


RESEARCH ARTICLE

# More than meets the ITT: A guide for anticipating and investigating nonsignificant results in survey experiments

John V. Kane 

New York University, New York, NY, USA

Email: [John.Kane@nyu.edu](mailto:John.Kane@nyu.edu)

## Abstract

Survey experiments often yield intention-to-treat effects that are either statistically and/or practically “non-significant.” There has been a commendable shift toward publishing such results, either to avoid the “file drawer problem” and/or to encourage studies that conclude in favor of the null hypothesis. But how can researchers more confidently adjudicate between true, versus erroneous, nonsignificant results? Guidance on this critically important question has yet to be synthesized into a single, comprehensive text. The present essay therefore highlights seven “alternative explanations” that can lead to (erroneous) nonsignificant findings. It details how researchers can more rigorously anticipate and investigate these alternative explanations in the design and analysis stages of their studies, and also offers recommendations for subsequent studies. Researchers are thus provided with a set of strategies for better designing their experiments, and more thoroughly investigating their survey-experimental data, before concluding that a given result is indicative of “no significant effect.”

**Keywords:** Surveys; experiments; null results; significance; bias

Survey experiments are an increasingly popular method for testing whether particular types of information and stimuli can causally affect politically relevant beliefs, attitudes, and behaviors (e.g., Druckman and Green 2021; Druckman 2022; Mutz 2011). However, little scholarship has attempted to concisely detail and address the various factors that can often *undermine* a researcher’s survey experiment – i.e., yield what appears to be a “non-significant” result (whether in the statistical and/or practical sense) despite a hypothesis actually being correct.

How can researchers be more confident that a nonsignificant result is actually indicative of “no effect” vis-à-vis a consequence of one or more of these undermining factors? An inability to address this crucial question stands to greatly diminish the theoretical and empirical value of one’s study. Thus, being able to rigorously investigate nonsignificant results is valuable, especially given the growing awareness of

the “file drawer problem” within scholarly research (Alrababa’h et al. 2022; Franco, Malhotra, and Simonovits 2014), as well as greater scholarly appreciation for nonsignificant results that occur in well-designed experiments and/or studies that conclude in favor of the null hypothesis (see Chambers and Tzavella 2022; Journal of Experimental Political Science 2023; The Journal of Politics 2022; Nature 2023).

A typical survey experiment randomly assigns respondents to one of at least two conditions within a survey. The condition to which one is assigned represents one value of the independent variable,  $X$ . Researchers then statistically test whether values of  $X$  are significantly associated with the values of an outcome ( $Y$ ) that is measured for all respondents. When conducted for the entire sample, researchers refer to this difference as an estimate of the intention-to-treat (ITT) effect (Gerber and Green 2012, 142).<sup>1</sup>

When is an ITT nonsignificant? Per the null hypothesis significance-testing (NHST) paradigm (e.g., see Gill 1999), researchers fail to reject the null hypothesis – i.e., deem a result “not statistically significant” – when a  $p$ -value exceeds a particular threshold (e.g.,  $\alpha = 0.05$ ). Apart from its *statistical* significance, an estimated effect can also be so small as to be *practically* (or “substantively”) nonsignificant (Rainey 2014).

Suppose a researcher fields a survey experiment and the ITT is nonsignificant (i.e., not statistically discernible from zero and/or substantively negligible in size). The researcher may infer that the hypothesis being tested, and/or underlying theory, is incorrect. This is one explanation for the result. However, there exist *alternative explanations* (AEs) that researchers should consider before concluding a treatment to be truly nonsignificant. Even if one disregards significance-testing and focuses only on the *range* of likely effect sizes (e.g., confidence intervals or credible sets see Gill 1999, 662–63; Rainey 2014), understanding these AEs is vital as they tend to lower the center of that range toward zero (i.e., toward practical nonsignificance). In short, the AEs identified here can undermine both hypothesis testing *as well as* point estimation of treatment effects, making it difficult for researchers to determine what can be learned from a nonsignificant result.

The key purpose of this essay, then, is to detail the variety of AEs that increase the likelihood of nonsignificant results in survey experiments. Knowledge of these AEs is relevant to researchers at two key stages of a research project: (1) when designing their survey experiments, and (2) when investigating the robustness of their nonsignificant findings after data collection. Ultimately, with this knowledge, researchers can minimize the threat of AEs before data are collected and better detect erroneously nonsignificant findings once data are in hand. In addition, researchers concluding *in favor of* a treatment having “no significant effect” can offer far more persuasive evidence by showing not only a nonsignificant ITT but also that these various AEs can be ruled out reasonably confidently.<sup>2,3</sup> Finally, this

<sup>1</sup>The ITT differs from the commonly-employed *average treatment effect* (ATE) insofar as the latter implicitly assumes full compliance (Harden et al. 2019, 200).

<sup>2</sup>Naturally, as this involves somewhat subjective determinations, motivated reasoning is a potential concern (e.g., in deciding whether a particular piece of evidence permits “a reasonable degree of confidence”). Nevertheless, *ceteris paribus*, being able to present more evidence in favor of a nonsignificant ITT is preferable to presenting less.

<sup>3</sup>It is worth emphasizing a common criticism of the NHST paradigm, which is that meaningfully large ITT estimates are often dismissed as “null” because the  $p$ -value was not below  $\alpha$ . In other words, the null

essay provides guidance for how researchers might design a subsequent study should they find evidence for one or more AEs in their initial study.

Importantly, the aim is not to provide experimentalists with more avenues to “find significant results,” raising the risk of Type I errors (“false positives”) in the process. To guard against this possibility, researchers should pre-register their studies and analysis plans, particularly the analyses they would conduct to detect and address the various AEs identified here (Blair et al. 2019; Druckman 2022, 143–44; Nosek et al. 2018).

This concern notwithstanding, given the great deal of effort that goes into designing and fielding a survey experiment, and the prevalence of obtaining “nonsignificant results,” knowledge of how to anticipate, investigate, and more confidently rule out these AEs can assist researchers in getting the most of out of their studies.

## Seven alternative explanations (AEs) for nonsignificant results

### **Alternative explanation #1: Respondent inattentiveness**

The more that a sample is inattentive to a treatment, the more that a treatment group’s experience resembles that of the control group. To the extent this occurs, we should expect a smaller difference in  $Y$  between the two groups. Because of respondent inattentiveness, therefore, a treatment effect (if one truly exists) will tend to be biased toward zero, making the ITT more difficult to precisely estimate, reliably detect, and attain statistical significance (given a particular sample size and level of  $\alpha$  (Bailey 2021)). By the same logic, inattentiveness will increase the likelihood of observing a failed manipulation (see AE #2 below). In short, attentiveness is often a *precondition* for having an efficacious treatment.

Researchers should therefore plan to investigate inattentiveness in their sample. There exists a variety of ways to conduct such an investigation using specialized “checks.” *Instructional manipulation checks* (i.e., “screeners” Berinsky, Margolis, and Sances 2014; Oppenheimer, Meyvis, and Davidenko 2009) and *factual manipulation checks* (Kane and Barabas 2019), for example, are questions with only one correct answer. The latter asks respondents about specific content that was manipulated across experimental conditions. Incorrect answers to such questions indicate insufficient respondent attentiveness. Overly fast response times (per question timers) are an additional method for gauging respondent inattentiveness.<sup>4</sup>

Existing studies using these techniques find substantial rates of inattentiveness, with estimates often ranging from approximately 15% to 40% (e.g., see Aronow et al. 2020; Kane and Barabas 2019). Importantly, there exists no agreed-upon *acceptable* level of inattentiveness in survey experiments. Such a level would depend greatly upon other aspects of the study (e.g., for NHST, sample size and treatment strength

---

hypothesis is “accepted” rather than merely “not rejected” (Gill 1999; see also Hartman and Hidalgo 2018). To more persuasively argue that there is no meaningful effect, one would ideally want to show a negligibly-sized ITT that is precisely estimated and/or (using a two one-sided test approach) that an ITT is smaller than – and statistically distinguishable from – the smallest substantively meaningful effect (e.g., see Rainey 2014).

<sup>4</sup>See the Supplemental Appendix A for extended discussion of attentiveness measures.

would also matter).<sup>5</sup> Nevertheless, the crucial point is that *any* amount of inattentiveness will tend to attenuate ITT estimates.

If substantial inattentiveness is found, what can be done? Researchers should first confirm whether treatment assignment substantially covaries with an item measuring attention to the treatment (Kane & Barabas 2019). This enables the researcher to more confidently conclude that respondents in one condition attended (on average) to *different* information than respondents in another condition – a crucial assumption of most survey experiments.<sup>6</sup>

Researchers can also employ *pre-treatment* attention checks and use these to analyze how treatment effects differ across varying levels of attentiveness (Druckman 2022, 56). Such checks enable the researcher to test whether – as one would expect if a treatment is truly effective – treatment effects are substantially larger for subsamples that were more likely to have attended to the treatment, and without the risk of post-treatment bias (see Kane, Velez, and Barabas 2023).<sup>7,8</sup> Critically, researchers should avoid using any *post-treatment* measure of attentiveness to re-estimate treatment effects (e.g., dropping respondents who fail post-treatment attention checks, or (per a timer) rushed through an experimental vignette). This practice has been shown to risk undermining random assignment and inducing statistical bias (Aronow, Baron, and Pinson 2019; Montgomery, Nyhan, and Torres 2018; Varaine 2022).

### ***Alternative explanation #2: Failure to vary the independent variable of interest***

A second potential AE is that one's independent variable was not actually varied by the treatment (Mutz and Pemantle 2015). For example, suppose that a researcher attempts to increase respondents' belief that a military draft is likely to be reinstated (to study if it affects respondents' support for war (Y)).<sup>9</sup> Whether the treatment *actually* accomplishes this objective is, of course, an empirical question.

The key recommendation is to include a classic manipulation check (Mutz 2021).<sup>10</sup> This is a survey item that (1) is asked after exposure to a control/treatment condition, and (2) measures the underlying construct that the researcher is attempting to manipulate. As in the above example, if a treatment is designed to make respondents perceive that a military draft is more likely, then the manipulation check should measure respondents' perceived likelihood of a draft; if a treatment is designed to make respondents feel more anxious, the manipulation check should measure anxiety, etc.

<sup>5</sup>These aspects can also matter in different ways. A larger sample and/or fewer conditions may still permit the detection of statistically significant effects even in the presence of high inattentiveness – however, the inattentiveness is still likely to bias the ITT downward.

<sup>6</sup>Importantly, this type of analysis is not possible for other attention check types, the answers to which do not depend upon treatment assignment.

<sup>7</sup>Researchers should report the extent to which the attentive subsample compositionally differs from the original sample.

<sup>8</sup>See Supplemental Appendix B for additional discussion.

<sup>9</sup>This example comes from Horwitz and Levendusky (2011).

<sup>10</sup>Kane and Barabas (2019) specifically refer to these as *subjective* manipulation checks as there are no correct/incorrect answers (unlike other types of manipulation checks).

Empirically, a researcher could then confirm that a treatment group significantly differs from the control group on this measure and, ideally, to a substantively large degree. When this occurs, it is additional evidence that the treatment accomplished what the researcher intended. It also indicates that, whatever the sample's level of inattentiveness (see *AE #1*), it was not substantial enough to prevent the researcher from successfully manipulating the independent variable of interest.<sup>11</sup>

As Mutz (2011, 84–85) argues, instances in which researchers should consider such manipulation checks “optional” are “relatively few and far between.” Yet, in spite of their simplicity and enormous value, such manipulation checks remain remarkably under-utilized: well under 50% of published experimental studies in political science feature a manipulation check (Kane and Barabas 2019; Mutz 2021).

What if treatment assignment is *not* significantly associated with the manipulation check (i.e., the manipulation “failed”)? Here, there exist several possible explanations. First, it may be a result of the inattentiveness problem described above: if respondents are not attentive to the treatment, we should expect an attenuated effect on the manipulation check (just as we expect an attenuated effect on *Y*).

If attentiveness levels are reasonably high (and, ideally, treatment assignment is found to correlate with a measure of attention to the treatment), a second possibility is that the manipulation check measure is flawed with respect to either its validity and/or reliability. In other words, we need to be confident that the manipulation check is a reasonably valid measure of the independent variable we intend to manipulate and also that it is not an overly “noisy” measure. These concerns can be investigated by testing whether theoretically relevant variables (e.g., education, age, and political attitudes) significantly correlate with the manipulation check measure. That is, we can investigate the manipulation check's *criterion validity* (e.g., Druckman 2022, 22–27). If substantial correlations are found, it suggests that the check is to some extent valid and reliable and, thus, should in principle be manipulable.

The third, more fundamental possibility is that the treatment is not actually manipulating what it is designed to manipulate. As an extreme example, imagine a single sentence of manipulated (i.e., treatment) material contained within several paragraphs of non-manipulated text, images, etc. Whatever this treatment might affect, its efficacy is potentially being overwhelmed and attenuated by the other information that respondents are being asked to process (e.g., see Mutz 2011, 58), even if this additional content is contextually relevant (see Brutger et al. 2023). In short, this treatment is not salient enough to induce variance in the independent variable. As a result, the treatment will not be significantly predictive of the manipulation check, nor is it likely, therefore, to predict *Y* (see Brophy and Mullinix 2023 for an applied example).

Being able to rule out one or more of these possibilities allows the researcher to better understand both their nonsignificant manipulation check result and, more broadly, their nonsignificant ITT result (see Supplemental Appendix C for additional details).

---

<sup>11</sup>Crucially, this kind of manipulation check *assumes* – rather than tests for – sufficient attentiveness. Thus, researchers can also conduct this check across different levels of attentiveness (measured pre-treatment) to determine whether the treatment significantly predicts manipulation check responses among the most attentive.

Like *AE #1*, this AE has important implications for the *design stage* of one's study. In addition to including a manipulation check, researchers must think carefully about what one's treatment actually involves, ideally pretesting alternative versions to determine which is most strongly associated with the manipulation check item (Chong and Junn 2011, 329–30; Mutz 2011, 65). Searles and Mattes (2015), for example, test several anger-induction techniques in one sample and then, based on the manipulation check results, use the best-performing technique for their subsequent study.

### **Alternative explanation #3: Pre-treated respondents**

A related AE is a “pre-treatment effect” (see Druckman and Leeper 2012; Gaines, Kuklinski, and Quirk 2007). Specifically, a treatment may be efficacious, but perhaps respondents have been treated *prior to the study*, in the real world, with information similar to what the researcher is employing in the experiment.

For example, at the height of the COVID-19 pandemic, a survey experiment that randomly assigned respondents to read information that the coronavirus is dangerous to one's health may have been thwarted by a pre-treatment effect: this information – though powerful – would have undoubtedly been absorbed by respondents *prior to* the experiment. Thus, the treatment may appear to have a nonsignificant effect on *Y*, not because the treatment is ineffective, but because it has already occurred. As Slothuus (2016, 303) writes of the effect of party cues, “paradoxically, experimenters will be most likely to find no relationship at the very time that the relationship is strongest outside the experimental context.” Indeed, pre-treatment will tend to bias treatment effects toward zero (see Gaines, Kuklinski, and Quirk 2007) and thus increase the likelihood of a nonsignificant result.<sup>12</sup>

Importantly, a pre-treatment effect should therefore *also* tend to result in a failed manipulation check. In the presence of nonsignificant ITT estimates and a failed manipulation check, therefore, researchers should carefully consider their treatment vis-à-vis what respondents may have been already exposed to in the real world prior to the experiment.

Again, this has important implications for how researchers design their survey experiments. The likelihood of pre-treatment will, naturally, depend upon what the researcher employs as treatment stimuli. In contexts wherein a pre-treatment effect is more likely, researchers might err on the side of having a relatively *stronger* treatment to compensate for the attenuating effects of pre-treatment. This should provide a better test of the hypothesis insofar as the experimental treatment is more powerful than what respondents have already been exposed to (though it may potentially diminish the external validity of the stimulus (see Druckman 2022, Ch.3)). Additionally, researchers can include a (pre-treatment) survey question that gauges whether substantial pre-treatment has occurred (e.g., asking how closely one has been following a particular topic in the news), and then estimate the ITT among those who are less likely to have been pre-treated (e.g., see Linos and Twist 2018). In the extreme case wherein *all* respondents are likely to have been heavily pre-treated,

<sup>12</sup>Notably, however, Linos and Twist (2018) find that pre-treatment can lead to *overestimation* of an effect when a treatment runs counter to information absorbed prior to the experiment.

researchers might consider an alternative research design to test the hypothesis of interest or postpone the study until the threat of pre-treatment subsidies.

#### **Alternative explanation #4: Insufficient statistical power**

Within the NHST paradigm, insufficient power is a well-established reason for statistically nonsignificant (i.e., “null”) results (Alrababa’h et al. 2022). Nevertheless, political science research remains severely underpowered (Arel-Bundock et al. 2022). A small sample – given the number of experimental groups and anticipated magnitude of the treatment(s) – is a common means by which null results become more likely in survey experiments, even when a treatment truly has an effect. In the extreme case, too small a sample will yield null findings no matter how efficacious one’s treatment is (e.g., see Perugini, Gallucci, and Costantini 2018).<sup>13</sup>

The most direct approach for guarding against this AE is to have a sufficient sample size. When designing their survey experiment, researchers should conduct statistical power analyses, and assume the *smallest* substantively meaningful effect size, to determine an appropriate sample size for the study (Lakens 2022). In so doing, researchers should also be mindful of (1) the possibility of substantial inattentiveness (AE #1), (2) the number of experimental conditions, and (3) whether any subgroup analyses will be performed, adjusting their power analyses accordingly.<sup>14</sup> A sample size of 1000, for example, may *seem* sufficiently powered. However, if the experiment involves five conditions, and will involve subsetting the data on racial identification, for example, the researcher could ultimately be estimating treatment effects among only a few dozen respondents (and only a fraction of these respondents will have actually attended to the experiment).

Thus, while researchers may have limited control over the total sample size (e.g., because of resource constraints), greater discretion can be exercised over (1) the number of experimental conditions (and whether some of these conditions can potentially be “collapsed” together because of a common element between them), and (2) the degree to which any subgroup analyses are necessary for testing a particular hypothesis.

Further, researchers can also improve statistical power by employing “blocking” to ensure that the experimental groups are perfectly balanced on a key covariate (e.g., Bailey 2021, 346–49; Dolan 2023; see Mousa 2020 for an applied example). Power can also be improved (via increasing precision of a point estimate) by estimating the ITT while controlling/adjusting for prespecified covariates (in particular, significant predictors of the dependent variable (see Gerber et al. 2014; Mutz and Pemantle 2015)). Wuttke et al. (2023) offer an applied example of this technique with survey-experimental data (see also Gerber et al. 2014). Finally, Clifford et al. (2021) find that utilizing pre *and* posttreatment measures of *Y* yields

<sup>13</sup>Of course, small samples matter for point estimation of effect sizes as well, with smaller samples yielding a wider variance of ITT estimates.

<sup>14</sup>In general, *a priori* power analysis is more informative than post-hoc, though the latter can help determine whether *n* was sufficient given the observed ITT, pooled SD of *Y*, and (predetermined) desired level of power (Perugini, Gallucci, and Costantini 2018). See Supplemental Appendix D for additional guidance.

similar ITT estimates to the more common between-subjects design, but with substantially greater precision and, thus, greater statistical power.

#### **Alternative explanation #5: Poor measurement of the dependent variable**

Another vexing alternative explanation is measurement error in the dependent variable ( $Y$ ). Assuming one is using a valid measure of  $Y$ , greater noise in this measure will often raise the likelihood of a result that, per the NHST paradigm, is not statistically significant. Specifically, measurement error in  $Y$  is expected to increase the ITT's standard error and thus increase the likelihood of null results when significance-testing (Bailey 2021, 148–50; Berry and Feldman 1985, 26–33). Yet measurement error in  $Y$  can potentially matter for point estimation as well. For example, per Clayton et al. (2023), measurement error in the dependent variable within conjoint experiments can bias treatment-effect estimates downward toward zero.

When designing their study, researchers should consider including *multiple indicators* of  $Y$  and then combining them into a single measure (e.g., an additive scale). Doing so can substantially reduce  $Y$ 's degree of random measurement error (e.g., see Mousa 2020 for an applied example). This practice therefore offers a notable advantage over using only one indicator of  $Y$  (Ansolabehere, Rodden, and Snyder 2008; Berry and Feldman 1985, 33–34). In addition, using measures of  $Y$  that have been previously validated (either in other studies or in pretests) is a wise strategy for having a dependent variable with the best signal-to-noise ratio possible.

Once data have been collected, researchers can investigate this AE by ensuring that other measures that should, theoretically, significantly correlate with  $Y$  actually do so. If substantial correlations are found, then the measure may be considered reasonably satisfactory, even if imperfect to some extent. This, of course, requires that, during the design stage, researchers include theoretically relevant covariates in their survey (pre-treatment). Existing literature provides detailed discussions of examining measurement properties of variables (Carmines and Zeller 1979; Druckman 2022, 22–27). See Supplemental Appendix E for additional discussion.

#### **Alternative explanation #6: Ceiling and floor effects**

Related to concerns involving the measurement of  $Y$ , another well-known problem in experiments is that of either a “ceiling” or a “floor” effect (e.g., see Mullinix et al. 2015, 116). A treatment may fail to produce a significant change in  $Y$  because values of  $Y$  in the control condition are already (on average) very high (a “ceiling effect”) or very low (a “floor effect”). This AE therefore restricts the substantive magnitude of the ITT estimate. Further, per NHST, this AE will tend to increase the  $p$ -value and thus the likelihood of a null result. As with all AEs, an inability to rule out this alternative explanation renders it more difficult to determine what a nonsignificant result indicates and, thus, more difficult to assess a study's value.

In the analysis stage, one simple technique for investigating this AE is to obtain descriptive statistics (e.g., means, proportions, etc.) of  $Y$  in the control condition. Ideally, one wants to observe moderate values, or values in the direction *opposite* the effect of concern (e.g., low values if the concern is a ceiling effect), which would indicate that  $Y$  had “room” to significantly change. Brierly and Pereira (2023,



endnote 9), for example, cite their outcome variable's mean being substantially below the highest value as evidence that "ceiling effects cannot explain the [nonsignificant] result." Though somewhat uncommon in survey experiments, accounting for this AE (specifically, right- and/or left-censoring of  $Y$ ) can take the form of specifying a Tobit regression model (Muthén 1989; see Bechtel and Scheve 2017 for an applied example).

In the design stage, there are several proactive strategies that can be employed. First, researchers should focus on preventing ceiling/floor effects that are *artifactual* – i.e., arising from the measure itself, rather than what is theoretically possible. To guard against this, researchers can feature survey measures of  $Y$  that will not have extreme means. In other words, researchers can utilize measures with a more (conceptually) extreme and/or less coarse range of response options. As Vanderweedt (2022, 833) writes of their nonsignificant findings, "effects . . . might have been clearer with more nuanced (less compressed) attitude response scales." Employing *multiple* measures of  $Y$  can often assist toward this end as it is unlikely that respondents will have an extreme value on every survey item comprising the scale. Researchers can also examine means of measures that have been featured in publicly available data to help ensure, *a priori*, that such measures will not induce ceiling/floor effects if used in their own survey experiments.

#### **Alternative explanation #7: Countervailing treatment effects**

Finally, a treatment may of course have substantially different-sized (i.e., heterogeneous) effects among different subgroups within the sample (e.g., Kam and Trussler 2017). Thus, a possible explanation for nonsignificant results is that one has a special case of heterogeneous effects wherein an overall ITT can be near zero because treatment effects occur in opposite directions for different subgroups.

Consider an example in which our treatment ( $X$ ) is whether or not a U.S. political candidate adopts a stance opposing access to abortion, and our outcome ( $Y$ ) is the perceived favorability of the candidate. Given Republicans' (Democrats') generally anti-abortion (pro-choice) views, we should expect the treatment to increase  $Y$  among Republicans but *decrease*  $Y$  among Democrats. Thus, if we fail to account for partisanship and simply estimate the ITT for the whole sample, the ITT may be extremely small. Again, this would not be because the treatment was inefficacious; rather, it is because the effect occurred in opposite directions for large subsets of the sample. We might therefore refer to this as a problem of *countervailing* effects.

One strategy for investigating this AE in the analysis stage (assuming the source of potential heterogeneity is unknown) would be to compare variances of  $Y$  across treatment and control groups. This can be done visually using overlaid histograms of  $Y$  over values of  $X$ , or statistically with tests of equivalent standard deviations of  $Y$  over values of  $X$  (see Bryk and Raudenbush 1988 and Ding, Feller, and Miratrix 2016). Continuing with the above example, we should observe that the variance of candidate favorability in the treatment condition is substantially larger than in the control condition, suggesting that our treatment pushed subgroups in opposite directions. As Bryk and Raudenbush (1988, 396) contend, "To ignore variance heterogeneity . . . is tantamount to interpreting main effects while concealing significant interaction effects." Coppock et al. (2020, Appendix C) offer additional

techniques for estimating the variance of effect sizes and formally testing for treatment-effect homogeneity.

Discovering heterogeneity in treatment effects naturally raises the question of where this heterogeneity is coming from. But when investigating this, ideally, researchers should theorize about such heterogeneous effects *a priori* (in the design stage). A measure of the theorized moderating variable (*M*) can then be included in the survey (pre-treatment), enabling researchers to identify the *source* of countervailing effects by exploring how the ITT estimate varies across *M* (e.g., via specifying an interaction in a regression model).<sup>15</sup>

Crucially, researchers should exercise great caution here because, with enough exploration of interactions, one is bound to find some statistically significant (yet spurious) interactive effect. Again, testing for heterogeneous effects should be theorized – and pre-registered – before data are collected. Further, researchers should first report the ITT as a matter of transparency and also explicitly state whether any countervailing effect (if discovered) was an exploratory – rather than hypothesized – finding.

## Discussion & conclusion

Survey experiments can and do yield nonsignificant ITTs, often much to the chagrin of researchers and, in some cases, potentially resulting in abandonment of the project (i.e., the “file drawer” problem). In other cases, researchers may point to a nonsignificant ITT as evidence that *X* has no causal effect upon *Y*. Yet while the lack of a true causal effect represents one explanation for a nonsignificant result, this essay stresses that there exist at least seven AEs for nonsignificant findings in survey experiments.

The purpose of this essay is to assist researchers with more thoroughly anticipating and investigating these AEs. Failure to do so means that one (or more) of the aforementioned AEs cannot be confidently ruled out, leaving open the possibility that a nonsignificant result is due to an AE rather than being indicative of no actual effect. Toward this end, Table 1 provides recommendations that researchers can implement in the *design stage* of their experiment (see second column) that will help (1) guard against, and (2) allow for investigation of, each of the seven AEs.

Table 1 also provides a simple checklist that can be employed during the *analysis stage* (see third column). If the researcher has sufficient reason to answer “No” to any of the questions in this checklist, an AE cannot be confidently ruled out, and thus an incorrect hypothesis (or theory) might not be the reason for a nonsignificant result. Alternatively, when a researcher can confidently answer “Yes” to each item in the checklist, they should more strongly suspect that *X* has no meaningful effect on *Y*.

As illustrated above, it is not wholly uncommon for survey experimentalists to report investigating some limited number of AEs. Wuttke et al. (2023), notably, report investigating inattentiveness (*AE#1*), manipulation of the independent variable (*AE#2*), and ceiling effects (*AE#6*). At present, however, the degree to which researchers habitually consider *each one of these* AEs – either in the analysis or design stages – remains unclear.

<sup>15</sup>See Hainmueller et al. (2019) for excellent guidance on specifying interaction models.

**Table 1.** Design recommendations & a checklist of potential alternative explanations for nonsignificant results

Alternative explanation (AE)	Recommended practices in design stage	Checklist for analysis stage	Recommendations for subsequent study
#1: Inattentiveness	<ul style="list-style-type: none"> <li>○ Include pre-treatment attention check(s) prior to random assignment</li> <li>○ Include a measure of attention to the experimental manipulation (after the outcome measure)</li> </ul>	<ul style="list-style-type: none"> <li>○ Majority in each experimental condition pass the attention check? (Yes or No)</li> <li>○ Treatment assignment correlates with measure of attention to the manipulation? (Yes or No)</li> <li>○ Treatment effect is roughly the same regardless of pre-treatment attention? (Yes or No)</li> </ul>	<ul style="list-style-type: none"> <li>○ Feature experiment relatively earlier in survey</li> <li>○ More salient treatment content</li> <li>○ Consider techniques to improve respondent attentiveness</li> <li>○ Use alternative survey company</li> <li>○ Increased (pre-treatment) screening-out of inattentive respondents</li> </ul>
#2: Failure to vary the independent variable	<ul style="list-style-type: none"> <li>○ Include a manipulation check after the outcome measure</li> </ul>	<ul style="list-style-type: none"> <li>○ Treatment assignment substantially associated with manipulation check? (Yes or No)</li> </ul>	<ul style="list-style-type: none"> <li>○ Make treatment content more salient (e.g., appear sooner and/or more frequently)</li> <li>○ Make treatment stronger (e.g., more direct, forceful language)</li> </ul>
#3: Pre-treatment effect	<ul style="list-style-type: none"> <li>○ Include a measure to gauge how pre-treated a respondent might be (implemented prior to random assignment)</li> </ul>	<ul style="list-style-type: none"> <li>○ Low risk of respondents having been treated prior to experiment? (Yes or No)</li> <li>○ If “No”, is treatment effect similar among those more vs. less likely to be pre-treated? (Yes or No)</li> </ul>	<ul style="list-style-type: none"> <li>○ Make treatment stronger (assuming no ceiling/floor effect)</li> <li>○ Postpone the study until treatment has lower salience in the real world</li> <li>○ Investigate using a non-experimental design</li> </ul>
#4: Statistical power	<ul style="list-style-type: none"> <li>○ Conduct power analyses to determine necessary <i>n</i> size (assuming smallest meaningful effect size)</li> <li>○ Choose sample size cognizant of number of conditions, subgroup analyses, and likely inattentiveness</li> <li>○ Use pre-registered covariates in model, blocking, or a within-subjects design</li> </ul>	<ul style="list-style-type: none"> <li>○ Large enough sample given the ITT, variation in <i>Y</i>, and (<i>a priori</i>) desired power? (Yes or No)</li> </ul>	<ul style="list-style-type: none"> <li>○ Aim to collect a larger sample</li> <li>○ Choose sample size cognizant of quantities learned from first study: effect size, SD of <i>Y</i>, level of inattentiveness, and number/size of subgroup analyses</li> <li>○ Consider an alternative design structure (e.g., within-subjects)</li> </ul>
#5: Poor measurement of the dependent variable	<ul style="list-style-type: none"> <li>○ Check <i>Y</i>'s criterion validity</li> <li>○ If possible, use existing (validated) measures of <i>Y</i></li> </ul>	<ul style="list-style-type: none"> <li>○ Dependent variable correlates with theoretically relevant socio-demographic variables? (Yes or No)</li> </ul>	<ul style="list-style-type: none"> <li>○ Use an alternative and/or multiple measures of <i>Y</i> to reduce measurement error</li> </ul>

(Continued)

Table 1. (Continued)

Alternative explanation (AE)	Recommended practices in design stage	Checklist for analysis stage	Recommendations for subsequent study
#6: Ceiling/Floor effect	<ul style="list-style-type: none"> <li>Use measure of <math>Y</math> that is unlikely to have an extremely high/low mean (e.g., scales with more extreme end-points or multi-item scales)</li> </ul>	<ul style="list-style-type: none"> <li>For a positive (negative) hypothesized effect, does control group have an average value of <math>Y</math> well below (above) the maximum (minimum)? (Yes or No)</li> </ul>	<ul style="list-style-type: none"> <li>Assuming population of interest remains the same, use a measure of <math>Y</math> with a more (conceptually) extreme range</li> <li>Consider postponing study if ceiling/floor is due to current context</li> </ul>
#7: Countervailing treatment effects	<ul style="list-style-type: none"> <li>Include a pre-treatment measure of the moderating variable, across which countervailing effects might occur</li> </ul>	<ul style="list-style-type: none"> <li>Treatment effect is roughly the same for groups that (theoretically) might respond to treatment in opposite ways? (Yes or No)</li> <li>Variance of <math>Y</math> is roughly equal across experimental groups? (Yes or No)</li> </ul>	<ul style="list-style-type: none"> <li>Pre-register a hypothesized interaction between treatment and moderator</li> <li>Include best possible (pre-treatment) measure(s) of moderating variable to improve precision</li> </ul>

Note: Within the “Checklist” column, if “No” is answered for any AE, it suggests that a nonsignificant ITT may not be entirely due to an incorrect hypothesis or theory. Note there is debate regarding the utility of post-hoc power analysis (see Perugini, Gallucci, and Costantini 2018). The final column discusses possibilities for follow-up study assuming all procedures in “Recommended Practices in Design Stage” column were followed.

This raises an additional point: experiments can be beset by *multiple* AEs. For example, Haas and Khadka (2020, 995) fielded a study in which a nonsignificant finding could be attributable to a pre-treatment effect (AE #3) or to a ceiling effect (AE #6). Thus, researchers must be mindful of the distinct pathways by which experimental hypothesis tests can be undermined, and how to both proactively and retroactively address them. On this point, the final column of Table 1 provides recommendations for researchers who wish to field a subsequent study after (1) discovering evidence for one or more AEs in their initial study, and (2) being unable to confidently determine the extent to which a nonsignificant finding is attributable to a particular AE. These latter recommendations are therefore designed to help researchers conduct an improved test of their hypothesis based on the AE(s) discovered in their first study (see Supplemental Appendix G for further elaboration).

Notably, this essay has focused on how these AEs apply to survey experiments. Yet, while inattentiveness (AE#1) may be most relevant to survey experiments, each of the other AEs can be problematic for other types of experiments. For example, field experiments may fail to manipulate the independent variable (AE#2), while lab-based experiments may suffer from insufficient statistical power (AE#4).

In sum, the value of one's study is undermined when there exist competing explanations for the same nonsignificant result. Importantly, preventing AEs is likely far more tractable than attempting to "correct for" them *ex post*. Thus, by becoming aware of these AEs, researchers can design their studies to be better safeguarded against, and (consequently) better equipped to investigate, them once results are in-hand. This stands to enable researchers to learn far more from their survey experiments than what a naïve, nonsignificant ITT alone can provide.

**Supplementary material.** The supplementary material for this article can be found at <https://doi.org/10.1017/XPS.2024.1>

**Acknowledgements.** I would like to express immense gratitude to Jamie Druckman, Carlisle Rainey, Yamil R. Velez, Jason Barabas, Brendan Nyhan, Daniel Lakens, Charles Crabtree, Yusaku Horiuchi, and the Dartmouth Department of Government for providing extremely thoughtful feedback on early drafts of this manuscript. In addition, the four anonymous reviewers and JEPS Editors provided truly invaluable guidance on ways to improve the manuscript's clarity, contribution, and usefulness to researchers. I am deeply thankful for their efforts.

**Competing interests.** The author declares that there were no conflicts of interest in this study.

**Ethics statement.** Approval from an Institutional Review Board was not needed as this study did not contain any collection or analysis of original data.

## References

- Alrababa'h, Ala' et al. 2022. "Learning from Null Effects: A Bottom-Up Approach." *Political Analysis* 31(3): 448–56.
- Ansolabehere, Stephen, Jonathan Rodden, and James M. Snyder, Jr. 2008. "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting." *American Political Science Review* 102(02): 215–32.
- Arel-Bundock, Vincent et al. 2022. *Quantitative Political Science Research Is Greatly Underpowered*. I4R Discussion Paper Series. Working Paper. <https://www.econstor.eu/handle/10419/265531> (May 18, 2023).

- Aronow, P. M., Joshua Kalla, Lilla Orr, and John Ternovski. 2020. "Evidence of Rising Rates of Inattentiveness on Lucid in 2020." <https://osf.io/preprints/socarxiv/8sbe4/> (December 29, 2022).
- Aronow, Peter M., Jonathon Baron, and Lauren Pinson. 2019. "A Note on Dropping Experimental Subjects Who Fail a Manipulation Check." *Political Analysis* 27(4): 572–89.
- Bailey, Michael A. 2021. *Real Stats: Using Econometrics for Political Science and Public Policy*. 2nd ed. New York, NY: Oxford University Press. <https://www.amazon.com/Real-Stats-Econometrics-Political-Science/dp/0190859547> (December 5, 2022).
- Bechtel, Michael M., and Kenneth F. Scheve. 2017. "Who Cooperates? Reciprocity and the Causal Effect of Expected Cooperation in Representative Samples." *Journal of Experimental Political Science* 4(3): 206–28.
- 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): 739–53.
- Berry, William D., and Stanley Feldman. 1985. *Multiple Regression in Practice*. Newbury Park, CA: SAGE Publications.
- Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *The American Political Science Review* 113(3): 838–59.
- Brierley, Sarah, and Miguel M. Pereira. 2023. "Women Bureaucrats and Petty Corruption. Experimental Evidence from Ghana." *Research & Politics* 10(1): 1–7.
- Brophy, Nathan, and Kevin J. Mullinix. 2023. "Partisan Motivated Empathy and Policy Attitudes." *Political Behavior*. <https://doi.org/10.1007/s11109-023-09890-x>.
- Brutger, Ryan et al. 2023. "Abstraction and Detail in Experimental Design." *American Journal of Political Science* 67(4): 979–95.
- Bryk, Anthony S., and Stephen W. Raudenbush. 1988. "Heterogeneity of Variance in Experimental Studies: A Challenge to Conventional Interpretations." *Psychological Bulletin* 104(3): 396–404.
- Carmines, Edward G., and Richard A. Zeller. 1979. *Reliability and Validity Assessment*. Beverly Hills, CA: SAGE Publications.
- Chambers, Christopher D., and Loukia Tzavella. 2022. "The Past, Present and Future of Registered Reports." *Nature Human Behaviour* 6(1): 29–42.
- Chong, Dennis, and Jane Junn. 2011. "Politics from the Perspective of Minority Populations." In *Cambridge Handbook of Experimental Political Science*, ed. James N. Druckman, Donald P. Greene, James H. Kuklinski and Arthur Lupia. Cambridge University Press, 320–35.
- Clayton, Katherine et al. 2023. "Correcting Measurement Error Bias in Conjoint Survey Experiments." <https://gking.harvard.edu/sites/scholar.harvard.edu/files/gking/files/conerr.pdf>.
- Clifford, Scott, Geoffrey Sheagley, and Spencer Piston. 2021. "Increasing Precision without Altering Treatment Effects: Repeated Measures Designs in Survey Experiments." *American Political Science Review* 115(3): 1048–65.
- Coppock, Alexander, Seth J. Hill, and Lynn 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): eabc4046.
- Ding, Peng, Avi Feller, and Luke Miratrix. 2016. "Randomization Inference for Treatment Effect Variation." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 78(3): 655–71.
- Dolan, Lindsay. 2023. "10 Things to Know About Randomization." EGAP. <https://egap.org/resource/10-things-to-know-about-randomization/> (January 2, 2023).
- Druckman, James, and Donald P. Green. 2021. *Advances in Experimental Political Science*. New York: Cambridge University Press.
- Druckman, James N. 2022. *Experimental Thinking: A Primer on Social Science Experiments*. New York, NY: Cambridge University Press. <https://faculty.wcas.northwestern.edu/~jnd260/pub/Druckman%20Experimental%20Thinking%20Fall%202020%20Submitted.pdf>.
- Druckman, James N., and Thomas J. Leeper. 2012. "Learning More from Political Communication Experiments: Pretreatment and Its Effects." *American Journal of Political Science* 56(4): 875–96.
- Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345(6203): 1502–5.
- Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15(1): 1–20.

- Gerber, Alan et al.** 2014. "Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee." *Journal of Experimental Political Science* 1(1): 81–98.
- Gerber, Alan S., and Donald P. Green.** 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton & Company.
- Gill, Jeff.** 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(3): 647–74.
- Haas, Nicholas, and Prabin B. Khadka.** 2020. "If They Endorse It, I Can't Trust It: How Outgroup Leader Endorsements Undercut Public Support for Civil War Peace Settlements." *American Journal of Political Science* 64(4): 982–1000.
- Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu.** 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* 27(2): 163–92.
- Harden, Jeffrey J., Anand E. Sokhey, and Katherine L. Runge.** 2019. "Accounting for Noncompliance in Survey Experiments." *Journal of Experimental Political Science* 6(3): 199–202.
- Hartman, Erin, and F. Daniel Hidalgo.** 2018. "An Equivalence Approach to Balance and Placebo Tests." *American Journal of Political Science* 62(4): 1000–1013.
- Horowitz, Michael C., and Matthew S. Levendusky.** 2011. "Drafting Support for War: Conscription and Mass Support for Warfare." *The Journal of Politics* 73(02): 524–34.
- "Journal of Experimental Political Science."** 2023. *Cambridge Core*. <https://www.cambridge.org/core/journals/journal-of-experimental-political-science/information/about-this-journal> (January 2, 2023).
- Kam, Cindy D., and Marc J. Trussler.** 2017. "At the Nexus of Observational and Experimental Research: Theory, Specification, and Analysis of Experiments with Heterogeneous Treatment Effects." *Political Behavior* 39(4): 789–815.
- Kane, John V., and Jason Barabas.** 2019. "No Harm in Checking: Using Factual Manipulation Checks to Assess Attentiveness in Experiments." *American Journal of Political Science* 63(1): 234–49.
- Kane, John V., Yamil R. Velez, and Jason Barabas.** 2023. "Analyze the Attentive and Bypass Bias: Mock Vignette Checks in Survey Experiments." *Political Science Research and Methods* 11(2): 293–310.
- Lakens, Daniel.** 2022. "8. Sample Size Justification." [https://lakens.github.io/statistical\\_inferences/08-samplesizejustification.html](https://lakens.github.io/statistical_inferences/08-samplesizejustification.html).
- Linos, Katerina, and Kimberly Twist.** 2018. "Diverse Pre-Treatment Effects in Survey Experiments." *Journal of Experimental Political Science* 5(2): 148–58.
- Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres.** 2018. "How Conditioning on Post-Treatment Variables Can Ruin Your Experiment and What to Do about It." *American Journal of Political Science* 62(3): 760–75.
- Mousa, Salma.** 2020. "Building Social Cohesion between Christians and Muslims through Soccer in Post-ISIS Iraq | Science." *Science* 369(6505): 866–70.
- Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese.** 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2(02): 109–38.
- Muthén, Bengt O.** 1989. "Tobit Factor Analysis." *British Journal of Mathematical and Statistical Psychology* 42(2): 241–50.
- Mutz, Diana C.** 2011. *Population-Based Survey Experiments*. Princeton: Princeton University Press.
- Mutz, Diana C.** 2021. "Improving Experimental Treatments in Political Science." In *Advances in Experimental Political Science*, ed. James N. Druckman and Donald P. Green. Cambridge University Press, 219–38.
- Mutz, Diana C., and Robin Pemantle.** 2015. "Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods." *Journal of Experimental Political Science* 2(02): 192–215.
- Nature.** 2023. "Nature Welcomes Registered Reports." *Nature*. <https://www.nature.com/articles/d41586-023-00506-2> (May 17, 2023).
- Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor.** 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences* 115(11): 2600–06.
- Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko.** 2009. "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power." *Journal of Experimental Social Psychology* 45(4): 867–72.
- Perugini, Marco, Marcello Gallucci, and Giulio Costantini.** 2018. "A Practical Primer To Power Analysis for Simple Experimental Designs." *International Review of Social Psychology* 31(1): 20.

- Rainey, Carlisle.** 2014. "Arguing for a Negligible Effect." *American Journal of Political Science* 58(4): 1083–91.
- Searles, Kathleen, and Kyle Mattes.** 2015. "It's a Mad, Mad World: Using Emotion Inductions in a Survey." *Journal of Experimental Political Science* 2(2): 172–82.
- Slothuus, Rune.** 2016. "Assessing the Influence of Political Parties on Public Opinion: The Challenge from Pretreatment Effects." *Political Communication* 33(2): 302–27.
- "The Journal of Politics: Registered Report Guidelines."** 2022. *The Journal of Politics*. <https://www.journals.uchicago.edu/journals/jop/registered-report-guidelines?doi=10.1086%2Fjop&publicationCode=jop> (January 2, 2023).
- Vandeweerdt, Clara.** 2022. "In-Group Interest Cues Do Not Change Issue Attitudes." *Politics, Groups, and Identities* 10(5): 828–36.
- Varaine, Simon.** 2023. "How Dropping Subjects Who Failed Manipulation Checks Can Bias Your Results: An Illustrative Case." *Journal of Experimental Political Science* 10(2): 299–305.
- Wuttke, Alexander, Florian Sichart, and Florian Foos.** 2023. "Null Effects of Pro-Democracy Speeches by U.S. Republicans in the Aftermath of January 6th." *Journal of Experimental Political Science*: 1–15. <https://doi.org/10.1017/XPS.2023.17>.

---

**Cite this article:** Kane JV. More than meets the ITT: A guide for anticipating and investigating nonsignificant results in survey experiments. *Journal of Experimental Political Science*. <https://doi.org/10.1017/XPS.2024.1>