**SUPPLEMENTAL APPENDIX**

**MORE THAN MEETS THE ITT:**

**A GUIDE FOR ANTICIPATING AND**

**INVESTIGATING NON-SIGNIFICANT RESULTS IN SURVEY EXPERIMENTS**

[**APPENDIX A: Al**](#_Toc151889191)[**ternative Types of Attentiveness Measures & Dealing With Inattentiveness in Practice 1**](#_Toc151889192)

**APPENDIX B: Alternative Methods For Examining Treatment Effects A**[**cross Varying Levels of Attentiveness 5**](#_Toc151889195)

**APPENDIX C:** [**Employing A (Subjective) Manipulation Check 6**](#_Toc151889197)

**APPENDIX D:** [**Additional Discussion of Statistical Power 8**](#_Toc151889199)

**APPENDIX E: Additional Discussion of Poor Measurement** [**of the Dependent Variable 11**](#_Toc151889202)

**APPENDIX F: Additional Discussion of Pre-Registreation, Pre-Tests** [**& When To Investigate AEs 14**](#_Toc151889205)

**APPENDIX G:** [**Table 1 Recommendations for Subsequent Study 17**](#_Toc151889207)

[**REFERENCES 25**](#_Toc151889208)

# APPENDIX A:

# ALTERNATIVE TYPES OF ATTENTIVENESS MEASURES & DEALING WITH INATTENTIVENESS IN PRACTICE

Across the social sciences, numerous strategies exist for measuring attentiveness in surveys. 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). However, these attention checks do not typically have response options that *relate to the experimental treatment itself*. Thus, researchers cannot examine correlations between the assignment variable and responses to the attention check.

Given this limitation, employing *factual manipulation checks* (FMCs)—and, in particular, treatment-relevant FMCs—is therefore a useful design choice (see Kane & Barabas 2019). These FMCs explicitly ask about content that is manipulated across conditions as means of measuring attentiveness to the treatment. As there is only one correct response option to the question (given the condition a respondent was assigned to), researchers can test for a significant correlation between treatment assignment and responses to the FMC (e.g., by creating a 2x2 cross-table and conducting a χ2 test).

An alternative to “checks” is the use of question timers (Harden, Sokhey, and Runge 2019; Niessen, Meijer, and Tendeiro 2016; Wood et al. 2017). These are timers placed on survey items (unseen by respondents) that record the amount of time a respondent spends on a particular screen and/or (by combining timers) the survey as a whole. Thus, low amounts of time (“latencies”) indicate insufficient attentiveness. However, again, there is often no reason to suspect *a priori* that the assignment variable will significantly correlate with the latency measure, which limits the researcher’s ability to know whether treatment information was sufficiently perceived by respondents.

Additionally, and in contrast to this timer-based 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.

A final 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). This again results in the researcher having to make more choices regarding how to measure attentiveness in a comparable fashion across respondents.

To supplement the manuscript, Table A1 provides a summary of AE #1 (inattentiveness), the kinds of checks that one can employ in the design stage of the study to measure inattentiveness, specific detection methods that can be employed in the analysis stage, and a technique to employ (using MVCs (see next Appendix)) if substantial inattentiveness is found.

**TABLE A1. Dealing With Alternative Explanation (AE) #1: Inattentiveness**

|  |  |  |  |
| --- | --- | --- | --- |
| **Potential AE** | **Example Checks Available** | **Detection Methods** | **If Problem Is Found…** |
|  |  |  |  |
| *Respondent**Inattentiveness* |  *Factual Manipulation**Checks (FMCs)**Instructional Manipulation Checks (IMCs)**Generic attention check**Mock Vignette Checks (MVCs)* | *a) Examine FMCs, IMCs, MVCs, or other attention checks 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 a treatment-relevant FMC, ensure responses significantly correlate with treatment assignment* |

A wholly alternative method for enabling researchers to potentially rule out inattentiveness as an explanation for non-significant is by including a *positive control* (see Hilgard 2021). In the design stage, the researcher would include a well-replicated study prior to the researcher’s experiment. If, in the analysis stage, the positive control replicates, it indicates that the sample is to some degree attentive—i.e., that the sample is (at least in part) willing and capable of engaging with the contents of an experiment. If, alternatively, the positive control experiment does not replicate, it suggests that inattentiveness—not incorrect theory—underlies the researcher experiment’s non-significant result. A potential challenge of the positive control, however, is that one must assume the positive control does not in any way affect the results of the experiment of interest. Additionally, the results of the positive control may be less informative to the extent that the true effect size of the positive control experiment is substantially larger than the true effect size of researcher’s experiment—i.e., inattentiveness may greatly attenuate both ITTs, but perhaps only the researcher experiment’s ITT will be statistically non-significant because its actual effect is substantially smaller than the positive control’s actual effect.

Lastly, note that the manuscript does not include “randomization checks” (Gerber et al. 2014), otherwise known as “balance tests”. This technique tests 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.

# APPENDIX B:

# ALTERNATIVE METHODS FOR EXAMINING TREATMENT EFFECTS

# ACROSS VARYING LEVELS OF ATTENTIVENESS

The mock vignette check (MVC) approach (Kane, Velez, and Barabas 2023) enables researchers to test for larger effects at higher levels of attentiveness—i.e., among respondents most likely to have attended to and absorbed the treatment. Crucially, the attentiveness is measured *pre-treatment* with factual questions (i.e., only one correct answer) about a non-manipulated vignette. As noted in Appendix A, one can also use instructional manipulation check (IMC) or other attention check performance in the same way (i.e., to estimate whether treatment effects are larger for more attentive respondents), so long as these measures appear pre-treatment.

However, recent research offers additional strategies. One alternative 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). This approach, though, comes with some complexity in terms of both deciding how to measure attentiveness as well as in the precision of CACE estimates (see 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 (see Read, Wolters, and Berinsky 2022).

The instrumental variable and timer approaches, however, feature *post-treatment* measures of attentiveness. Thus, concerns about post-treatment bias may be warranted. Researchers should exercise greater caution when interpreting results should they decide to employ either of these particular techniques (see Montgomery, Nyhan, and Torres 2018).

# APPENDIX C:

# EMPLOYING A (SUBJECTIVE) MANIPULATION CHECK

Manipulation checks (in particular, “subjective” manipulation checks (SMCs)) allow researchers to test whether the experimental manipulation actually varied the independent variable of interest. Without such a check, a non-significant result becomes more difficult to learn from: one potential reason for the result is simply that the experiment did not vary the independent variable it was trying to vary.

To supplement the manuscript, Table C1 provides: (a) a summary of AE #2 (failed manipulation of the independent variable); (b) the check that one can employ in the design stage of the study to measure whether the independent variable was significantly manipulated; (c) a specific detection method that can be employed in the analysis stage; and (d) various techniques to employ if it does not appear that the independent variable was significantly manipulated by the treatment.

The purpose of the final column is to help adjudicate between different possibilities for the failed SMC. As noted in the manuscript, one reason for the failed SMC is because of inattentiveness. A second reason is because the SMC was not a valid and/or reliable measure of the independent variable of interest (see also Appendix E below). A third reason is that the design of the manipulation was not adequate. This might occur, for example, because the treatment information was not sufficiently clear and/or potent: perhaps it was written in a confusing way, or buried within too much non-manipulated text that ultimately undercut the treatment information. Thus, being able to rule out the first two reasons helps to discern that the third reason underlies the failed manipulation check. This discovery obviously comes with important implications for how useful the non-significant findings are, as well as for how the experimental design can be improved upon in future iterations.

**TABLE C1. Dealing With Alternative Explanation (AE) #2: Failed Manipulation of the Independent Variable**

|  |  |  |  |
| --- | --- | --- | --- |
| **Potential AE** | **Check Available for Design Stage** | **Detection Methods****in Analysis Stage** | **If Problem Is Found…** |
| *Failed Manipulation of the Independent Variable* | *Subjective Manipulation Check (SMC)* | *Examine SMC to ensure Control and Treatment groups substantially**differ* | 1. *Check whether the failed manipulation may also be due to inattentiveness (e.g., check if manipulation was effective among the attentive)*
2. *Test for construct validity between other, theoretically-related items and SMC*
3. *Rethink treatment (e.g., for clarity, cognitive overload, and potency)*
4. *Possibly pre-test different versions of treatment to ensure effect on SMC*
 |

# APPENDIX D:

# ADDITIONAL DISCUSSION OF STATISTICAL POWER

Statistical power—defined as 1 minus β, where β is equal to the probability of a *false negative* (failing to reject the null when the null is false)—is something experimentalists often desire to maximize as it reduces risk of error within the null hypothesis significance testing (NHST) paradigm. UCLA (2023) provides a very accessible introduction to statistical power and power analysis.

Power itself is determined by several factors. One that is perhaps in the most direct control of the researcher during the design stage is sample size (*n*) selection. As noted in the manuscript, too small an *n* size will render it more difficult to detect a statistically significant result even if a true non-zero effect exists. Thus it is worthwhile to conduct power analyses *prior to* data collection. This can be done to determine either (1) the necessary sample size (at a given level of power), or (2) the level of power (for a given sample size). Lakens (2022) offers especially practical advice for how to select and justify one’s sample size, as well as clear guidance on deciding a level of power beyond using the conventional level of .80 (which would mean that β=.20=the error rate for failing to reject the null hypothesis when it should be rejected).

Whether the goal is to determine sample size or power, researchers can state a difference of means along with a pooled variance (or standard deviation (SD)) of *Y*. These quantities are often inherently challenging to specify *a priori* with any degree of confidence. Somewhat alternatively, researchers can state an anticipated *effect size*, which is typically a standardized effect size, such as *Cohen’s d* or *r* (see Perugini, Gallucci, and Costantini (2018) for guidance on how to convert between these). This solves the challenge of having to know group means and the SD of *Y*, but still requires specifying an often unknown quantity. A very useful recommendation is to state the smallest possible effect size that would still be of substantive interest (Lakens 2022). This is a conservative choice insofar as it will effectively require the largest sample size possible (at a given power level). In other words, smaller effects will require larger samples in order to detect an effect at a given significance threshold (α).

Thus, if researchers cannot use existing scholarship (e.g., single studies or meta-analyses) or pre-test data to specify an anticipated effect size (see Lakens 2022), they may opt to specify a very small (yet still substantively meaningful) effect size in order to determine the sample size needed. Perugini et al. (2018, Table 1), for example, describes a Cohen’s *d* of .20 (which converts to an *r* value of .10) as small. Such figures may serve as a useful starting point for researchers thinking about the smallest meaningful effect size, though Lakens (2022) stresses the impracticality of trying to establish any such benchmark given the heterogeneity in typical effect sizes across fields.

As stated in the manuscript, the decision regarding sample size or power should also be made knowing number of conditions one plans to employ, as well as any subgroup and/or interactive analyses that will be conducted. Ideally, and to allow such analyses to be maximally informative, one should be determining if power and/or sample size is sufficient not just for the main analysis, but for each of the subgroup analyses as well. (Perugini et al. (2018) also offer specific guidance on how conduct power analysis for regression analyses with interactions.) The researcher should also be mindful that any anticipated effect sizes may well be smaller to the extent that a sample is inattentive.

Conducting *post hoc* power analyses to determine statistical power is not viewed as especially useful (Lakens 2022; Perugini, Gallucci, and Costantini 2018). However, once the ITT and pooled SD are estimated in one’s experiment, a power analysis may potentially be helpful for determining whether one’s sample size amounted to an under-powered study—yet another useful piece of information when trying to determine whether a non-significant result is informative (though note that such an analysis is a bit circular insofar as it is based upon the observed ITT). Lastly, power/sample size analysis can be done using free resources like G\*Power[[1]](#footnote-1) or with packages in common statistical software (e.g., the *pwr* or *pwrss*[[2]](#footnote-2) package in *R*; the *power* suite of commands in Stata).

Beyond investigating whether a sample size was sufficient, the analysis stage also allows for improving power via including covariates in one’s model—specifically, improving the precision of the ITT estimate (see Bowers 2011, 460–61). Nevertheless, considerations regarding statistical power—and efforts to improve power in one’s experiment—should largely occur in the design, rather than analysis, stage of an experimental study.

# APPENDIX E:

# ADDITIONAL DISCUSSION OF POOR MEASUREMENT

# OF THE DEPENDENT VARIABLE

Among the many challenges of survey experiments, the survey questions that researchers employ may, in fact, be poor measures of the underlying concept of interest. Hence, poor measurement of the dependent variable (*AE#5*) may also be responsible for erroneous non-significant results.

This section extends the discussion appearing in the manuscript to provide researchers with practical advice for avoiding (in the design stage) and investigating (in the analysis stage) poor measurement of *Y*. Ideally, researchers would not have to construct their own outcome measures. Within political science research, a large number of relevant outcomes—e.g., vote choice, feelings toward candidates or other political targets, and policy attitudes—likely have validated measures already in existence, or at least very similar measures that require only a small amount of adaptation. Such measures might come from published studies or large publicly available data sets (e.g., the American National Election Study (ANES), Congressional Election Study, or Democracy Fund Voter Study Group). Beyond these, tools for locating survey questions *by topic*, across a wide variety of polling firms, are readily available online (e.g., Roper iPoll[[3]](#footnote-3)). For psychological outcomes, the American Psychological Association has an online repository (“PsycTests”) of instruments for various psychological constructs.[[4]](#footnote-4) A wide range of measures of personality and other individual differences can be found at International Personality Item Pool (IPIP).[[5]](#footnote-5) The key advantage of using such measures is that they can be validated prior to fielding one’s study (see points below) and incorporated into one’s own study, relieving the researcher of the need to develop brand new questions and corresponding response options.

When researchers need to construct their own outcome measure, several pieces of advice can improve the quality of the measure. As Druckman (2022, 21) recommends, the reliability and accuracy of a measure that involves a scale (e.g., a 7-point scale measuring opposition versus support for a policy) can be greatly improved when the scale points are individually labeled. By omitting labels for all but the extreme values, for example, respondents must now guess at what each point on the scale substantively means. This “guessing” becomes a source of statistical noise, raising the risk of a non-significant finding. Druckman also recommends the use of multiple items that can then be formed into a scale (e.g., researchers can take an average of similarly scaled items, combine all items into a single additive scale, or factor-analyze all of the items to generate a single factor score for each respondent). Again, the aim is to reduce statistical noise: a respondent may misinterpret a single outcome measure (either the question wording and/or the response options), but with multiple measures, the possibility that a given respondent is misinterpreting *all* items is low.

Regardless of the measure of *Y* that is ultimately chosen, a non-significant ITT opens the possibility that the measure of *Y* was invalid and/or unreliable/noisy. In such cases, a researcher would ideally want to investigate the construct validity of *Y* (*Y*’s degree of correlation with another measure that is theoretically related to *Y*). Ideally, the researcher could also investigate the criterion validity (*Y*’s degree of positive correlation with an accepted alternative measure of *Y*) and/or convergent validity (*Y*’s degree of positive correlation with a set of alternative measures of *Y*; see McDonald (2005) for additional details). Criterion and convergent validity may not be possible to investigate within one’s own survey, however (doing so would imply that the researcher already has valid measures of *Y*). However, if the measure is being adapted from existing publicly available data, it may be possible to investigate all forms of validity to determine whether a particular measure of *Y* is useful to include in one’s own study.

Thus, in one’s own survey experiment, it is essential in the design stage to include covariates that are theoretically relevant to *Y*. So long as these are measured pre-treatment, their inclusion allows for an unbiased test of construct validity in the analysis stage. If a non-significant ITT is found, yet *Y* correlates substantially with theoretically-relevant items, and in the correct direction, it suggests that an invalid and/or overly noisy measure of *Y* cannot explain the non-significant finding—a crucial insight that better enables the researcher to determine the theoretical and empirical value of their study.

# APPENDIX F:

# ADDITIONAL DISCUSSION OF PRE-REGISTRATION, PRE-TESTS,

# & WHEN TO INVESTIGATE AEs

Each of the alternative explanations (AEs) detailed in the manuscript can potentially undermine one’s survey experiment in the form of yielding (erroneously) non-significant results. Ideally, researchers should anticipate—and guard against—these AEs when *designing* their study. When this has not occurred, researchers can still use some of the techniques outlined in the manuscript to investigate the robustness of their non-significant findings, though likely to a far lesser extent than if the study had been designed with AEs in mind. That said, a number of points are worth keeping in mind.

## *Pre-Registration and Pretesting*

Obviously, the investigation of these various AEs can require additional analyses (e.g., estimating an ITT among a subgroup who performed well on a pre-treatment measure of attentiveness or who is unlikely to have been pre-treated). Yet, with a greater number of analyses being conducted, the risk of a *false positive* naturally increases. Thus, at a minimum, researchers should *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 these analyses.

Finally, as noted in the manuscript, a number of the aforementioned AEs can potentially be studied using pre-tests. When possible, pre-testing various manipulations, measures, and samples would indeed be ideal for avoiding AEs before an experiment is fielded. However, resource constraints may prevent researchers from being able to field a sufficiently-powered study in advance of the primary study. As such, it is understandable that a pre-test—though ideal—may not always be feasible. Second, it is worth stressing the various AEs and techniques outlined here *can apply equally to pre-test data*. Pre-test data, in other words, should not be considered an alternative to investigating AEs; indeed, pre-test data can just as easily be confounded by the same AEs discussed in the manuscript.

## *When to Investigate AEs?*

Non-significant results are, generally speaking, likely an undesired outcome in most research projects. In such cases, a researcher should be interested in designing their study so as to be able to more explicitly investigate AEs as it will help adjudicate between the hypothesis and/or theory being incorrect (one explanation) vis-à-vis one (or more) of the seven AEs identified in the manuscript.

As noted in the manuscript, however, researchers can conduct these same procedures when arguing *in favor of* a non-significant result—e.g., in “null-by-design” experiments (Druckman 2022, 49).[[6]](#footnote-6) Demonstrating that a non-significant finding is unlikely to be due to insufficient power (*AE#4*) or attentiveness (*AE#1*), nor to a failed manipulation of the independent variable (*AE#2*), for example, provides more compelling evidence that a treatment has no meaningful effect; evidence that is substantially more robust than if one merely reports the ITT and/or its corresponding *p*-value alone.

Further, researchers can even employ these procedures in the presence of *significant* results. Finding, for example, a significant effect *combined with* evidence of stronger effects among the attentive (*AE#1*) *and* successful manipulation of the independent variable (*AE #2*), strengthens the case against a significant finding being a spurious result. Along similar lines, finding significant results *despite* encountering AEs (e.g., substantial inattentiveness) implies that one’s ITT estimate is likely an *underestimate* of the actual treatment effect size (e.g., see Gerber and Green 2012, 141–51). An important implication of this latter point is that early experimental work that (in effect) assumed all AEs to be non-existent may have routinely reported underestimates of treatment effects and/or erroneously concluded a treatment to have “no effect”.

# APPENDIX G:

# TABLE 1 RECOMMENDATIONS FOR SUBSEQUENT STUDY

The final column of Table 1 in the manuscript provides recommendations for researchers who wish to field a subsequent study after (1) discovering one or more alternative explanations (AEs) in their initial study, and (2) being unable confidently determine, with the data at hand, the extent to which a given AE is responsible for a non-significant finding. For example, a researcher might discover that inattentiveness in the sample is so substantial that it leaves too few (attentive) respondents to meaningfully test for a significant ITT. Or, a researcher might discover substantial evidence that a non-significant result is due to a “ceiling effect.” Upon learning this, the researcher may desire to redesign the study so as to conduct an improved test of the hypothesis on a new sample.

This section provides some additional elaboration on the recommendations listed in Table 1. Before proceeding, however, it is important to emphasize that *these recommendations assume the researcher has already implemented the “Recommended Practices in Design Stage”* (see the second column of Table 1). If these were not implemented in the initial study, then naturally any subsequent study should first consider these recommendations. Second, it is worth reiterating that (if possible) pre-testing is a useful option before fielding a subsequent study. For example, pre-testing can be used to determine whether a revised treatment is better able to pass a manipulation check (AE #2), whether a sample continues to show evidence of pre-treatment (AE #3), or which measure(s) of *Y* have the highest reliability and best construct validity (AE #5).

## *AE #1: Inattentiveness*

Beginning with AE #1 (inattentiveness), a subsequent study should aim to garner a more attentive sample, particularly a sample that is more attentive to the experimental content itself. To accomplish this goal *within their survey*, researchers can employ several strategies. For example, the experiment of interest can be moved to an earlier point in the study to reduce the possibility of respondent fatigue and, thus, inattentiveness to the contents of the survey experiment. Along similar lines, the treatment content can be made more salient by, for example, including it earlier in a vignette’s text, repeating it throughout the vignette, and featuring it in bolded, italicized, or otherwise more noticeable text. Researchers can also make explicit attempts, prior to the randomly assigned vignette, to prompt respondents to be attentive. This could take the form of gentle requests (e.g., “Please read the following information carefully.”) or more forceful warnings. For example, Clifford and Jerit (2015, 798) found that an “audit” message (which warned respondents that participants would only be accepted if they “clearly demonstrate that they have read and understood the survey”, and subsequently asked respondents to confirm that they understand) to be a consistent method for increasing respondent attentiveness. However, as the authors note, such warnings may also risk inducing socially desirable responding. As such, researchers should exercise caution when attempting to motivate respondents’ attentiveness, particularly for outcomes that are susceptible to social desirability bias.

Outside of what can be accomplished *within* the survey, researchers may opt to choose an alternative survey company to field the survey (i.e., one that is expected to perform better in terms of providing attentive respondents). In addition to this (or perhaps alternatively, if the initial survey firm is used again), researchers may impose more stringent screening-out procedures at the start of the survey (prior to random assignment) to ensure that respondents who complete the experiment have performed satisfactorily on one or more measures of attention.

## *AE #2: Failure to Vary the Independent Variable*

As discussed in the manuscript, the standard approach to determining whether the independent variable has been effectively manipulated is via a manipulation check (specifically, a “subjective manipulation check” (see Kane and Barabas 2019)). If the initial study yielded a “failed” manipulation check, the most obvious recommendation for the subsequent study is to make the treatment content both more salient (see above sub-section) and/or more potent (i.e., stronger). Salience can be increased by having the treatment appear sooner, more frequently, and/or more noticeably. Separately, the potency (i.e., strength) of treatment content—when it appears—can be increased in a variety of ways—e.g., using stronger, more direct, and more compelling language or imagery; having the content be attributed to a more credible, authoritative source, etc. When experimental treatments have low salience and/or potency, respondents may not only fail to attend to the treatment (see AE #1) but, as a separate matter—and per the present AE—fail to process and absorb the treatment as it was intended. Thus, for a subsequent study, researchers should aim to ensure that the treatment is not only more salient (i.e., more difficult for respondents to overlook), but also more potent in terms of how compelling respondents are likely to find it. (Of course, researchers should consider whether the stronger treatment means that the experiment will have lower external validity, and whether this is problematic for the aims of the study (see Druckman 2022, ch.3.) Again, pretests can serve as a useful way for researchers to construct and test a variety of possible versions of a treatment before fielding the subsequent study.

Naturally, there are other factors that could be driving a failed manipulation check that have nothing to do with how researchers have designed the treatment. A floor or ceiling effect might be the cause of a failed manipulation (i.e., the manipulation check item has too low, or too high (respectively), a value in the control group). Similarly, a manipulation check item with low measurement validity and/or reliability, low statistical power, and a pre-treatment effect could all plausibly yield a non-significant manipulation check and, thus, should also be considered before fielding a subsequent study.

## *AE #3: Pre-Treatment Effect*

 Assuming that a non-significant ITT is unlikely to be due to a floor or ceiling effect, and assuming the researcher finds evidence for a pre-treatment effect in their initial study (e.g., large ITT among those least likely to have been pre-treated), researchers can potentially better guard their *subsequent* study against pre-treatment via implementing a stronger treatment. If the treatment in the experiment, in other words, is comparable to how respondents tended to experience the treatment “in the real world” prior to the experiment, then we should expect a non-significant ITT: the treatment, as it appears in the experiment, was already “absorbed” prior to the experiment. However, if the treatment takes the form of a “stronger dose” in the experiment (relative to what was experienced in the real world), then it is reasonable to suspect that it will exert an *additional* influence upon respondents’ beliefs, attitudes, and/or behaviors beyond what was already absorbed in the real world.

However, there may specific contexts in which the risk and/or strength of pre-treatment is so great that the experiment—regardless of how the treatment is designed—is unlikely to substantially change respondents’ beliefs, attitudes, or behaviors. In such circumstances, researchers might consider postponing the experiment until the context has changed—i.e., until the risk of pre-treatment has subsided to a substantial degree.

Of course, if the source of pre-treatment is unknown, a survey experiment on the topic may not be possible at all. Along the similar lines, researchers might investigate the feasibility of testing their hypothesis via non-experimental means (e.g., a cross-sectional or difference-in-difference analysis using publicly available survey data).

## *AE #4: Statistical Power*

If the initial study suffers from low statistical power, any subsequent study should of course aim to increase the sample size substantially and/or reduce the number of experimental conditions. Fortunately, researchers can use their initial study to determine a variety of important quantities: the effect size, the amount of variation in *Y*, the amount of respondent inattentiveness, and the final number of experimental conditions and subgroup analyses. This can give the researcher greater insight into what total sample size is necessary to conduct all analyses with sufficient power, including those with only a subset of respondents, some share of which will likely be inattentive. For example, whatever the required sample size to conduct an analysis among a particular subgroup at power=.80, the researcher might instead aim to have a sample size—for that particular analysis—equal to the required sample divided by (1 – (proportion inattentive)).[[7]](#footnote-7)

Somewhat alternatively, the structure of the design can be fundamentally changed. For example, researchers might consider a *within*-subjects experimental design rather than a between-subjects design, thereby substantially increasing statistical power via having multiple observations per respondent (see Clifford, Sheagley, and Piston 2021).

## *AE #5: Poor Measurement of Y*

If the initial experiment’s measure of *Y* is found to have low construct validity, then naturally any subsequent study should aim to use an improved measure of *Y*. This can potentially be accomplished by locating a measure that has been validated elsewhere (e.g., in an existing study or in publicly available survey data from a large polling firm). In addition, the researcher should consider using multiple measures of *Y,* particularly if the initial study did not already do so.

Again, pre-testing can assist researchers with examining various measurement properties of potential measures of *Y* before the subsequent study. Researchers can examine, for example, how well each measure correlates with other items that, theoretically, it should correlate with (i.e., construct validity), as well as how well multiple measures scale together. Such a process can help ensure that the subsequent study ultimately features a more valid, informative, and reliable measure of *Y*.

## *AE #6: Ceiling / Floor Effect*

When the initial study finds substantial evidence for a ceiling or floor effect, it is reasonable to assume that, if the same population is targeted in the *subsequent* study, the experiment will continue to be at risk of encountering the same ceiling or floor effect. (Of course, if the population of interest changes, or the researchers anticipate a demographically dissimilar sample in the subsequent study, then a ceiling or floor effect may not be as great of a concern.) Thus, assuming that the population of interest remains constant, researchers should aim in the subsequent study to employ a measure of *Y* with a more conceptually extreme range. As a simple example, if an item is a five-point scale ranging from “Disagree” to “Agree”, this could be expanded to a seven-point scale in which the two extreme values are “Completely Disagree” and “Completely Agree.” The logic is that only a fraction of control group respondents who, for example, would have selected “Agree” in the initial study will opt to select “Completely Agree” in the subsequent study, thus bringing the mean of *Y* down relative to its maximum possible value. Similarly, if using multiple items to create a scale, researchers might consider including an additional item for which control respondents are less likely to select a high value (if concerned about a ceiling effect) or low value (if concerned about a floor effect).

An alternative factor that might result in a ceiling or floor in the initial study, of course, is that of the particular context in which the initial study was fielded. It may prove difficult to increase respondents’ interest in a particular issue, for example, if that issue is highly salient in mass media during the time of the study and respondents are already extremely interested in it before beginning the experiment. In such cases, researchers might consider postponing the study until the context changes and the salience of the issue subsides. Pre-tests can potentially assist researchers with determining when the context has changed sufficiently, and also with assessing whether changes to the measure of *Y* are effective in lowering or raising the mean among control group respondents (relative to what was observed in the initial study).

## *AE #7: Countervailing Treatment Effects*

If researchers discover a countervailing treatment effect in the initial study—i.e., that a moderating variable (*M*) yields treatment effects in opposite directions, such that the overall ITT is non-significant—researchers might take the following steps should they decide to field a subsequent study. First, the researchers should aim to pre-register a hypothesized interaction. In other words, researchers should transparently declare that, based upon the results of the initial study and/or any relevant theory, they anticipate an interaction between treatment assignment and *M*.

Second, knowing that an interaction will be specified, researchers should attempt to include the best possible (pre-treatment) measure(s) of *M*. When *M* is a common categorical variable (e.g., party identification, racial identification, etc.) locating a single measure that is valid and reliable is reasonably simple. However, when *M* is a more complicated construct, and particularly when it is continuous in nature, more care should be given to identifying *multiple* valid and reliable measures. Echoing the discussion of *AE #5* above, an overly noisy measure of *M* will make it more difficult to reliably detect a significant countervailing effect, especially insofar as specifying an interaction is, by its very nature, already adding multicollinearity to one’s analysis, increasing the probability of a Type II error.

Lastly, beyond merely pre-registering such an analysis and featuring valid and reliable measures of *M*, researchers also might consider altering the design so as to better understand *why* the countervailing effect is occurring. For example, imagine that a binary measure of gender shows evidence of a countervailing effect, such that men and women in the sample exhibit treatment effects in opposite directions. The researchers might consider including (pre-treatment) measures that can help reveal, for example, whether such countervailing effects are due to gender differences in knowledge, differences in lived experience or socialization, differences in specific attitudes, or some other kind of difference. In other words, while the initial study revealed that gender plays a moderating role, a subsequent study can potentially go further by helping to shed light upon the underlying reason(s) for this moderating effect.[[8]](#footnote-8)

# REFERENCES

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

Aronow, P. M., Joshua Kalla, Lilla Orr, and John Ternovski. 2020. “Evidence of Rising Rates of Inattentiveness on Lucid in 2020.” SocArXiv. https://doi.org/10.31235/osf.io/8sbe4.

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.

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

Bowers, Jake. 2011. “Making Effects Manifest in Randomized Experiments.” In *Cambridge Handbook of Experimental Political Science*, edited by James N. Druckman, Donald P. Greene, and James H. Kuklinski, 1st ed., 459–80. New York, NY: Cambridge University Press.

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.

———. 2015. “Do Attempts to Improve Respondent Attention Increase Social Desirability Bias?” *Public Opinion Quarterly* 79 (3): 790–802.

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.

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.

Gerber, Alan, Kevin Arceneaux, Cheryl Boudreau, Conor Dowling, Sunshine Hillygus, Thomas Palfrey, Daniel R. Biggers, and David J. Hendry. 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.

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.

Hilgard, Joseph. 2021. “Maximal Positive Controls: A Method for Estimating the Largest Plausible Effect Size.” *Journal of Experimental Social Psychology* 93 (March): 104082.

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. “Sample Size Justification.” *Collabra: Psychology*. https://doi.org/10.31234/osf.io/9d3yf.

McDonald, Michael P. 2005. “Validity, Data Sources.” In *Encyclopedia of Social Measurement*, edited by Kimberly Kempf-Leonard, 939–48. New York: Elsevier. https://doi.org/10.1016/B0-12-369398-5/00046-3.

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.

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 (August): 1–11.

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.

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.

UCLA. 2023. “Introduction to Power Analysis.” UCLA Statistical Methods and Data Analytics. 2023. https://stats.oarc.ucla.edu/other/mult-pkg/seminars/intro-power/.

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.

1. See <https://stats.oarc.ucla.edu/other/gpower/>. [↑](#footnote-ref-1)
2. This is also freely available as an online application for non-R users: <https://cran.r-project.org/web/packages/pwrss/vignettes/examples.html> . An additional resource online for calculating necessary sample sizes given a particular effect size is: https://sample-size.net/sample-size-means/ [↑](#footnote-ref-2)
3. Located at https://ropercenter.cornell.edu/ipoll/ [↑](#footnote-ref-3)
4. Located at https://www.apa.org/pubs/databases/psyctests [↑](#footnote-ref-4)
5. Located at https://ipip.ori.org [↑](#footnote-ref-5)
6. See Rainey (2014) for excellent additional guidance on arguing for null/weak effects. [↑](#footnote-ref-6)
7. For example, if an analysis among a particular subgroup requires *n*=250 in order to have power=.80, yet the initial study found 30% of respondents to be inattentive (using a pre-treatment attention check), then the researcher might instead aim to have a sample size equal to $250 ÷\left(1-0.30\right)=357$ respondents for that particular analysis in the subsequent study. [↑](#footnote-ref-7)
8. Of course, specifying an interaction will tend to diminish statistical power, which means that researchers may need to adjust accordingly when calculating their required sample size (Perugini, Gallucci, and Costantini 2018, 12–15). [↑](#footnote-ref-8)