MORE THAN MEETS THE ITT:
A GUIDE FOR INVESTIGATING NULL RESULTS

John V. Kane
New York University
jvk221@nyu.edu

Draft Manuscript
(05/19/23)

Word Count: 7275

ABSTRACT

Experiments can often yield non-significant and/or substantively weak intention-to-treat (ITT) effects. In recent years, there has been a commendable shift toward publishing such “null results”, either to avoid the “file drawer problem” and/or to encourage “null-by-design” studies. But how can researchers be more confident that null results are actually evidence of “no effect”? An inability to rigorously address this crucial question stands to diminish the theoretical and empirical value of one’s study. Yet, in contrast to the extant literature on field experiments, there exists little explicit guidance on how researchers can more rigorously investigate weak ITT estimates in survey-based experiments. In response, this essay highlights seven “pitfalls” that can lead survey experiments to yield null findings, and draws upon recent research to show how various “checks” and other techniques can be employed to detect and address them. While some of these pitfalls may be relatively familiar to researchers doing survey-experimental work, two such “pitfalls”—namely, (1) respondent inattentiveness, and (2) failed manipulation of the independent variable—have only recently received substantial scholarly focus. Researchers are thus provided with a set of simple strategies, including a checklist, for troubleshooting, diagnosing, and more thoroughly investigating their survey-experimental data before concluding that a given treatment has “no effect.”

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. In response, scholars have begun dedicating more careful thought toward the choices that researchers face when designing survey experiments and analyzing their experimental results (e.g., Druckman and Green 2021; Druckman 2022; Mutz 2011).

However, relatively little scholarship has yet to detail and address the various formidable factors that can often *undermine* a researcher's survey experiment. This stands in notable contrast to literature on other forms of experiments (e.g., field experiments), in which there exists explicit guidance on how to grapple with common challenges (e.g., noncompliance; see Gerber and Green 2012; Karlan and Appel 2016). Yet knowledge of such factors is extremely important. First and foremost, researchers can benefit greatly from knowing whether a "null result" (e.g., p >.05) is actually evidence of no treatment effect or, in contrast, simply a consequence of one or more of these undermining factors. Not being able to address this crucial question stands to greatly diminish the theoretical and empirical value of one's study. Second, being able to rigorously investigate null results is especially valuable 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 null results that occur despite an experiment being well-designed (Chambers and Tzavella 2022; Journal of Experimental Political Science 2023; The Journal of Politics 2022; Nature 2023).

In its most basic form, an online survey experiment randomly assigns individuals to one of at least two distinct conditions within a survey. The condition to which one is assigned is regarded as a 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.

For example, a researcher might examine whether Y's mean when X=0 differs from Y's mean when X=1 at conventional levels of statistical significance (e.g., p<.05). 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).[1]

Suppose that a researcher has fielded a survey experiment and the intention-to-treat (ITT) effect is non-significant at conventional levels and/or very close to zero in magnitude. The researcher may infer from this that the hypothesis being tested is incorrect. Perhaps it is. This result may also cast doubt upon the underlying theory from which the hypothesis was derived. Perhaps it should. However, before arriving at such an interpretation, there are various alternative explanations—i.e., "pitfalls"—that the researcher might consider *before* concluding that the survey experiment "failed" or interpreting the results as evidence of the null hypothesis being true.

This essay highlights tools that researchers can use to more thoroughly prepare for, troubleshoot, and investigate null and/or weak survey-experimental results. Specifically, it details a variety of pitfalls that can increase the likelihood of null findings, even when a hypothesis and/or an underlying theory *is correct*. The essay concludes with a checklist that researchers can use to better diagnose their experimental findings, especially when these findings are "null", disappointingly weak or, alternatively, welcomed enthusiastically as evidence that a particular hypothesis is false (i.e., when a researcher is anticipating a null result).

Of course, the aim of this essay is not to provide experimentalists with additional "researcher degrees of freedom"—i.e., more avenues to find "significant" results, raising the risk of Type I ("false positive") errors in the process. To guard against this possibility, researchers should

[1] As Harden et al. (2019, 200) explain, the ITT differs from the commonly-employed *average treatment effect* (ATE) insofar as the latter implicitly assumes full compliance, whereas the former does not.

strongly consider pre-registering their studies and analysis plans, particularly the analyses they would conduct in order to detect and address the various pitfalls identified here (Druckman 2022, 143–44; Nosek et al. 2018). Researchers should also first report ITT results in the interest of full transparency.

These concerns notwithstanding, given the great deal of effort that often goes into designing and fielding a survey experiment, knowledge of how to investigate and rule out the various pitfalls that can undermine an experiment stands to aid researchers in getting the most of out of their studies.

TWO PITFALLS THAT CAN BE INVESTIGATED WITH "CHECKS"[2]

Two pitfalls that are common to survey experiments—(1) *respondent inattentiveness*, and (2) *failure to manipulate the independent variable*—have received greater scholarly focus in recent years (Harden, Sokhey, and Runge 2019; Kane and Barabas 2019; Kane, Velez, and Barabas 2023; Mutz and Pemantle 2015; Peyton, Huber, and Coppock 2022). These two particular pitfalls are also unique in that they can be directly investigated by including specialized "checks" within one's survey.

Pitfall #1: Respondent Inattentiveness

Respondent inattentiveness to a treatment is a form of noncompliance. Crucially, when a sample is insufficiently attentive to a treatment, a treatment effect (if there actually is one) will

[2] This essay does not include "randomization checks" (Gerber et al. 2014), otherwise known as "balance tests", which test whether particular pre-treatment measures are significantly predictive of treatment assignment (ideally, *no* pre-treatment measure will be significantly predictive of treatment assignment). While such checks are indeed useful for experimentalists (because they help ensure that relevant predictors of Y are balanced across experimental conditions), there is no *a priori* reason to expect that a failed randomization will necessarily bias ITT estimates *downward* toward null results (i.e., they may just as easily bias estimates *upward*). In short, researchers may perform a randomization check regardless of the ITT estimate they obtain.

tend to be biased toward zero (Bailey 2021; Kane, Velez, and Barabas 2023). Put simply, the more a (randomly assigned) treatment group fails to attend to the treatment, the less they are (in reality) a *treatment* group. To the extent this is true, we should expect a smaller difference in the outcome measure between the treatment and control groups. In other words, we should expect a smaller ITT, and thus a greater likelihood of null results. By the same logic, inattentiveness is therefore also likely to increase the likelihood of observing a failed manipulation (see *Pitfall #2* below). In short, attentiveness is often precondition for having a successful treatment.

Researchers fielding survey experiments should therefore plan to investigate whether the sample was sufficiently attentive. There exists a variety of ways to conduct such an investigation using specialized "checks". *Factual Manipulation Checks* (FMCs; Kane and Barabas 2019) are factual questions which ask respondents about content that was shown in the experimental stimulus (e.g., a vignette), with incorrect answers indicating insufficient attentiveness to the experiment. *Mock Vignette Checks* (MVCs; Druckman 2022, 56; Kane, Velez, and Barabas 2023) are also factual questions, but which ask about information contained in a non-varying, *pre-treatment* vignette that is shown to all respondents (again, incorrect answers indicate insufficient attentiveness). Generic "attention checks" (Aronow et al. 2020), Instructed Response Items (IRI; Alvarez et al. 2019), and "bogus items" (Clifford and Jerit 2014; Meade and Craig 2012) either explicitly instruct respondents to indicate a particular response as a way of demonstrating attention to the question, or feature an unambiguously incorrect response option that, if selected, also indicates insufficient attentiveness. Similarly, *Instructional Manipulation Checks* (IMCs; Berinsky, Margolis, and Sances 2014; Oppenheimer, Meyvis, and Davidenko 2009), also known as "screeners", embed specific instructions as to how respondents should answer before asking respondents an ostensibly generic survey question (i.e., akin to a "trick" question).

Somewhat alternatively, question timers (Harden, Sokhey, and Runge 2019; Niessen, Meijer, and Tendeiro 2016; Wood et al. 2017) record the amount of time a respondent spends on a particular screen and/or the survey as a whole. Thus, low amounts of time ("latencies") indicate insufficient attentiveness. In contrast to this latter technique, the former techniques afford the researcher a discrete measure of attentiveness (by virtue of there existing only one correct response option), whereas the use of question timers requires the researcher to establish some decision rule regarding what constitutes a "sufficient amount of time" to be considered attentive.[3]

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). While no agreed-upon *acceptable* level of inattentiveness exists[4], it is important reiterate that *any* amount of inattentiveness will tend to weaken ITT estimates.[5]

If substantial (e.g., 20% or more) inattentiveness is found (either for the sample as a whole and/or within each experimental condition), what can be done?[6]

[3] An additional consideration when using question timers is *which* timers to use for gauging attentiveness, particularly when different respondents see different versions of the survey (e.g., some respondents might be shown follow-up questions on a separate screen because of a response option they chose, whereas other respondents will not have seen those additional questions).

[4] An "acceptable" level of inattentiveness would depend greatly upon other aspects of the study (e.g., the size of the sample, number of experimental conditions, and strength of the treatment (see Pitfall #3 below for additional details)).

[5] In their discussion of estimating the complier average causal effect (CACE) in field experiments with one-sided non-compliance, for example, Gerber and Green (2012, 151) show that the CACE is equal to the ITT divided by the proportion of respondents who were actually treated. By implication, then, having survey respondents who fail to attend to an assigned treatment will result in an ITT estimate that is always smaller in magnitude than the CACE.

[6] It is worth noting that, in some cases, researchers may wish to know the ITT insofar as it represents a better estimate of how effective the treatment would be in the real world (i.e., in the real world, we would expect substantial noncompliance among those assigned to the treatment group). In other words, while models that account for inattentiveness possess likely greater *internal* validity, the ITT may represent an estimate that possesses greater *external* validity.

First, researchers can use employ the use of a treatment-relevant FMC (see Kane and Barabas 2019). While substantial inattentiveness may exist, if a researcher has also included a treatment-relevant FMC in the study, they are able to test whether attentiveness to the manipulated information significantly covaries with treatment assignment. (This is not the case for other kinds of attention checks, the answers of which do not depend upon treatment assignment.) If so, it demonstrates that, at a minimum, respondents in one condition attended (on average) to different information than respondents in another condition. Conversely, if there is no such covariation, it indicates that inattentiveness was severe enough that the experimental groups did not attend to different information to a significantly large extent.

Second, researchers can employ the use of MVCs in the analysis of treatment effects. Specifically, because MVCs are measured pre-treatment, researchers are able to test whether treatment effects are larger for subsamples that are more (versus less) attentive (see Kane, Velez and Barabas 2023). This enables the researcher to test whether (as one would expect) treatment effects are larger for those that were the most likely to attend to the information in the experiment.[7]

Recent research has identified additional techniques for conducting such a test. One approach uses random assignment as an instrument for attentiveness (measured either by a manipulation check or by a question timer), which permits estimation of a *complier average causal effect* (CACE; Harden, Sokhey, and Runge 2019; cf. Kane, Velez, and Barabas 2023 Appendix J). An alternative approach uses many question timers within a survey to create an individual response-time score that can then be used for estimating treatment effects on more (versus less) attentive subsets of survey respondents (Read, Wolters, and Berinsky 2022).

[7] Researchers should report whether the attentive subsample compositionally differs, to any substantive extent, from the sample as a whole.

Yet, regardless of the specific technique that is used, researchers should avoid using any *post*-treatment measure of attentiveness to conduct such an analysis as it runs the risk of undermining random assignment and inducing statistical bias (Aronow, Baron, and Pinson 2019; Montgomery, Nyhan, and Torres 2018; Varaine 2022).[8] This includes the common practice of removing (or statistically controlling for) respondents who spent too little time viewing the content in their experimental condition (per a question timer) as well as respondents who incorrectly answered a post-treatment attention check, FMC, or IMC.

In sum, if a researcher discovers a high rate of inattentiveness but did *not* feature measures of attentiveness, their options are severely limited: they will have little sense of how attentive the sample was, cannot perform an unbiased test of whether the treatment was attended to, and cannot test whether the treatment was at least efficacious among those respondents who actually attended to it. As a consequence, the researcher cannot confidently rule out that a given null result was simply due to inattentiveness, therein limiting the value of their study. Table 1 summarizes the tools that can be used to detect this potential pitfall, as well as procedures that can be used should the problem be found.

[8] The instrumental variable and timer approaches (described above) technically do feature post-treatment variables. Thus, researchers should perhaps exercise greater caution when interpreting results should they decide to employ either of these particular techniques.

TABLE 1. Pitfall #1: Inattentiveness (Noncompliance)

Potential Pitfall	Checks Available	Detection Methods	If Problem Is Found…
Respondent Inattentiveness	*a) FMCs* *b) MVCs* *c) IMCs* *d) Attention Checks* *e) Instructed Response Items* *f) Question timers*	*a) Examine FMCs, MVCs, IMCs, other checks, and/or question timers for sample as a whole and for each experimental condition*	*Use MVCs (or alternative pre-treatment measures) to test for larger treatment effects among attentive vis-à-vis inattentive*
		b) If using an FMC-TR, ensure responses significantly correlate with treatment assignment	

Pitfall #2: Failure to Vary the Independent Variable of Interest

Regardless of whether there is an inattentiveness problem in the sample, a second potential pitfall is that one's independent variable was not actually varied by the treatment (Mutz and Pemantle 2015). Imagine a treatment in which we try to increase respondents' belief that a draft is likely to be reinstated (to study if it affects respondents' support for war (Y)).[9] We may believe our treatment will accomplish this task, but whether it *actually* does so is an empirical question.

What to do? The key recommendation here is to feature a classic manipulation check, also known as a *Subjective Manipulation Check* (SMC; Kane and Barabas 2019).[10] An SMC is a survey item that (1) follows a condition to which one was randomly assigned, 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 a military draft as being more likely, then the

[9] This example comes from Horwitz and Levendusky (2011).

[10] These are labeled "subjective" because, unlike factual manipulation checks, there is no single correct answer, nor is the measure asking about factual content.

SMC should measure respondents' perceived likelihood of a draft; if a treatment is designed to make people feel more anxious, we should have an SMC that measures anxiety, and so on.

Empirically, we want to then confirm that a treatment group significantly differs from the control group on this SMC measure and, ideally, to a substantively large degree. If this is the case, it suggests the treatment accomplished what the researcher intended. It also indicates that, whatever level of inattentiveness existed in the sample, it was not substantial enough to prevent the researcher from successfully manipulating the independent variable of interest.[11]

Despite their simplicity and enormous utility, such manipulation checks remain in surprisingly low usage: well under 50% of published experimental studies in political science feature a manipulation check (Kane and Barabas 2019; Mutz 2021). Again, as these measures provide crucial information about a common pitfall, researchers would do well to regularly feature them in their experiments. As Mutz (2011, 84–85) argues, instances in which researchers should consider such manipulation checks "optional" are "relatively few and far between."

What if treatment assignment is *not* significantly associated with SMC responses? 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 SMC (just as we would expect an attenuated effect on Y).

If we instead determine that inattentiveness is not overly problematic (e.g., because attentiveness levels were sufficient and treatment assignment correlates substantially with FMC responses), *and* can reasonably assume that the SMC is a valid measure of the independent variable

[11] It is worth stressing, though, that an SMC *assumes*—rather than tests for—sufficient attentiveness. Researchers can, of course, perform manipulation check analyses across different levels of attentiveness (measured pre-treatment) to determine whether SMC results are significantly stronger for the attentive (and, among the attentive, whether the effect of the treatment on the SMC results is statistically distinguishable from zero).

the researcher is intending to manipulate, then a null relationship between random assignment and the SMC may suggest one of two possibilities. First, the SMC may simply be too "noisy" a measure. This can be investigated by testing whether theoretically-relevant variables (e.g., education, age, political attitudes) significantly correlate with the SMC. That is, we can investigate the SMC's *criterion validity* (e.g., Druckman 2022, 22–27). If substantial correlations are found, it indicates that the measure is not entirely noise and thus should, in principle, be manipulable.

If the aforementioned possibility can be ruled out, then the second 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 buried within many paragraphs of non-manipulated text, images, etc. Whatever this manipulated content might affect, its efficacy is potentially being crowded out by the other information that respondents are processing (e.g., see Mutz 2011, 58). In short, the treatment is not nearly potent enough to actually induce variance in the independent variable. As a result, the treatment will not be significantly predictive of the SMC, nor is it likely, therefore, to predict a downstream outcome of interest (Y). Remedying this would require the researcher to think more carefully about what one's treatment actually involves, perhaps even pre-testing alternative versions of the treatment to determine which is most strongly associated with the SMC (Chong and Junn 2011, 329–30; Mutz 2011, 65). Table 2 summarizes the tools that can be used to detect this potential pitfall, as well as procedures that can be implemented should the problem be found.

TABLE 2. Pitfall #2: Failed Manipulation of the Independent Variable

Potential Pitfall	Check Available	Detection Methods	If Problem Is Found…
Failed Manipulation of the Independent Variable	*Subjective Manipulation Check(s)*	*Examine SMC to ensure Control and Treatment groups substantially differ*	*a) Check whether the failed manipulation may also be due to inattentiveness (e.g., check if manipulation was effective among the attentive)* *b) Test for criterion validity between other, theoretically-related items and SMC* *c) Rethink treatment (e.g., for clarity, cognitive overload, and potency)* *d) Possibly pre-test different versions of treatment to ensure effect on SMC*

FIVE PITFALLS THAT CAN BE ADDRESSED BY OTHER MEANS

Beyond inattentiveness and failure to manipulate the independent variable, there exists a variety of additional pitfalls that can undermine a researcher's experiment. While they cannot be investigated using the various "checks" described above, there are nevertheless potential methods via which the researcher can better diagnose—and prevent—each pitfall.

Pitfall #3: Insufficient Statistical Power

Power considerations are a well-established culprit of null results (Alrababa'h et al. 2022). Nevertheless, one recent study finds that political science research is often severely underpowered, and yet also that most methodologists *overestimate* the share of studies that are sufficiently powered (Arel-Bundock et al. 2022). In particular, a small sample—given the number of experimental groups and anticipated magnitude of the treatment(s)—is a common way in which null (i.e., the conventional p >.05) findings become more likely, even when a treatment is indeed efficacious. In the extreme case, no matter how potent one's treatment, too small a sample will yield null findings (e.g., see Perugini, Gallucci, and Costantini 2018).

Obviously, the most direct approach for overcoming a lack of statistical power is to increase statistical power via increasing the sample size. More concretely, researchers should be mindful of how many conditions exist in their experiment and whether any subgroup analyses will be performed. A sample size of 1000, for example, may initially seem sufficient with respect to power. However, if the experiment involves five conditions, and will involve subsetting the data on party identification (e.g., Democrats, Independents, and Republicans), the researcher could ultimately be estimating treatment effects among groups of only several dozen respondents.[12] Thus, while researchers may only have partial control over the total sample size (e.g., because of resource constraints), greater discretion can be exercised over the number of experimental conditions and necessary subgroup analyses.

Researchers can also improve power by employing "blocking" to ensure that the experimental groups are perfectly balanced on a potentially confounding variable (e.g., see Bailey 2021, 346–49). For example, Mousa (2020) employed a pretreatment survey to allow for blocking on baseline attitudes. This practice minimizes covariation between our assignment variable and the confounding variable, thus improving the precision of our ITT estimate and lowering the possibility of a Type II error (e.g., see Dolan 2023). Along similar lines, researchers can also improve precision by estimating the ITT while controlling (i.e., "adjusting") for pre-specified covariates (in particular, significant predictors of the dependent variable (see Mutz and Pemantle

[12] For example, if 30% of respondents in the sample are Republican, with five conditions, an analysis of Republicans only will yield groups of approximately $n = 60$.

2015)), though such a modeling decision should ideally be pre-registered before the data are collected rather than *ex post* (Gerber et al. 2014).

Pitfall #4: Poor Measurement of the Dependent Variable

Another vexing problem is measurement error in *Y*. A "noisy" measure of Y, and/or a measure that is not actually measuring the underlying concept of interest, may make it more difficult for researchers to find that a treatment is efficacious. Specifically, measurement error in the dependent variable is expected to increase the size of a treatment estimate's standard error, thus increasing the likelihood of null results (Bailey 2021, 148–50; Berry and Feldman 1985, 26–33). In addition, Clayton et al. (2023) stress that measurement error in the dependent variable within conjoint experiments can also bias treatment-effect estimates downward.

What to do? Once data have been collected, researchers can check for criterion validity 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 valid, even if imperfect to some extent. Existing literature provides detailed discussions of examining measurement properties of variables (Carmines and Zeller 1979; Druckman 2022, 22–27).

When researchers are designing their study, including *multiple indicators* of *Y*—and then combining them into a single measure (e.g., a scale or index of some kind)—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 pre-tests) is also a wise strategy for having a dependent variable with the best signal-to-noise ratio possible.

Pitfall #5: Heterogeneous (i.e., Countervailing) Treatment Effects

A treatment may have substantially different-sized (i.e., heterogeneous) effects among different subgroups in the sample. Such a scenario is commonly known as "moderation" (i.e., the variable identifying the subgroups *moderates* the treatment effect) and is commonly analyzed via specifying interactions in regression analysis (e.g., Kam and Trussler 2017). Thus, when an experiment yields null results, a possible explanation is that one has a special case of heterogeneous effects wherein an overall ITT effect can be near 0 (or, at least, substantively very weak) because treatment effects go in opposite directions for different subgroups.[13]

Consider an example in which our treatment *(X)* is whether or not a U.S. respondent reads a news article that communicates negative information about a Republican Party president, and our outcome *(Y)* is trust in news media. Because such information will likely be congenial to Democrats, but uncongenial to Republicans, this treatment may cause Democrats to increase trust in the news media, yet cause Republicans to *decrease* trust. Thus, if we fail to take partisanship into account and simply examine the ITT for the sample as a whole, we may find the ITT estimate to be extremely small. But this is not because the treatment was inefficacious; rather, it is because the treatment effect occurred in opposite directions for large subsets of the sample. We might call this a problem of "countervailing effects" insofar as a treatment is efficacious, yet in such a way that it serves to cancel itself out.

What to do? One relatively simple strategy for investigating this pitfall would be by comparing variances, rather than means, across treatment and control groups. This can be done visually (using, for example, overlaid histograms or kernel density plots of *Y* across values of *X*) or

[13] Similarly, effects might not be oppositely signed, but still substantially different enough to result in a weak effect overall. For example, the effect for a majority of the sample may be near zero, and the effect may be substantial for a minority of the sample, leading to an overall weak effect.

statistically (using, for example, tests of equivalent standard deviations of *Y* across values of *X*). Continuing with the above example, we should see that the variance of trust in news media *(Y)* in the treatment condition is substantially larger than in the control condition—a clue that our treatment may have pushed subgroups in different directions.

Additionally, we can attempt to identify the *source* of these countervailing effects by exploring interactions.[14] In other words, we can examine whether a theoretically-relevant variable—e.g., party identification of the respondent—does indeed moderate the treatment effect to a substantial extent, with treatment effect estimates going in opposite directions depending upon the value of the moderating variable. (For the same reasons noted above, such a variable should be measured *pre*-treatment.) Because such a process is inherently *post hoc*, though, researchers should first report the ITT as a matter of transparency, and also explicitly state that any significant result was an exploratory—rather than hypothesized—finding.[15]

Pitfall #6: Ceiling and Floor Effects

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 increase in *Y*

[14] Researchers should exercise caution here because, as is well known, with enough exploration of interactions, one is bound to find some significant treatment effect among some subgroup (adding to this, with enough exploration, one will likely conduct some analysis using a variable that was not balanced across conditions, raising the risk of a spurious rather than real finding in the sample itself). Researchers conducting such investigations should be transparent about this exploratory process when reporting their findings. As a separate matter, featuring interactions will tend to reduce statistical power insofar as the interaction terms 1) require an additional degree of freedom to estimate, and 2) will tend to enlarge standard errors via collinearity with their constitutive terms (Kam and Franzese Jr. 2007).

[15] Importantly, finding a "significant moderator" does not imply *causal* moderation: if the moderator is not manipulated, it may simply be correlated with the real causal moderator (Kam and Trussler 2017).

because values of *Y* in the control condition are already (on average) very high (a "ceiling effect"). Conversely, a treatment may fail to show a significant decrease in *Y* because values of *Y* in the control condition are already (on average) very low (a "floor effect"). This restriction on the magnitude of the ITT will, *ceteris paribus*, therefore increase the *p*-value and thus the likelihood of a null result. A treatment may therefore appear ineffective precisely because, in essence, there exists very little "room" for *Y*'s average to move any further. Again, an inability to rule out this alternative explanation for a null result renders it more difficult to determine the value of one's study.

The obvious approach is to analyze descriptive statistics (e.g., means, proportions, etc.) of *Y* in the control condition. Ideally, the researcher would want to see moderate values, or values in the direction opposite to 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. While little can be done *post hoc* to correct for ceiling/floor effects, when designing an experiment, researchers might consider using an alternative measure of *Y* that would ideally not have as high, or low, a mean. Somewhat alternatively, researchers can aim for a measure of *Y* that at least yields greater variation across respondents (e.g., by offering finer-grained response options and/or using multiple measures).

Pitfall #7: Pre-Treated Respondents

A final potential pitfall is that of "pre-treatment effect" (see Druckman and Leeper 2012; Gaines, Kuklinski, and Quirk 2007). Specifically, a treatment may indeed be efficacious, but perhaps respondents have largely been treated *prior to the study*, by the real world, with the same information that the researcher is employing as the treatment.

For example, at the height of the COVID-19 pandemic, a survey experiment in which we 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 will appear to have a small, perhaps even non-significant effect on *Y*, not because the treatment is actually ineffective, but because it has already occurred prior to the experiment. Indeed, pre-treatment will, except under special circumstances, tend to bias treatment effects toward zero (see Gaines, Kuklinski, and Quirk 2007) and thus also increase the likelihood of a null result.

Importantly, a pre-treatment effect should *also* result in a failed subjective manipulation check (see Pitfall #2 above). In the presence of surprisingly weak results and a failed manipulation check, therefore, researchers should consider whether respondents may have been, in effect, already treated by the real world prior to beginning the experiment. The likelihood of this pitfall will, naturally, depend very much upon what the researcher is employing as treatment stimuli. In situations where a pre-treatment effect is more likely, the researcher may benefit from having respondents answer a survey question (prior to random assignment) that could assist with determining whether substantial pretreatment has occurred (e.g., answering the degree to which one has been following a particular topic or story in the news), and then testing for an effect among those who are less likely to have been pre-treated.

Summary

Each of the pitfalls detailed above can potentially undermine one's experiment. By becoming familiar with each one of them, researchers can more confidently rule out the possibility that null or weak results are driven by insufficient attentiveness, failure to manipulate the independent variable, insufficient statistical power, measurement error in the dependent variable, countervailing effects, a ceiling/floor effect, and/or pre-treatment. Researchers therefore stand to increase the theoretical and empirical value of their experimental studies by investigating each of these possible pitfalls.

Toward this end, Table 3 provides a checklist for researchers. If the researcher has sufficient reason to answer "No" to *any* of the questions in the checklist, it suggests that an incorrect hypothesis (or theory) might not be responsible for underwhelming experimental results. Put differently, only when a researcher is able to answer "Yes" to every item in Table 3 should they begin to more strongly suspect that the hypothesis, and/or underlying theory itself, is incorrect.

Of course, experiments can be beset by *multiple* pitfalls. For example, Haas and Khadka (2020, 995) fielded a study in which a null finding could have been attributable to a ceiling effect (Pitfall #6) *or* a pretreatment effect (Pitfall #7) in the form of prior familiarity with particular politicians' policy stances. As this example illustrates, researchers need to be mindful of the distinct pathways by which experimental hypothesis tests can be undermined, as well as how to both proactively and retroactively address them.

Importantly, the investigation of—and possible corrections for—these various pitfalls obviously require additional analyses. Yet, with a greater number of analyses being conducted, the risk of a *false positive* naturally increases. Thus, at a minimum, researchers should make every

TABLE 3: A Checklist of Potential Pitfalls for Experimentalists

Checklist	Yes	No	Potential Concern if "No"
Majority in each experimental condition attentive to attention check, mock vignette check, and/or factual manipulation check?			*Inattentiveness to survey in general*
Treatment assignment correlates with responses to factual manipulation check (FMC)?			*Inattentiveness to the manipulated content*
Treatment effect is roughly the same regardless of mock vignette check (MVC) performance?			*Inattentiveness biasing ITT downward*
Treatment assignment significantly predicts subjective manipulation check (SMC)?			*Failure to manipulate the independent variable*
Large enough sample given the anticipated effect size?			*Insufficient statistical power*
Dependent variable correlates with theoretically-relevant socio-demographic variables?			*Measurement error / invalid measure of the dependent variable*
Treatment effect is roughly the same for groups that, a priori, should (versus should not) be susceptible to the treatment?			*Heterogeneous (i.e., countervailing) treatment effects*
For a positive hypothesized effect, does control group have an average value of Y well below the maximum?			*Ceiling effect*
For a negative hypothesized effect, does control group have an average value of Y well above the minimum?			*Floor effect*
Low risk of respondents having been "treated" prior to experiment?			*Pretreatment effect*

Note: If "No" is answered for any applicable pitfall, it suggests that null or weak ITT results may not be entirely due to an incorrect hypothesis or theory.

possible attempt to *pre-register* the various checks and additional analyses they plan to investigate and/or implement in their study, as well as *the conditions under which* they will perform specific analyses (e.g., see Blair et al. (2019)).

CONCLUSION

Experiments can and do yield null and/or surprisingly weak results. This can often be disappointing for researchers, particularly when a great deal of time and effort has been expended to design and conduct the experiment. This essay emphasizes that the reasons for such results may have little to do with the hypothesis being tested or the underlying theory. Rather, any number of "pitfalls" may be responsible. A key consequence is that any given null result can have multiple explanations, which severely diminishes the theoretical and empirical value of one's study. By becoming aware of these pitfalls, researchers can design their studies to be better safeguarded against, and better equipped to permit investigation into, these pitfalls once results are in-hand (e.g., by including measures of attentiveness (Pitfall #1) and subjective manipulation checks (Pitfall #2)).

Importantly, researchers can conduct these same procedures when the argument is *in favor of* the null—e.g., in "null-by-design" experiments (Druckman 2022, 49). In such experiments, critical readers may question whether the researcher obtained a null result in a fair test of the hypothesis or, conversely, whether the null result was simply due to one or more of the many pitfalls described above. Much like being able to rule out that a null result is simply due to insufficient power, demonstrating that a null finding is not merely due to inattentiveness, nor to a failed manipulation of the independent variable, nor to a ceiling or floor effect, for example, would provide compelling additional evidence in favor of the null hypothesis beyond the non-significant ITT estimate alone (e.g., see Vandeweerdt 2022).

In addition, a number of the aforementioned issues can potentially be studied using pre-tests. When possible, pre-testing various manipulations, measures, and samples would indeed be an ideal for avoiding pitfalls before an experiment is fielded. However, there are two points worth emphasizing. First, resource constraints may prevent some researchers from being able to field a sufficiently-powered study in advance of the primary study. Second, the various techniques outlined here can be equally applied to pre-test data. Pre-test data, in other words, can just as easily be undermined by the various pitfalls discussed above.

It is also worth stressing that even when a researcher *does* find significant ITT estimates, it remains useful to investigate these various pitfalls. Finding a significant effect *combined with* (1) evidence of substantial attentiveness (and perhaps stronger effects among the highly attentive (e.g., see Kane, Velez, and Barabas 2023)), and (2) evidence of successfully varying the independent variable, for example, would serve to bolster the case against the alternative explanation that the significant finding was merely a "chance" result.

Further, finding significant results *despite* substantial inattentiveness, for example, implies that one's ITT estimates are likely an *underestimate* of the actual treatment effect size (e.g., see Gerber and Green 2012, 141–51). An important implication of this point is that early experimental work that assumed pitfalls to be non-existent may have routinely reported underestimates of treatment effects, or erroneously concluded a treatment to be inefficacious simply because of a non-significant *p*-value.

There are, of course, numerous other issues worth investigating in the presence of a statistically significant ITT: confirming covariate balance across groups (Gerber et al. 2014), ensuring the treatment is not confounded (Dafoe, Zhang, and Caughey 2018), determining generalizability to more representative populations (Mullinix et al. 2015), and, if relevant, assessing whether the effect is likely to be observable in real-world, natural settings with more complex information environments (Barabas and Jerit 2010; McDermott 2011).[16] Though immensely important in their own right, such issues are beyond the scope of the present essay. Instead, by highlighting tools and procedures that can be used for rigorously investigating *null* results specifically, the aim of this essay is to assist researchers with getting more out of their experiments than what a naïve, non-significant ITT alone can provide.

REFERENCES

Alrababa'h, Ala' et al. 2022. "Learning from Null Effects: A Bottom-Up Approach." *Political Analysis*: 1–9.

Alvarez, R. Michael, Lonna Rae Atkeson, Ines Levin, and Yimeng Li. 2019. "Paying Attention to Inattentive Survey Respondents." *Political Analysis* 27(2): 145–62.

Ansolabehere, Stephen, Jonathan Rodden, and James M. Jr. Snyder. 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).

[16] In addition, it may be of great practical importance to examine the extent to which significant treatment effects can be recovered long after a treatment is implemented vis-à-vis being largely ephemeral (e.g., see Gerber et al. 2011).

Barabas, Jason, and Jennifer Jerit. 2010. "Are Survey Experiments Externally Valid?" *American Political Science Review* 104(02): 226–42.

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*. SAGE.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *The American Political Science Review* 113(3): 838–59.

Carmines, Edward G., and Richard A. Zeller. 1979. *Reliability and Validity Assessment*. 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*, eds. 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, and Jennifer Jerit. 2014. "Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies." *Journal of Experimental Political Science* 1(2): 120–31.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." *Political Analysis* 26(4): 399–416.

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*. 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., James G. Gimpel, Donald P. Green, and Daron R. Shaw. 2011. “How Large and Long-Lasting Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment.” *American Political Science Review* 105(01): 135–50.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton & Company.

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.

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.

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.

Karlan, Dean, and Jacob Appel. 2016. Failing in the Field *Failing in the Field: What We Can Learn When Field Research Goes Wrong*. Princeton, N.J: Princeton University Press. https://www.degruyter.com/document/doi/10.1515/9781400883615/html (January 3, 2023).

McDermott, Rose. 2011. “Internal and External Validity.” In *Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Greene, James H. Kuklinski, and Arthur Lupia. Cambridge University Press, 27–40.

Meade, Adam W, and S Bartholomew Craig. 2012. “Identifying Careless Responses in Survey Data.” *Psychological methods* 17(3): 437–55.

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.

Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton: Princeton University Press.

———. 2021. “Improving Experimental Treatments in Political Science.” In *Advances in Experimental Political Science*, eds. 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).

Niessen, A. Susan M., Rob R. Meijer, and Jorge N. Tendeiro. 2016. “Detecting Careless Respondents in Web-Based Questionnaires: Which Method to Use?” *Journal of Research in Personality* 63: 1–11.

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): 2600–2606.

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.

Peyton, Kyle, Gregory A. Huber, and Alexander Coppock. 2022. "The Generalizability of Online Experiments Conducted During the COVID-19 Pandemic." *Journal of Experimental Political Science* 9(3): 379–94.

Read, Blair, Lukas Wolters, and Adam J. Berinsky. 2022. "Racing the Clock: Using Response Time as a Proxy for Attentiveness on Self-Administered Surveys." *Political Analysis* 30(4): 550–69.

"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. 2022. "How Dropping Subjects Who Failed Manipulation Checks Can Bias Your Results: An Illustrative Case." *Journal of Experimental Political Science*: 1–7.

Wood, Dustin, P. D. Harms, Graham H. Lowman, and Justin A. DeSimone. 2017. "Response Speed and Response Consistency as Mutually Validating Indicators of Data Quality in Online Samples." *Social Psychological and Personality Science* 8(4): 454–64.