other. We define political placebos as those tagged as having to do with U.S. or global politics, and apolitical placebos as those pertaining to all other topics. As shown in Table 5, the APE for political placebos is 4% of a standard deviation (SE = .04, p = .33) and −2% of a standard deviation for apolitical placebos (SE = .04, p = .58). Though the gap between political and apolitical APEs is .06 control group standard deviations, this difference is not statistically significant at conventional levels of statistical significance (p = .19). However, the opposite signs for political and apolitical APE's underscore the possibility that one might draw different inferences depending on the kind of placebo that is used. To illustrate this possibility, we estimate the ATE (and corresponding confidence intervals) for the student loans message—along with the APE's for political and apolitical placebo conditions—and plot them in Figure 1.

If one were to use the political placebo as the relevant baseline, the estimated PSCATE would be .09 standard deviations (± .11 standard deviations), and not statistically distinguishable from zero at conventional levels of statistical significance. When an apolitical placebo is used, however, this estimate jumps to .15 standard deviations (± .11 standard deviations), reflecting a 33% increase in effect size, and is now statistically significant. Therefore, placebo selection can impact effect size estimates and, consequently, whether or not a significant treatment effect is recovered. Our evidence suggests that, at least in political science survey experiments, political placebos may be more likely to yield null effects than apolitical placebos and thus might serve as more conservative tests. This should not be especially surprising, as it corroborates the common assumption that placebos are most useful when they hold some aspect of the treatment constant (thereby removing an NSE).

https://doi.org/10.1017/pan.2021.16 Published online by Cambridge University Press

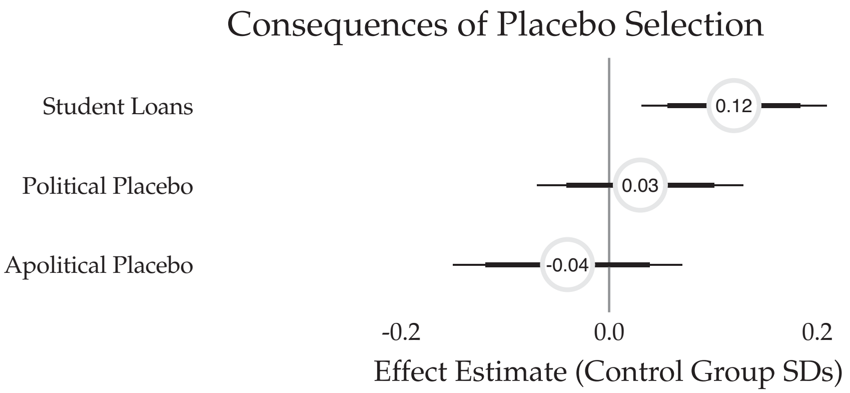

Figure 1. The importance of placebo selection. Effect estimates for placebos and the student loans treatment. Ninety-five percent (narrow bar) and 84% (heavy bar) confidence intervals are shown. Eighty-four percent confidence intervals (heavy bar) allow for visual tests of equality across conditions; the use of 95% confidence intervals results in Type II errors when comparing visible coefficients.

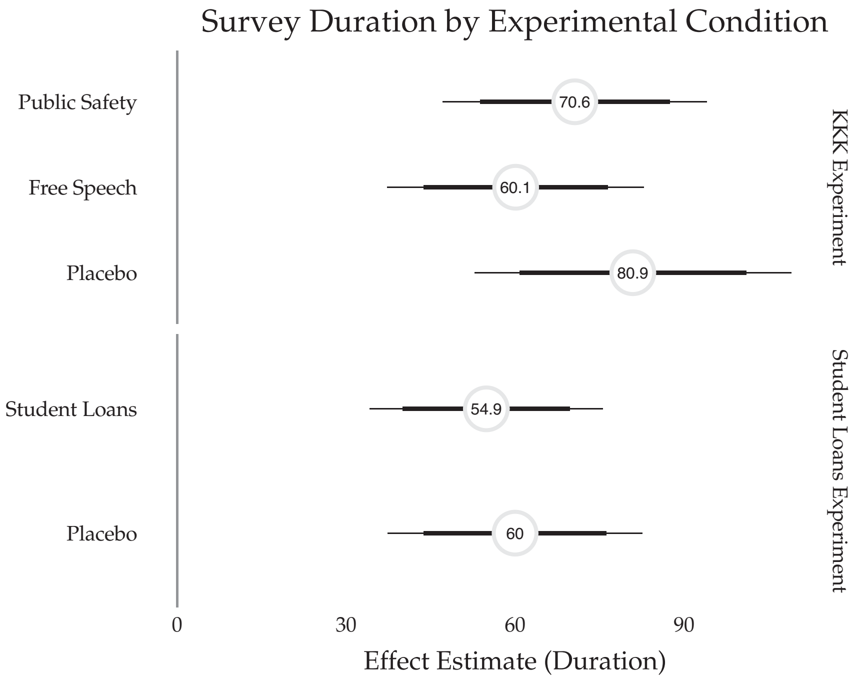

Figure 2. Effect of experimental conditions on survey duration in seconds. Ninety-five percent (narrow bar) and 84% (heavy bar) confidence intervals are shown. Eighty-four percent confidence intervals (heavy bar) allow for visual tests of equality across conditions; the use of 95% confidence intervals results in Type II errors when comparing visible coefficients.

4.1 Do Computer-Generated Placebos Suffer from Noncompliance?

One potential risk of relying on computer-generated placebos is that individuals assigned to placebo conditions may deem those placebos unrealistic and fail to comply as a result. Estimates of placebo effects in our experiments may be minimized due to noncompliance. $APEs$ could thus be negligible due to compliance issues, rather than indicating a null effect of the NSE *per se*. We assess this by examining whether those assigned to the placebo condition spend less time on the survey relative to those assigned to other conditions (Harden, Sokhey, and Runge 2019). As shown in Figure 2, participants assigned to experimental arms versus the pure control condition spend more time completing the survey (which is captured by the positive and significant ATE estimate for all conditions). However, none of the differences in survey duration estimates are

https://doi.org/10.1017/pan.2021.16 Published online by Cambridge University Press

significant between experimental arms and placebo conditions. In sum, we do not find evidence of differential noncompliance due to the nature of the placebo conditions.

5 API for Computer-Generated Placebos

To expand the number of potential NSEs placebos control for, and to minimize the possibility that researchers' choice of placebos will affect their estimates, we recommend that those who wish to use placebos rely on the automated text-generation process used above. Here, we provide tools that can be readily used and adapted by scholars. Given that this is an area of ongoing research, we encourage future scholars to improve upon the general method we describe here. Adopting and improving upon these tools can limit the possibility that placebo selection will unduly influence treatment effect estimates. For scholars interested in adopting our placebo design with minimal modifications, we have created a Qualtrics-compatible API that accepts an integer between 1 and 4,696 and returns a computer-generated placebo vignette (see online Appendix B for full list of placebo vignettes and implementation instructions).

Using our tool, researchers can assign a random integer to respondents in the placebo condition, call the API, and pipe text from the API into the relevant survey block. This API can be accessed using http://ourlocalcommunities.com/placebo.php?num=x, where *x* is the assigned integer. For those who desire more customization or would like to alter the prompt for generating the placebo vignettes, a Google Colaboratory notebook can be accessed at https://bit.ly/placebo_tools.[11] This code can be modified to (1) produce shorter or longer vignettes by varying the *length* parameter, (2) produce specific kinds of placebo texts by using different prompts (e.g., "Today in sports"), and (3) generate more "surprising" text by increasing the *temperature* parameter. The generated text can then be directly used in Qualtrics or uploaded to a database where placebo vignettes can be retrieved via database queries (as in our API above).

6 Recommendations and Discussion

Although placebos play a familiar role in political science survey experiments, little empirical evidence or theoretical guidance informs their usage. In contrast, the medical literature has identified precise neurobiological mechanisms that lead to placebo effects (Colloca and Barsky 2020). Despite the absence of similarly identified mechanisms in political science, survey experimenters make frequent use of them. Our review of the literature, however, shows that researchers use placebos in inconsistent ways, with some explicitly expecting placebos to generate significant effects, while others expect effects to be indistinguishable from control. In some cases, researchers use placebo terminology to describe tests of alternative hypotheses. The discretion that researchers have in choosing placebos may lead to systematic Type I and Type II errors, particularly when assumptions are violated. In addition, relying on a single placebo requires NSEs to be equivalent to NSEs in the treatment condition.

To investigate means of minimizing researcher discretion in placebo selection, we replicated two well-known survey experiments while relying on a corpus of computer-generated text placebos. Respondents assigned to placebo conditions were exposed to placebos created by GPT-2, a generative language model, with refinements made, so that the placebos more closely approximated those typically used in survey experiments. Using this approach, the APE in our study—which can be thought of as the average NSE of reading a text vignette—was both indistinguishable from zero and quite small. Furthermore, the use of apolitical placebos resulted in significant treatment effects, whereas the use of political placebos did not.

Relying on computer-generated placebos thus serves multiple purposes. First, as it minimizes the researcher role in placebo selection, it also proves capable of yielding effects that are not

[11] This script draws heavily on Max Woolf's GPT-2 Google Colaboratory script (Woolf 2020).

https://doi.org/10.1017/pan.2021.16 Published online by Cambridge University Press

distinguishable from zero. The APE of the GPT-2-generated placebos indicates that such an approach results in a mostly neutral placebo that neither amplifies nor diminishes the treatment effect estimate to a substantial degree. Second, to the extent that researchers are concerned about adjusting for NSEs, computer-generated placebos can be evaluated in ways that permit them to control for such confounds, again while minimizing the researcher's role.

Scholars operating under resource constraints might only have funds to use one type of baseline condition. Under what circumstances might pure controls be preferable to placebos, and vice versa? In online Appendix D, we find that using a placebo condition as the baseline often produces a smaller treatment effect estimate than using a pure control. To the extent that a placebo condition protects one against the criticism of not accounting for NSEs and provides a more conservative test of hypotheses, using a placebo condition may be advisable. However, pure controls might be preferred over placebos if NSEs are assumed to be relatively small *a priori* or when there is no desire to "unbundle" a compound treatment.[12] If resources permit, the use of both a placebo and pure control condition enables scholars to benchmark the effect size of their intervention against a representative set of alternative experimental conditions, even when placebo design assumptions do not hold. However, we recognize that this, along with any design choice, should be balanced against other considerations such as statistical power.

Researchers who wish to emulate our approach should begin by articulating NSEs they believe may come to affect their treatment effect estimates. If they do not have specific expectations about NSEs, but expect that reading a text vignette will exert an independent effect on the outcome, they can rely on the general "today" prompt we used above. If, however, they have more precise expectations, they can change the GPT-2 prompt to correspond with NSEs they anticipate, and rely on the resulting corpus. For example, if researchers who are studying how race affects views toward payment of college athletes wish to separate the effects of reading about college athletics from outcome measures, they could seed GPT-2 with a college sports-related prompt. If the presumed NSE is tied to reading a *political* news article, a more specific prompt about politics might be preferable. In any event, relying on automated processes will reduce the possibility that, however unintentionally, estimates are sensitive to the placebo chosen, while increasing the possibility that the placebo condition better represents the pool of possible placebos. Finally, we recommend that any baseline condition intending to rule out NSEs be referred to as "placebo conditions," reserving "pure controls" for baseline conditions that lack any informational content. As demonstrated by our analysis of the existing literature, there appears to be a lack of common terminology. Whether they choose to follow our agnostic approach to placebo construction or not, scholars should clearly distinguish between the two kinds of baseline conditions.

Which Y_0 should survey experimenters use? Ultimately, we recommend that researchers minimize their own role in answering the question. Instead, they can rely on automated processes to create many placebos and average over them. Such an approach can be tailored to accommodate the precise needs of researchers, who can modify the text-generation process to their liking. Scholars using video or images could also adopt the general principles of drawing placebos from a larger corpus when implementing placebo designs. For example, political advertising treatments could be compared against a corpus of the most popular advertisements in a given time period. Following this approach will help researchers avoid choosing placebos that affect whether a significant treatment effect is detected, and whether the theoretically relevant portion of the intervention or NSEs are responsible for those effects.

[12] Our estimates of placebo effects could also be used to inform priors. However, we caution that these effect sizes may not generalize beyond the domains of free speech and support for student loan forgiveness.

https://doi.org/10.1017/pan.2021.16 Published online by Cambridge University Press

Acknowledgments

We would like to thank Riki Conrey, Alex Coppock, Naoki Egami, John Kane, Josh Kertzer, Patrick Kraft, Brendan Nyhan, Michael Peress, and seminar audiences at GW for their feedback, and Jacob Liedke and Mark McKibbin for research assistance. All mistakes are our own. Author names appear in alphabetical order.

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2021.16.

References

Broockman, D., and J. Kalla. 2020. “When and Why Are Campaigns’ Persuasive Effects Small? Evidence from the 2020 US Presidential Election.” OSF Preprints. doi:10.31219/osf.io/m7326.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al. 2020. “Language Models Are Few-Shot Learners.” Preprint, arXiv:2005.14165.

Colloca, L., and A. J. Barsky. 2020. “Placebo and Nocebo Effects.” *New England Journal of Medicine* 382(6):554–561. http://doi.org/10.1056/NEJMra1907805.

Colloca, L., and F. Benedetti. 2005. “Placebos and Painkillers: Is Mind as Real as Matter?” *Nature Reviews Neuroscience* 6(7):545–552.

Coppock, A., S. J. Hill, and L. Vavreck. 2020. “The Small Effects of Political Advertising Are Small Regardless of Context, Message, Sender, or Receiver: Evidence from 59 Real-Time Randomized Experiments.” *Science Advances* 6(36):1–6.

Dickersin, K., S. Chan, T. Chalmers, H. Sacks, and H. Smith, Jr. 1987. “Publication Bias and Clinical Trials.” *Controlled Clinical Trials* 8(4):343–353.

Franco, A., N. Malhotra, and G. Simonovits. 2014. “Publication Bias in the Social Sciences: Unlocking the File Drawer.” *Science* 345(6203): 1502–1505.

Geersa, A. L., and F. G. Miller. 2014. “Understanding and Translating the Knowledge About Placebo Effects: The Contribution of Psychology.” *Current Opinion Psychiatry* 27(5):326–331.

Gerber, A. S., D. P. Green, E. H. Kaplan, and H. L. Kern. 2010. “Baseline, Placebo, and Treatment: Efficient Estimation for Three-Group Experiments.” *Political Analysis* 18(3):297–315. http://doi.org/10.1093/pan/mpq008.

Gilbert, C., and H. Eric. 2014. “Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.” In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, 81–82. Palo Alto, CA: AAAI Press.

Harden, J. J., A. E. Sokhey, and K. L. Runge. 2019. “Accounting for Noncompliance in Survey Experiments.” *Journal of Experimental Political Science* 6(3):199–202.

Litman, L., J. Robinson, and T. Abberbock. 2017. “Turkprime.com: A Versatile Crowdsourcing Data Acquisition Platform for the Behavioral Sciences.” *Behavior Research Methods* 49(2):433–442.

Montgomery, G. H., and I. Kirsch. 1997. “Classical Conditioning and the Placebo Effect.” *Pain* 72(1–2):107–113.

Mullinix, K. J., T. J. Leeper, J. N. Druckman, and J. Freese. 2015. “The Generalizability of Survey Experiments.” *Journal of Experimental Political Science* 2(2):109–138.

Mummolo, J., and E. Peterson. 2019. “Demand Effects in Survey Experiments: An Empirical Assessment.” *American Political Science Review* 113(2):517–529. http://doi.org/10.1017/S0003055418000837.

Nelson, T. E., R. A. Clawson, and Z. M. Oxley. 1997. “Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance. *American Political Science Review* 91(3): 567–583.

Nickerson, D. W. 2005. “Scalable Protocols Offer Efficient Design for Field Experiments.” *Political Analysis* 13(3):233–252. http://doi.org/10.1093/pan/mpi015.

Nosek, B., and Open Science Framework . 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349(6251): aac4716.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. “Language Models Are Unsupervised Multitask Learners.” *OpenAI Blog* 1(8):9.

Rubin, D. B. 2005. “Causal Inference Using Potential Outcomes: Design, Modeling, Decisions.” *Journal of the American Statistical Association* 100(469):322–331.

Snowsill, T., I. Flaounas, T. De Bie, and N. Cristianini. 2010. “Detecting Events in a Million New York Times Articles.” In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 615–618. Berlin: Springer.

Vambheim, S. M., and M. A. Flaten. 2017. “A Systematic Review of Sex Differences in the Placebo and the Nocebo Effect.” *Journal of Pain Research* 10:1831–1839.

Velez, Y., and E. Porter. 2021. “Replication Data for: Placebo Selection in Survey Experiments: An Agnostic Approach.” https://doi.org/10.7910/DVN/4PYOXP, Harvard Dataverse, V1, UNF:6:tg08CtjaYGUVLsprGUDh1A== [fileUNF].

https://doi.org/10.1017/pan.2021.16 Published online by Cambridge University Press

Wells, G. L., and P. D. Windschitl. 1999. “Stimulus Sampling and Social Psychological Experimentation.” *Personality and Social Psychology Bulletin* 25(9):1115–1125.

White, A., A. Strezhnev, C. Lucas, D. Kruszewska, and C. Huff. 2018. “Investigator Characteristics and Respondent Behavior in Online Surveys.” *Journal of Experimental Political Science* 5(1):56–67.

Woolf, M. 2020. “Gpt-2-Simple, a Python Package.” https://github.com/minimaxir/gpt-2-simple.

Zellers, R., A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. 2019. “Defending Against Neural Fake News.” In *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, 9054–9065.

https://doi.org/10.1017/pan.2021.16 Published online by Cambridge University Press