

Non-Compliance in Survey Experiments with Placebo Treatments

Nan Zhang¹ · Delia Baldassarri²

¹University of Mannheim

²New York University

Survey experiments often utilize placebo controls which aim to mimic the treatment condition while leaving the key explanatory variable unaffected. We illustrate the problems that can arise under placebo-controlled designs when compliance with treatment assignment is imperfect (e.g., when respondents are inattentive to experimental manipulations). We argue that estimating complier average causal effects (CACEs) under these conditions requires either (i) the inclusion of a “pure control” condition involving no experimental manipulation, or (ii) the strong assumption that compliance profiles are balanced across the treatment and placebo arms. We illustrate these points with data from a “failed” survey experiment on how the performance of ethnic minorities affects processes of social categorization.

Keywords: survey experiment; compliance; inattentive responding

1 Introduction

Survey experiments are widely used across the social sciences. By randomly assigning respondents across experimental conditions, researchers can estimate causal effects by simply comparing “treatment” and “control” group means. While some studies use “pure controls” in which no manipulation is employed, others utilize “placebo controls” which are similar to the treatment condition *except that the actual treatment variable remains unaffected*. In this way, researchers can “distinguish the effects of the putative treatment from ancillary factors associated with treatment delivery” (Porter and Velez 2022, p 481).

While placebo controls offer attractive design features, they may also present problems when compliance with treatment assignment is imperfect. Generally, scholars are concerned with whether respondents engage with or understand the experimental manipulations to which they have been assigned. Failure to attend to experimental stimuli means that respondents do not “take” their intended treatments. Additionally, however, patterns of non-compliance in placebo-controlled designs may differ across the treat-

ment and placebo arms, thereby destroying the *ex ante* balance created by randomization.

We illustrate these issues through a discussion of our own “failed” survey experiment in which we sought to examine how the performance of ethno-racial minorities affects whether they are classified in terms of their race or their nationality. In the context of international football, our treatment consisted of real video-clips of four different penalty shootouts in which the English national team was defeated: in the placebo condition white players missed the penalties; in the treatment condition non-white players failed their shots. Unexpectedly, however, attention to our videos varied widely across treatment arms, leading to an unanticipated situation of differential non-compliance. We also find that respondents who assimilated the intended information differed in baseline characteristics across various video treatments. Importantly, this imbalance in terms of who “took the treatment” in the experimental arms constitutes an experimental failure which precludes identifying a *complier average causal effect* (CACE) (Angrist et al. 1996).

Turning to “lessons learned,” we review how instrumental variables (IV) provide a ready solution to compliance problems. These strategies are particularly useful in cases such as ours where compliance profiles differ across experimental arms in unanticipated ways. In these instances, IV can be used to recover separate CACEs for each treatment

Corresponding author: Nan Zhang, University of Mannheim, Mannheim, Germany (Email: n.zhang@uni-mannheim.de)

and placebo arm. However, this solution is only available if researchers have included a *pure control condition* as a part of their design. Thus our key takeaway and recommendation: practitioners would do well to include a pure control in placebo-controlled experiments.

Our paper links the literature on survey experiments to a broader discussion of experimental non-compliance. While both placebo designs (Broockman et al. 2017; Nickerson 2005) and non-compliance (Gerber and Green 2012) have been explored within the literature on *field experiments*, such issues have largely flown under the radar of survey researchers. For instance, in her seminal book on survey experiments, Mutz (2011, p 110) briefly mentions non-compliance in field experimental research but then argues that “in population-based survey experiments, respondents may opt out of taking a survey altogether, but they cannot opt into or out of a specific experimental condition; thus, non-compliance is not a threat to internal validity.”

Of course, with the ever greater reliance on self-administered online opt-in panels, survey researchers too have become increasingly cognizant of problems relating to inattentive survey respondents and the consequences of this inattention for the attenuation of treatment effects (see e.g. Berinsky et al. 2014; Harden et al. 2019; Kane and Barabas 2019; Mutz 2021). At the same time, however, this literature has largely overlooked the problems that can arise when attention varies not only *across respondents*, but also *across different arms of the experiment*.¹ As we demonstrate below, such between-condition variability can generate imbalances in the “types of respondents” who attend the experimental manipulations, thereby posing a challenge for the analysis of placebo-controlled designs.

Importantly, while we urge scholars to do everything in their capacity to standardize treatments and ensure comparability in compliance across treatment arms, this article is not about how to improve compliance in survey experiments.² Instead, our aim is to offer general advice about what to do when respondents do not “take the treatment” regardless why non-compliance occurs. We view differential non-compliance as something that cannot be fully prevented, either because it is not always predictable, or because other important aspects of a research design im-

pose compromises. In particular, whenever scholars opt for “realistic” designs—for instance, by exposing subjects to actual videos or newspaper articles rather than artificial ones—they are, to some extent, deliberately giving up control over comparability since it is difficult, if not impossible, to find treatments and placebos in the real world that can be matched according to all plausible sources of variation. When this happens, a pure baseline could be the solution.

This paper proceeds as follows. Section 2 discusses placebo-controlled designs and introduces the problem of differential non-compliance across experimental arms. Section 3 illustrates these issues in the context of our own “failed” survey experiment. Section 4 then discusses potential solutions, focusing in particular on how IV can be used in placebo-controlled experiments if a pure control condition is included. Finally, Section 5 concludes with some practical advice for the broader survey research community.

2 Placebo-Controlled Designs and Non-Compliance

Placebo-controlled experiments take their name from medical trials where researchers randomly assign subjects to receive either an experimental drug or a placebo and then compare the outcomes of all those who complete the prescribed regime (Nickerson 2005). One key advantage of placebo-controlled designs relates to the issue of “compound treatments.” In the medical context, researchers might be worried that subjects are responding to a combination of the active properties contained in the “real” drug, plus the psychological effects of taking medication. These two effects would be confounded in a comparison of the treatment group against a “pure” control group that receives no pill from the researchers. On the other hand, the use of a placebo group allows researchers to account for these psychological effects and thereby isolate the active effects of the “real” drug.

As Nickerson (2005) notes, medical researchers analyze placebo-controlled experiments by comparing all subjects who complete the prescribed regime—that is, they compare the average outcomes for drug-takers to average outcomes for placebo-takers. But, of course, some subjects may *not* complete the prescribed regime; for example, they may forget to take either the “real drug” or the placebo. Thus, despite random assignment to treatment *conditions*, subjects can still determine their own treatment *status*. How then is a comparison of drug-takers to placebo-takers valid?

The validity of the comparison hinges on the assumption that *compliance profiles are independent of treatment assignment*. In other words, the “types of people” who forget to take the “real” drug are, on average, the same “types of people” who forget to take the placebo. Under this assumption, the group of drug-takers remains comparable to

¹ Kane and Barabas (2019) examine the passage rates of manipulation checks in each experimental arm of several studies, but the between-condition variability in passage rates was mostly minimal. Further, these authors do not explore how to handle situations when substantial between-condition variability does arise.

² In fact, our paper largely sidesteps the issue of why compliance differed across our video treatments in the first place. Admittedly, we remain in the dark about this, although we offer some conjectures in Section 3 and refer the reader to a specialized literature on how individuals interact with audio-visual stimuli (see e.g., Couper et al. 2004; Eifler and Petzold 2022).

the group of placebo-takers. Taking the difference-in-means across these two groups provides us with an unbiased estimate of the *complier average causal effect* (CACE)—i.e. the isolated effect of the drug on the subset of individuals who took their medication as instructed.³ On the other hand, if compliance profiles differ across experimental arms, then drug-takers and placebo-takers are no longer identical in expectation and a comparison of their outcomes would be biased by baseline differences between the two groups.

Given the importance of this assumption, it may be instructive to reflect on the situations where (non-)compliance patterns are plausibly similar across experimental arms. Medical trials provide a good example: since subjects are unaware of whether they have been given the “real drug” or the placebo, it is hard to imagine a scenario where forgetfulness should systematically differ between experimental groups. Another example is provided by recent field experiments on prejudice reduction through interpersonal conversations with door-to-door canvassers (Broockman and Kalla 2016; Kalla and Broockman 2020). Subjects were randomly assigned to have a conversation about either a marginalized group (unauthorized immigrants and transgender people) or a placebo issue (recycling). Since compliance was defined as simply being at home when the canvasser came knocking, and subjects did not know why someone was at their door, it is highly plausible that the same “types of people” were exposed to both conversations.

Nickerson (2005) cautions, however, that the assumption of *independent compliance profiles* may not hold in many social science applications. The problem arises because good placebos should have no “active ingredients”—i.e. they should have no independent effects on the outcome. Yet, such a placebo is unlikely to induce the same compliance profile as the treatment of interest. To take Nickerson (2005, p 247)’s example: “[S]uppose a researcher seeks to determine how political documentaries change pre-existing beliefs ... [A] randomly selected placebo group is set up to watch the romantic comedy *Shakespeare in Love*. While *Shakespeare in Love* is unlikely to change a subject’s political beliefs, the set of subjects who will sit through the romantic comedy and the political documentary are likely to be very different ...” In this example, a comparison of outcomes among people who actually watched the movie to which they were assigned would risk confounding the effect of political documentaries with baseline differences that may be correlated with people’s movie preferences.

So far, such issues have received scant attention amongst survey researchers because it is commonly assumed that compliance with survey treatments is high (Mutz 2011;

Porter and Velez 2022). However, as research on survey attentiveness has shown, there is no guarantee that respondents will actually engage with and assimilate the treatments to which they have been assigned (see e.g. Berinsky et al. 2014; Harden et al. 2019; Kane and Barabas 2019; Mutz 2021). Moreover, attentiveness may vary across experimental arms, giving rise to a situation akin to the differential non-compliance described by Nickerson (2005). The next section illustrates these problems in the context of our own “failed” experiment.

3 Our Survey Experiment

Our initial research question—which, importantly, is *not* the focus of this article—concerned how racial minorities in the UK are perceived by white survey respondents.⁴ More specifically, would white respondents focus on minorities’ race or their nationality? Drawing on seminal theories of group categorization in social psychology (e.g., Cialdini et al. 1976), we hypothesized that respondents would be more likely to apply a racial—as opposed to national—lens when minorities were perceived as under-performing.

To test this hypothesis, we conducted a survey experiment with an online-access panel involving approximately 3000 “white British” respondents. Respondents were recruited via the survey firm Dynata between November - December 2022. We implemented quotas for age, gender and education based on marginal distributions from the UK census. Table 1 provides summary statistics of the sample.

We situated our experiment in the domain of international football, where it was easy to signal both a player’s citizenship and ethno-racial background. To manipulate perceptions of minority under-performance, we created 2–3 minute videos of the complete penalty shootouts from the 2012 and 2020 European Championships. In both of these matches, the English national team was eliminated because *all of the non-white players*, and none of the white players, missed their penalties. These two videos thus constituted our treatments of interest. We also created two placebo treatments featuring videos of the penalty shootouts from the 1996 European Championship and 2006 World Cup. In these placebo matches, only white players had taken and missed shots, resulting in an England loss.⁵

Our survey began with a few introductory questions about respondent demographics, national identity, and interest in sport. Respondents were then randomly assigned to watch one of the four videos, resulting in a between-

³ Importantly, this population of compliers cannot be assumed to be “representative” of the overall study population (Gerber and Green 2012).

⁴ Details on the motivation behind our initial research can be found in the Appendix.

⁵ Videos, data and replication code can be found at <https://osf.io/huyxz/files>.

Table 1*Summary Statistics of Survey Respondents*

Characteristic	<i>n</i>	% Sample
Male	1460	48
Female	1541	51
Non-binary	17	1
Age:		
18–24	307	10
25–34	438	15
35–44	467	15
45–54	551	18
55–64	526	17
65+	729	24
Degree:		
Yes	967	32
No	2051	68
Identify as English:		
Very strongly	2100	70
Fairly strongly	679	22
Not very strongly	195	6
Not at all	44	1
Interest in football:		
Huge fan	644	21
Regular fan	581	19
Irregular fan	280	9
Tournaments only	662	22
No interest	511	17
Hate football	340	11
Party Id:		
Conservative	807	27
Labour	911	30
LibDem	307	10
UKIP, Reform, BNP	148	5
Other	161	5
DK, no vote	684	23
Total observations	3018	100

Percentages may not sum to 100 because of rounding.

subjects design.⁶ Importantly, we did not include a pure control condition where respondents were shown no video. The videos were immediately followed by a manipulation check in which respondents were asked to identify the

players who missed their penalties from a row of headshots.⁷ Finally, respondents undertook a categorization task to measure their usage of national vs. racial classification schemes.

For the purposes of this article, we will focus on respondents' performance in the *manipulation check* while reserving description of the full experiment for the Appendix. We interpret passage of this manipulation check as our measure of treatment compliance. The idea is that, if respondents cannot correctly pick out the players who missed their penalties, then they did not assimilate the intended information about the racial background of the "under-performers." In other words, respondents who failed the manipulation check did not "take their treatments."

To support this interpretation, we first show that passage of the manipulation check is correlated with two additional measures of respondent attentiveness: (1) the proportion of respondents who navigated away from the survey page during the video, and (2) the share of the video which was actually streamed. As shown in Fig. 1, respondents who failed the manipulation check also performed worse on both metrics.

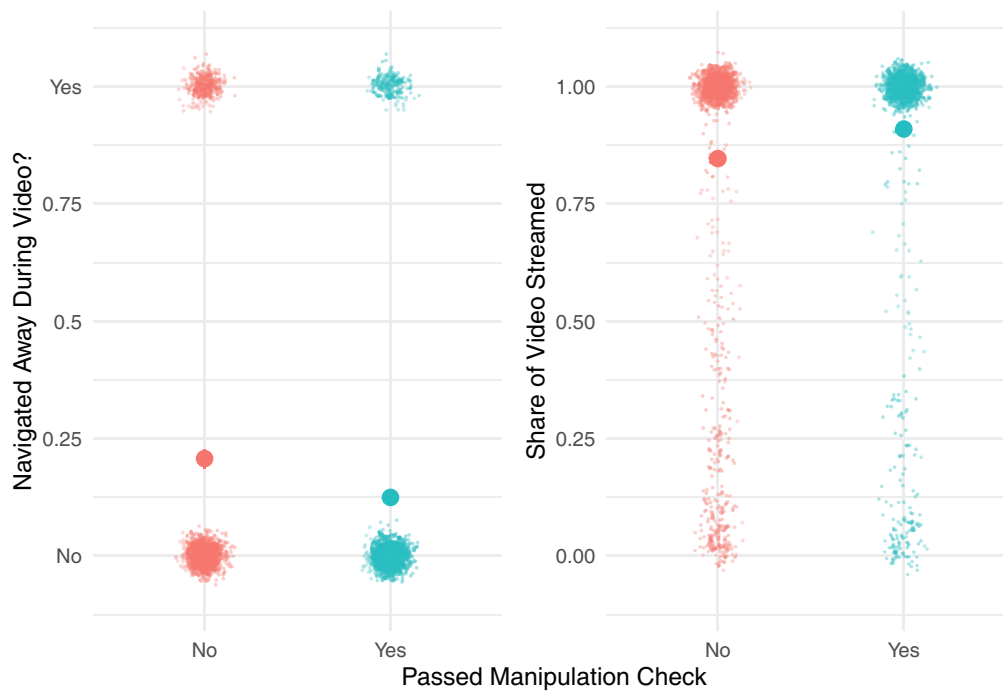
We next turn to an analysis of the manipulation check across treatments. Our outcome here is a dummy variable indicating whether the respondent successfully picked out the players from the shootout video who missed their penalties. The results are presented in Fig. 2. Note that if compliance profiles are indeed independent across treatments, we would expect to observe similar rates of passage in each treatment arm. This is analogous to the medical context where, if the same "types of people" forget to take their medication independently of experimental arm to which they have been assigned, then we should expect similar rates of non-compliance in the placebo and treatment arms. However, compliance rates in Fig. 2 vary widely: about 60% of respondents passed the manipulation when shown the 1996 and 2012 videos, compared to only about 40% in the 2006 and 2020 videos. We therefore have some first evidence of differential non-compliance.

Figure 3 provides further evidence that compliance profiles are not identical across experimental arms. Here, we present standardized mean differences⁸ for several pre-treatment covariates across treatment arms for both (i) the full

⁷ We thus employ what Kane and Barabas (2019) term a "factual manipulation check."

⁸ This is the difference-in-means across treatment arms divided by the standard deviation of the variable in the overall sample. Following Ho et al. (2007) and Imai et al. (2008), we opt for the standardized mean difference as opposed to more traditional balance tests because (i) balance is a property of the sample, not of a population from which the sample was drawn, and (ii) the power of balance tests depends on the sample size, which obviously changes as we move from the full sample to the subsample of "compliers."

⁶ We opted for two treatments and two placebos to ensure that our results were not sensitive to the particulars of any one video itself.

**Fig. 1**

The figure displays correlations between passing the manipulation check and two alternative measures of respondent attention: (1) an indicator for whether the respondent navigated away from the survey page during the video (left panel), and (2) the share of the video which was actually streamed (right panel).

sample, and (ii) the subsample of respondents who passed the manipulation check. We examine demographic characteristics such as gender, age and educational attainment, as well as partisan political preferences, national identity, and interest in football. The variables are described in Table 1. In many cases, we observe that focusing on respondents who pass the manipulation check results in greater imbalance, with differences-in-means across treatment arms surpassing the customary threshold of 0.1 standard deviations of the variable in the overall sample. Imbalances are particularly evident for variables related to political preferences, national identity and interest in football.

These imbalances are problematic since they indicate that the individuals in the treatment and placebo arms *who also received their intended treatments* are no longer “the same” in expectation. And although we might control for *observed* imbalances, we cannot rule out further sources of omitted variable bias from covariates *that we did not measure*. More broadly, the larger problem is conceptual: since respondents themselves choose whether or not to pay attention to the videos, subsetting the sample according to the manipulation check allows self-selection to creep into the experiment.

These patterns of selective non-compliance or inattention may reflect a combination of factors which we are unable to disentangle. Admittedly, since we used actual broadcast footage of the penalty shootouts, we were unable to perfectly standardize video quality, length, etc. It is plausible these factors influenced respondents’ engagement with the video stimuli (Couper et al. 2004; Eifler and Petzold 2022). Neither could we standardize the number of players who took (and missed) penalties in each match, which obviously affected the difficulty of the manipulation checks.

While acknowledging these trade-offs, however, we nonetheless chose to use “real” videos because we thought the benefits of “natural treatments” outweighed the costs of loss of control. In addition, we were initially relatively unconcerned about differential non-compliance: after all, our videos were all about the same length, showed the same content featuring penalty shootouts in the knockout phases of major international football tournaments, and ended with an England loss. We therefore did not anticipate that they would also elicit such different patterns of inattention.

Unfortunately, we were wrong. However, this crucial fact motivates our discussion below in which we try to reason through how we might have “rescued” our experiment while retaining natural, albeit non-standardized, video treatments.

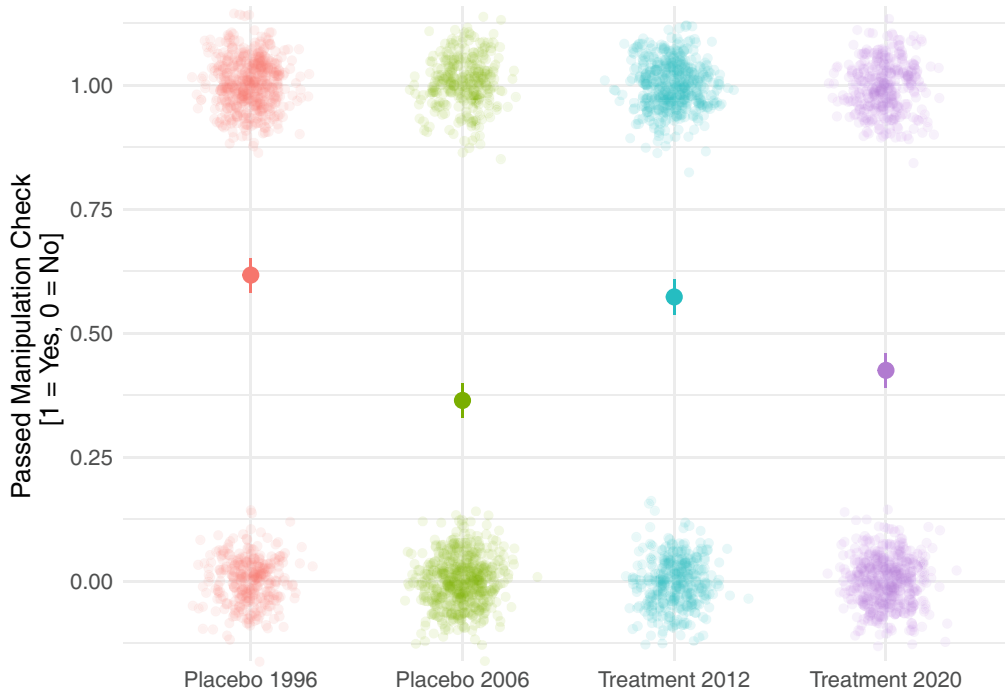


Fig. 2

The figure displays the fraction of respondents who passed the manipulation check in each treatment arm. We observe that this fraction differs substantially across videos, indicating differential non-compliance.

We intend this discussion to be particularly useful for researchers navigating trade-offs similar to the ones we faced.

4 Potential Solutions to Non-compliance

One obvious solution for researchers in our situation would be to simply ignore the manipulation check and instead estimate an intent-to-treat (ITT) effect by comparing outcomes across *assigned* experimental arms. Certainly, the ITT would be unbiased. Further, in field experimental studies with non-compliance such as Gerber et al. (2008)’s “get-out-the-vote” experiments, the ITT may actually be a substantive quantity of interest since some amount of “failure to treat” is simply “built in” to how the intervention works in the real world. By contrast, our purpose here was to test a specific hypothesis about the effects of changing a theoretical quantity of interest—namely, the perception of non-white players’ poor performance. Given these research goals, estimating the effect of *treatment assignment* misses the mark; we need to know the effect of *actually receiving treatment*.

In lieu of the ITT, we advocate instead for an instrumental variables (IV) approach where *treatment assignment* is used as an instrument for *treatment status* (Gerber and

Green 2012). Importantly, such strategies recover the causal effect of actually *taking the treatment*, albeit only for the subgroup of “compliers.” However, this approach requires the inclusion of a pure control condition which is unfortunately absent from our own experiment. Nonetheless, we can illustrate the logic via a simplified example.

Suppose we had only a single treatment group plus a pure control group. In this case, the average outcomes in the group assigned to treatment (Y^T) would be given by the formula:

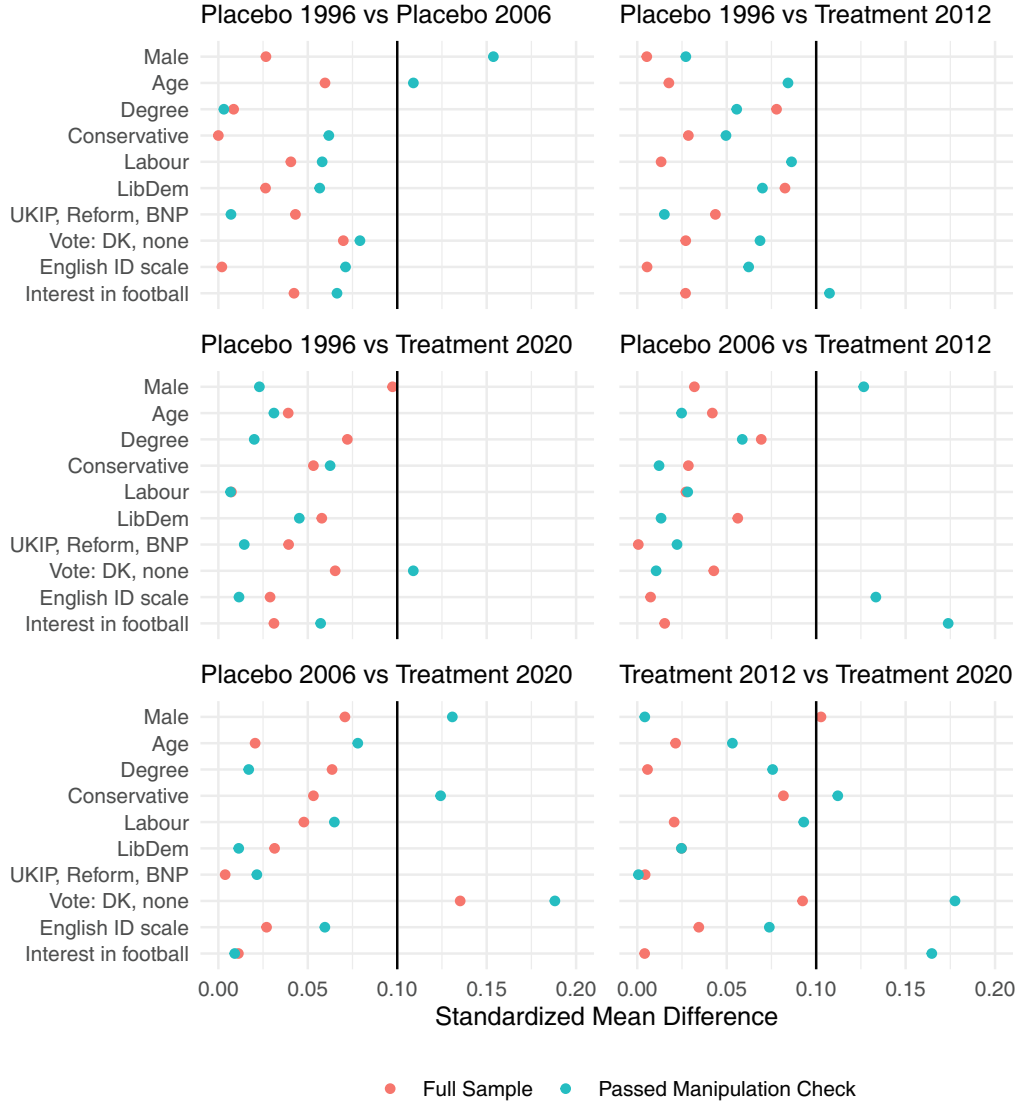
$$Y^T = \pi_c^T(B_c^T + \Delta) + \pi_{nc}^T(B_{nc}^T) \quad (1)$$

where π_c^T and π_{nc}^T respectively denote the proportion of compliers and non-compliers in the treatment group, B_c^T and B_{nc}^T represent the baseline outcomes amongst compliers and non-compliers in the treatment group, and Δ denotes the CACE.

The average of the pure control group outcomes (Y^P) is given by:

$$Y^P = \pi_c^P(B_c^P) + \pi_{nc}^P(B_{nc}^P) \quad (2)$$

Since we observe compliance behavior in the group assigned to treatment, we can calculate π_c^T and π_{nc}^T . Further,

**Fig. 3**

The figure displays standardized mean differences for demographic and attitudinal variables across treatment arms, broken down by respondents in the full sample versus the subsample of respondents who passed the manipulation check. Variable codings are provided in Table 1. We observe greater imbalance (surpassing the customary 0.1 threshold) among the subset of respondents who passed the manipulation check, particularly on attitudinal variables related to political preferences, national identity and interest in football.

random assignment assures us that (i) the same proportions of compliers and non-compliers are present in the control group—i.e. $\mathbb{E}(\pi_c^T) = \mathbb{E}(\pi_c^P)$ and $\mathbb{E}(\pi_{nc}^T) = \mathbb{E}(\pi_{nc}^P)$ —and (ii) compliers and non-compliers in the treatment arm will have similar baseline characteristics as their respective counterparts in the control arm—i.e. $\mathbb{E}(B_c^T) = \mathbb{E}(B_c^P)$ and $\mathbb{E}(B_{nc}^T) = \mathbb{E}(B_{nc}^P)$. We can therefore drop the superscripts ^T

and ^P from Eqs. (1) and (2). Re-arranging our two equations yields the well-known Wald estimator for the CACE Δ):

$$\Delta = \frac{Y^T - Y^P}{\pi_c} = \frac{ITT}{\text{Compliance Rate}} \quad (3)$$

It is crucial to note however that this simple solution is not available in a placebo-controlled design with *no pure*

control. Under this scenario, the average of the placebo group outcomes (denoted Y^*) is:

$$Y^* = \pi_c^*(B_c^* + \delta) + \pi_{nc}^*(B_{nc}^*) \quad (4)$$

where δ denotes the placebo effect amongst compliers in the placebo group. Ideally, Δ minus δ yields the “isolated” treatment effect of interest (which we denote as α). It is important to remember, however, that Δ is defined with respect to π_c^T —i.e. people who actually take the substantive treatment of interest when assigned to do so—whereas δ is defined with respect to π_c^* —i.e. people who actually take the placebo when assigned to do so. Unless one is willing to assume that these are more or less the same “types” of people, one cannot simply substitute $\Delta = \delta + \alpha$ into Eq. 1 above and solve for α .

Instead, if the experiment contains *both* a pure control and a placebo arm, it is possible to separately estimate Δ and δ using Eq. 3. In this case, one would have two distinct CACEs: Δ for the treatment of substantive interest, and δ for the placebo. Evidence supporting the hypothesized treatment effect of interest could be adduced if Δ were large and statistically significant and δ were small and statistically insignificant.⁹

Unfortunately, this strategy was not available to us because we neglected to include a pure control condition in our design. Consequently, the importance of having a pure control condition constitutes the key “lesson learned” from our failed experiment.

5 General Discussion

Compliance issues compromise what we can learn from randomized experiments. When subjects are able to select out of or into treatment, then the individuals who are actually treated are likely to differ from people who are not, thereby destroying the balance created by random assignment. In survey experiments, non-compliance comes in the form of inattention to experimental stimuli. While it is generally believed that most respondents in survey experiments do indeed “take the treatment,” we demonstrated via our own experiment that “failures to treat” do occur, and moreover, at different rates in different experimental arms.

What should researchers do in such cases? In hindsight, we should obviously have subjected our video treatments to more extensive pretesting. That said, we are admittedly not experts in human-video interactions and, we suspect, nei-

ther are most applied researchers. The fact that we remain “in the dark” about how people engaged with our video manipulations does not invalidate the fact that there are still lessons to be drawn from our experience. Moreover, given our preference for naturalistic treatments, there exist some issues that no amount of pre-testing could have resolved: for example, we could not hold constant the number of players who missed their penalty shots across the four videos.

For the purposes of this paper, we view these limitations as “features” rather than “bugs.” In fact, our reflections below are applicable precisely in cases where researchers *have little theoretical knowledge about how treatments are cognitively processed and/or cannot perfectly calibrate their treatments*. In these instances, our advice can still help to “rescue” placebo-controlled experiments with differential non-compliance.

First, it is vital that researchers measure whether respondents actually “took the treatment” *in each of the experimental arms*. In other words, it is important to confirm that “experimental stimuli alter a latent independent variable of interest” (Kane and Barabas 2019, p 234). Treatments that do not work due to respondent inattention will lead to underestimates of the “true” treatment effect (Kane 2025).

Second, if researchers detect non-compliance with treatment assignment, instrumental variables methods can be used to estimate the CACE *provided that the design includes a pure control condition*. This consideration is especially relevant to placebo-controlled experiments in which researchers may be tempted to use the placebo as a sole control. In such cases, however, a comparison between treatment-takers and placebo-takers is invalid if compliance profiles are different across experimental arms. For this reason, a pure control condition serves as an important “backup” because it allows for the estimation of separate CACEs for each of the treatment and placebo arms.

Acknowledgements We are grateful to Lena Hipp, Sebastian Wenz and two anonymous reviewers for helpful comments on prior drafts of this article.

References

- Angrist, J., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444–455. <https://doi.org/10.2307/2291634>
- Berinsky, A., Margolis, M., & Sances, M. (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. <https://doi.org/10.1111/ajps.12081>

⁹ Even here, however, we would caution against directly comparing the CACEs since they are estimated with respect to two different complier populations. Comparisons may therefore reflect both α as well as baseline differences ($B_c^T - B_c^*$).

- Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, 352(6282), 220–224. <https://doi.org/10.1126/science.aad9713>
- Broockman, D., Kalla, J., & Sekhon, J. (2017). The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs. *Political Analysis*, 25(4), 435–464. <https://doi.org/10.1017/pan.2017.27>
- Cialdini, R., Borden, R., Thorne, A., Walker, M.R., Freeman, S., & Sloan, L.R. (1976). Basking in reflected glory: Three (football) field studies. *Journal of Personality and Social Psychology*, 34(3), 366–375. <https://doi.org/10.1037/0022-3514.34.3.366>
- Couper, M., Tourangeau, R., & Kenyon, K. (2004). Picture this! Exploring visual effects in web surveys. *Public Opinion Quarterly*, 68(2), 255–266. <https://doi.org/10.1093/poq/nfh013>
- Eifler, S., & Petzold, K. (2022). Fear of the dark? A systematic comparison of written vignettes and photo vignettes in a factorial survey experiment on fear of crime. *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology*, 16(2), 201–234. <https://doi.org/10.12758/mda.2022.01>
- Gerber, A., & Green, D. (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York, NY: W.W. Norton & Company.
- Gerber, A., Green, D., & Larimer, C. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, 102(1), 33–48. <https://doi.org/10.1017/S000305540808009X>
- Harden, J., Sokhey, A., & Runge, K. (2019). Accounting for noncompliance in survey experiments. *Journal of Experimental Political Science*, 6(3), 199–202. <https://doi.org/10.1017/XPS.2019.13>
- Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3), 199–236. <https://doi.org/10.1093/pan/mpi013>
- Imai, K., King, G., & Stuart, E. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 171(2), 481–502. <https://doi.org/10.1111/j.1467-985X.2007.00527.x>
- Kalla, J., & Broockman, D. (2020). Reducing exclusionary attitudes through interpersonal conversation: Evidence from three field experiments. *American Political Science Review*, 114(2), 410–425. <https://doi.org/10.1017/S0003055419000923>
- Kane, J. (2025). More than meets the ITT: a guide for anticipating and investigating nonsignificant results in survey experiments. *Journal of Experimental Political Science*, 12(1), 110–125. <https://doi.org/10.1017/XPS.2024.1>
- Kane, J., & Barabas, J. (2019). No harm in checking: Using factual manipulation checks to assess attentiveness in experiments. *American Journal of Political Science*, 63(1), 234–249. <https://doi.org/10.1111/ajps.12396>
- Mutz, D. (2011). *Population-based Survey Experiments*. Princeton, NJ: Princeton University Press.
- Mutz, D. (2021). Improving experimental treatments in political science. In J. Druckman & D. Green (Eds.), *Advances in Experimental Political Science* (pp. 219–238). Cambridge, UK: Cambridge University Press.
- Nickerson, D. (2005). Scalable protocols offer efficient design for field experiments. *Political Analysis*, 13(3), 233–252. <https://doi.org/10.1093/pan/mpi015>
- Porter, E., & Velez, Y. (2022). Placebo selection in survey experiments: An agnostic approach. *Political Analysis*, 30(4), 481–494. <https://doi.org/10.1017/pan.2021.16>