Journal of Experimental Political Science (2024), **11**, 147–161
doi:10.1017/XPS.2023.22

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

A Practical Guide to Dealing with Attrition in Political Science Experiments

Adeline Lo, Jonathan Renshon and Lotem Bassan-Nygate

Department of Political Science, University of Wisconsin-Madison, Madison, WI, USA
Corresponding author: Adeline Lo; Email: aylo@wisc.edu

Abstract

Despite admonitions to address attrition in experiments – missingness on Y – alongside best practices designed to encourage transparency, most political science researchers all but ignore it. A quantitative literature search of this journal – where we would expect to find the most conscientious reporting of attrition – shows low rates of discussion of the issue. We suspect that there is confusion on the link between *when* attrition occurs and the type of validity it threatens when present, and limited connection to and guidance on which estimands are threatened by different attrition patterns. This is all exacerbated by limited tools to identify, investigate, and report patterns attrition. We offer the **R** package – attritevis – to visualize attrition over time, by intervention, and include a step-by-step guide to identifying and addressing attrition that balances post hoc analytical tools with guidance for revising designs to ameliorate problematic attrition.

Keywords: experiments; attrition; missingness; visualization

The use of experimental methods combined with online samples, already popular in political science, has accelerated in recent years. Yet, even while a cottage industry has developed around subjecting our methods to empirical scrutiny (Brutger et al. 2022; Coppock 2019; Dafoe, Zhang, and Caughey 2018; Jerit, Barabas, and Clifford 2013; Kertzer 2020; Mullinix et al. 2015; Mummolo and Peterson 2019), the issue of attrition – missingness on the outcome – has been mostly ignored by practitioners (Gerber et al. 2014; Zhou and Fishbach 2016). Psychology has already faced a reckoning regarding attrition (Zhou and Fishbach 2016, 495), and our reading of the state of the field in experimental political science is that we do not fare much better.

This article has earned badges for transparent research practices: Open Data and Open Materials. For details see the Data Availability Statement.

Data and methods described in this paper can be accessed at https://github.com/lbassan/attritevis, along with a paired vignette. Michael Hatfield provided excellent research assistance, and we thank the feedback from participants at MPSA, APSA, and JEPS anonymous reviewers. All errors remain our own.

© The Author(s), 2023. Published by Cambridge University Press on behalf of American Political Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press

This is true for attrition as "missingness on the outcome," and even more so for an expansive definition that we use here which includes pretreatment respondent drop-off.

While standards of best practices have been set (Gerber et al. 2014, 97), habits are hard to change: attrition is rarely inspected or discussed, despite the known issues caused if the missingness correlates with potential outcomes of those who drop out of the study (Druckman et al. 2011, 19). Helpful advances have been made (e.g., Coppock et al. 2017), but they focus almost exclusively on *ex-post* solutions such as double sampling or extreme value bounds, which, though valuable, do not help with the issue of easily *identifying* attrition that results in threats to inference or *preventing* it from occurring in the first place in the design stage.

Combining the dictum of Fisher to "analyze as you randomize" with the advice of Coppock (2021) to "visualize as you randomize," our contribution is to offer experimentalists a way to do both. Specifically, we provide a "holistic approach" to addressing attrition, beginning with a quantitative literature search to illuminate the scope of the problem. Our search of all articles published in *Journal of Experimental Political Science* yields discouraging results: in the journal where we would most expect to see systematic and transparent discussions of attrition, 60% of the empirical articles published contained no mention at all.

Our contribution centers on an **R** package – attritevis – that provides diagnostic visualizations and corresponding tests for investigating and addressing attrition. A central output is an "attrition over time" plot that provides a question-by-question, over-time snapshot of an experiment along an axis with the corresponding amount of attrition at each of these moments, across treatment conditions. This is paired with a respondent-level visualization by treatment condition for detailed inspection of patterns of attrition. The package and the guidance in this article are designed around three central questions researchers ought to ask: (1) is there attrition (if so, where)? (2) what kinds of threats to inference are there? If there are threats, (3) what adjustments can be made to account for them, or to preempt them in future studies? Our goal is to provide guidelines and descriptive statistics to help researchers pinpoint the nature and scope of attrition to alter experimental designs and minimize or preempt problems in future studies. In service of that goal, we connect patterns of nonresponse in experiments to more general concerns about internal and external validity, while still focusing on what estimands remain recoverable even when there is heavy attrition.

The scope of the problem

Several patterns are evident from a brief review of the literature. First, attrition can (in theory) pose problems for inference and extant work suggests that (in practice) our fears may be justified, even when considering only reported attrition (presumably lower than actual attrition, Musch and Reips 2000; Zhou and Fishbach 2016). Second, despite the importance of detecting attrition, it appears as though "ignorance is bliss" for most researchers: in a systematic quantitative literature search within political science, Gerber et al. (2014, 88) find that 58% of sampled experimental articles did not report subject size in each treatment group for which there is missing outcome data (see also Mutz and Pemantle 2015, 13 and Zhou and Fishbach 2016, 495). Finally, the

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press

options typically suggested to address attrition are inadequate on their own since they typically focus solely on *ex-post* solutions.

To set the stage and update previous reviews of the problem, we coded every experimental article published in *Journal of Experimental Political Science* from its inception in 2014 (Volume 1) to 2021 (Volume 8). The resulting population consisted of 131 articles (our unit of analysis).[1] As the flagship journal for experimental studies of politics (which has published reporting standards recommending best practices for discussing attrition, Gerber et al. 2014), JEPS is the most likely place to see evidence of scholars taking the problem seriously.[2]

Results of our quantitative literature search – presented graphically in Figure 1 with each square representing a single article – suggest there is considerable room for improvement (and we suspect the patterns we observe would be significantly worse in other journals). First, we find that the modal experimental paper – 60% (78 papers, in gray) – published in JEPS contains *no mention of attrition* (this worsens when setting aside the 8% of the papers studying attrition directly). Second, of the 40% that do mention attrition, nearly half – 17% of the total – note no attrition in their studies, suggesting the possibility of an adverse selection problem whereby attrition is only mentioned if there is no evidence of it being a problem. Finally, for papers transparent about attrition occurring (33 papers, or 25% of the total), only three clearly analyze and account for attrition.[3]

Of course, respondent "cost to attrite" varies across study type: these costs are high in lab experiments, for example, which typically feature zero attrition (some types of field experiments may pose similar costs). To further analyze the data, we distinguished between survey and non-survey (e.g., field or lab) experiments and find that of the 78 papers to not mention attrition and 58 were survey experiments (and only $\frac{14}{78}$ were lab and $\frac{6}{78}$ were field experiments). Though well-known approaches exist to address the issue, our argument is that attrition is considered less often than it ought to be – even where professional and cultural incentives are strong – and that one significant stumbling block for scholars is a clear, easy way to determine if and when attrition is occurring in their studies.

Current approaches to addressing attrition

Current solutions cluster around two approaches: reducing attrition in the design stage and addressing it *post hoc* (i.e., analytically). The first camp asks the practical question of what might reduce attrition, testing and evaluating specific ideas such as using different survey modes (Morrison et al. 1997), or appealing to respondents' conscience (Zhou and Fishbach 2016). Another popular approach utilizes monetary incentives (Göritz 2014), with some research (Castiglioni, Pforr, and Krieger 2008) suggesting the utility of conditional incentives. Other work has focused on question length and relevance (McCambridge et al. 2011) as well as adding "warm-up" tasks. Warm-up tasks are intended to increase respondents' "sunk costs" (Horton, Rand,

[1]Many articles contained multiple studies, so using article – not study – in practice translates to relative leniency in our coding: if any studies in a paper discussed attrition, that counts.

[2]In fact, current practices at JEPS ask authors to be explicit about attrition.

[3]Callen et al. (2016), Boas (2016), Boas (2016) and Green and Zelizer (2017) are notable examples that explicitly discuss, quantify and adjust for attrition.

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press

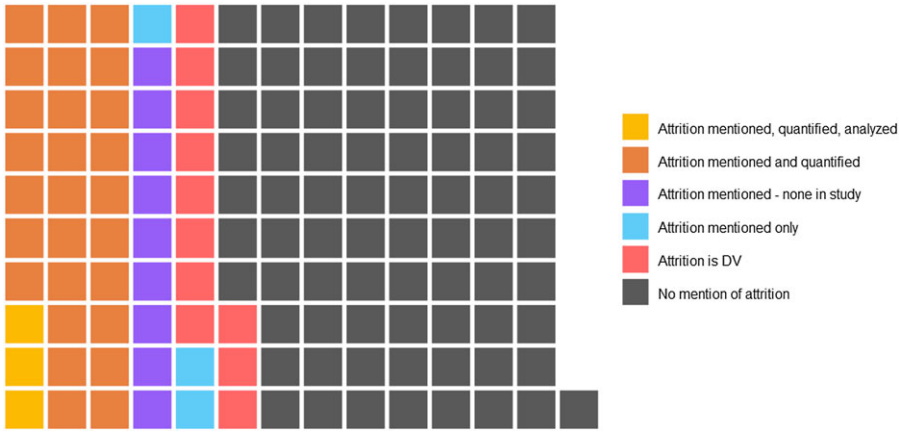


Figure 1. Experimental papers in full *JEPS* corpus and their discussion of attrition.

and Zeckhauser 2011), and Reips (2000) finds some evidence that they work in online survey contexts.

While the bulk of the research on reducing attrition comes from panel/longitudinal settings (Lynn 2018), it still points to some generalizable lessons – e.g., the utility of incentives and the importance of debriefing questions. However, the fatal flaw in focusing solely on preempting attrition by design is that one never knows how successful the effort has been in the particular context in which it was used. That is, utilizing design choices – whether cash payments or subtle "nudges" – to minimize attrition does not obviate the necessity of having a way to understand when and why attrition is occurring. After all, using incentives to lower attrition from some counterfactual baseline does not solve one's inferential problems if it still occurs and is causally related to treatments.

Other approaches address attrition *ex-post*, such as through the use of extreme value bounds ("Manski Bounds," see Manski 1995, 2009 and Coppock 2021, 333) or inverse probability weighting ("IPW," Wooldridge 2007), sometimes visualized to highlight the difference between observed data and imputed best/worst-case scenarios (as in Coppock 2021, 333). These bounds assume that all attriters exposed to treatment would have had the highest possible outcome level while all attriters in control would have had the lowest values on the outcome (a related approach requires a stronger assumption that treatment only affects attrition in one direction, Lee 2009). Another approach involves double sampling (or "refreshment samples" in panel studies, Deng et al. 2013) and is designed to address attrition through randomized followups among attrited respondents. Recently, some advances suggest combining multiple approaches (Coppock et al. 2017; Gomila and Clark 2020), allowing researchers to partially identify the ATE of an experiment even when assumptions about missingness in follow-up contact attempts are relaxed.[4]

[4]Early approaches to double-sampling relied on an assumption of "no missingness among follow-up subjects" to point-identify the ATE. See Hansen and Hurwitz (1946), Kaufman and King (1973) and discussion in Coppock et al. (2017, 189).

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press

In sum, attrition represents a significant threat to inference and current approaches to addressing it are either *ex-post* – requiring strong assumptions in order to estimate treatment effects – or involve implementing design choices that rely on confusing conventional wisdom about what "works." Moreover, Gomila and Clark (2020, 1) highlight a more fundamental problem by noting that one set of solutions (IPW) is suggested for "mild" attrition while another (double-sampling) is suggested for "severe" cases. Our question is: *how are experimentalists to know the difference?*

Our argument is that whether one wishes to preempt attrition through design choices or address it in the analysis stage, researchers would benefit from a way to understand when and why attrition is occurring in the first place. Our proposed set of solutions, detailed below, aid in the design stage by allowing researchers to pinpoint and design around problematic questions and treatments – heading off attrition in studies before they are fielded – and in the analysis stage by providing an understandable method for understanding when attrition is occurring and when it poses threats to inference.

Diagnosing attrition in your experiment

The typical approach to identifying or addressing attrition in an experiment focuses on its *levels*. Researchers ask *if* attrition occurred, or in rarer cases, *how much* attrition is present, calculating the number of respondents who finished the study as a proportion of the total. We argue that a more informative way to think about attrition in experiments focuses attention on *when* it occurs and *what* implications it has for the recovery of causal estimands we are interested in and the assumptions upon which they rest. The following sections outline a series of practical questions and steps researchers can take – facilitated by attritevis – to address attrition, broadly summarized in Figure 2.

Overview

Our temporal approach enables researchers to visualize the attrition that occurs throughout their study by treatment arm. The first step is to load one's data, ensuring that columns in the dataframe are ordered by occurrence, and to specify key moments in the study (e.g., treatment and DV, mediators).[5] Following that, researchers can use plot_attrition to to create the "Attrition timeline" plot, which highlights variation, by intervention arm, in the over-time levels of attrition as well as its relationship to critical moments in the study.

The x-axis in the timeline plot represents all items in the experiment in the order in which they occurred while the y-axis indicates either *attrited count* (how many respondents attrite at each question) or *proportion attrited*[6]. The first outcome is helpful for detecting whether large (absolute) numbers of respondents drop out of

[5]Some survey vendors will allow researchers to use their own Qualtrics instrumentation, in which case all the advice below applies. For others, e.g., YouGov, researchers typically receive only "completes." In these cases, we recommend researchers negotiate *ex ante* with the company to provide at least basic information regarding subjects who started – but did not complete – studies. In a perfect world, researchers would receive the complete dataset including attritors, but any information provided would go a long ways towards addressing the inferential questions we discuss below.

[6]Both are options in plot_attrition, as is missingness by question.

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press

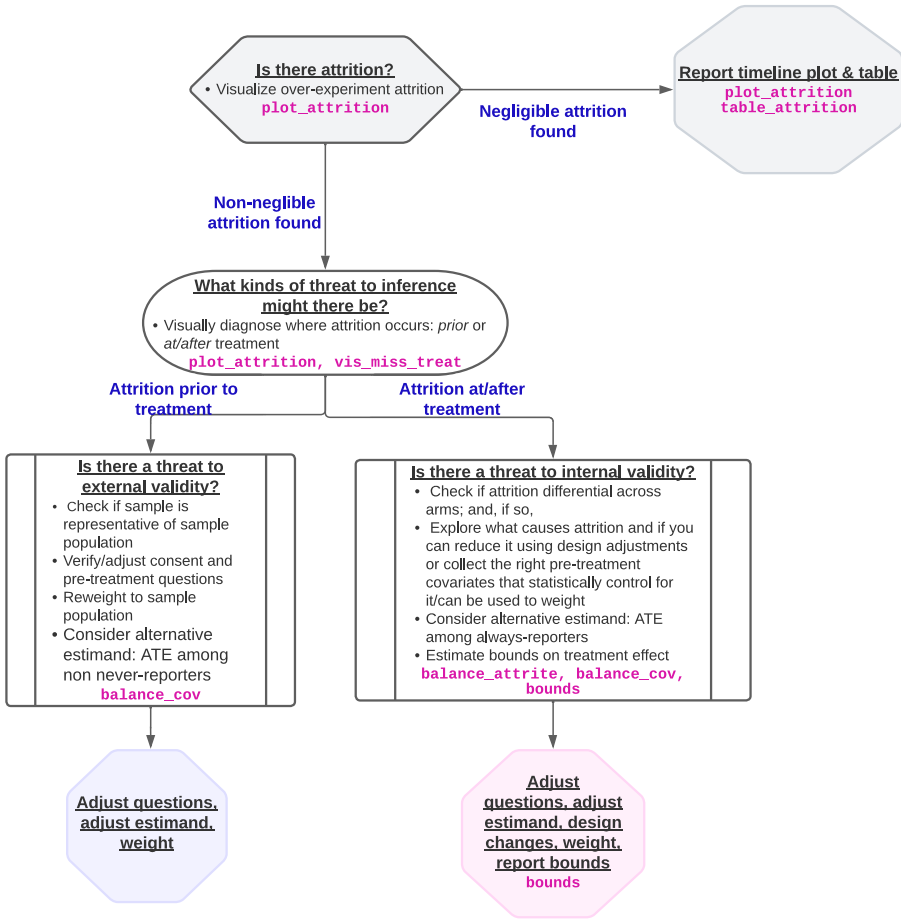


Figure 2. Organizing schematic for assessing and handling attrition in an experimental study. Functions from attritevis that can be utilized at each query stage are in pink.

the survey at certain moments in time, while the second takes into account the baseline number of respondents who still remain at that point in time.

Is there attrition?

Once plotted, the graphic is "know-it-when-you-see-it:" any amount of attrition (in any of the experimental arms) is clearly visible. Consider four different phases of a relatively straightforward (but popular) family of experimental designs: pretreatment, treatment (including the immediate aftermath of the treatment), outcome measurement, and post-outcome. Our package adds a vertical line demarcating the moment of delivery of treatment to highlight pretreatment and posttreatment periods. Figure 3 presents toy examples of experiments that experience varying levels of attrition at different stages. For example, Figure 3(a) suffers from mild attrition, with very low-level attrition distributed more or less equally across the

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press

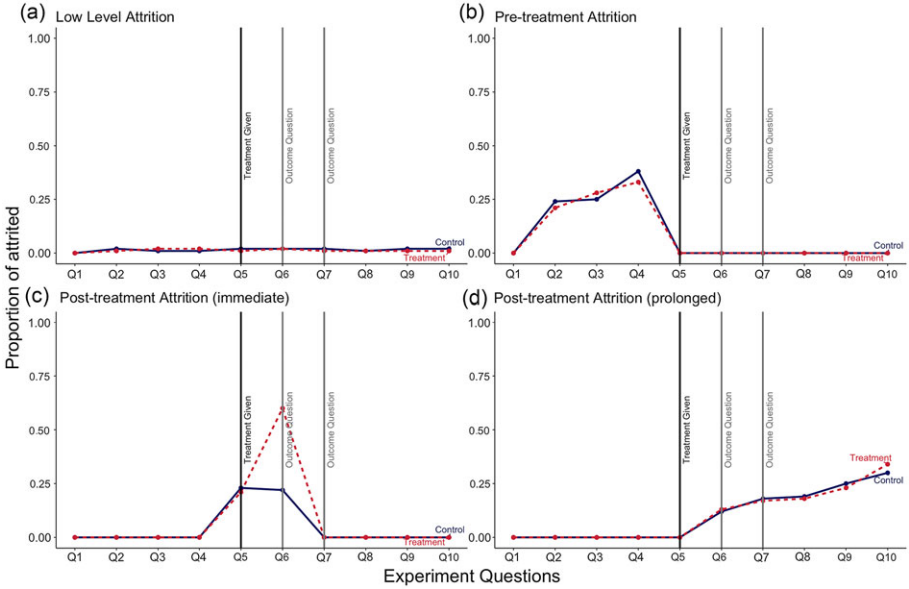


Figure 3. Attrition timeline visualizations: Four toy examples of attrition are presented: (a) low levels of attrition throughout the survey, with little variation across experimental arms; (b) pretreatment attrition, with little variation across arms; (c) attrition right after treatment, with differential attrition across arms; and (d) prolonged posttreatment attrition, with limited variation across arms. We assume treatment in all toy examples is assigned when respondents enter the study and delivered at Q5 (marked with a dark vertical line). The plot_attrition function in attritevis also allows plotting of attrition for all respondents (across all possible treatment groups in the study). This allows users to consider attrition pretreatment, when treatment assignment occurs mid-study. The function further permits users to plot questions by number of responses, rather than attrition, and defaults to gray scale. Users may plot by as many experimental arms as they would like and may specify plot colors.

experiment and across arms – there does not appear to be any particular "choke point" in the design that is generating attrition.

More broadly, how do researchers know if they have attrition worth investigating further? Put simply: for any study with either zero attrition or negligible levels, it should be sufficient to include the Attrition Timeline plot or a standard table – using table_attrition – that attritevis outputs. This is not as unlikely as it seems, as attrition is an uncommon occurrence in some experimental contexts (e.g., some lab studies). Our hope is that the easy-to-use **R** package makes it more likely for researchers to report this quantity even when attrition does not represent a threat to inference. Moreover, the "over-time" dimension of both plot and table represent improvements over the current standard in which only total attrition is reported (an improvement that becomes even more meaningful when there *is* a threat to inference, as discussed below).

Threats to inference and solutions

Experiments are – in an extreme but common stereotype – presumed to deliver clean estimates of causal effects but fall short in allowing scholars to derive lessons applicable "outside the lab." Indeed, the stated rationale for turning to experiments

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press

in the first place is often a concern for internal validity (McDermott 2002, 38). A similar presumed dichotomy is at work in considering the effects of attrition in experiments. Pretreatment missingness implicates the external validity of a study while posttreatment attrition threatens its internal validity: the former because attrition that occurs exclusively pretreatment does not impinge on our ability to randomly assign respondents to treatment arms and achieve probabilistic equivalence, the latter because of the concern that exiting the study is correlated with respondents' potential outcomes. While this is not a misleading heuristic, it is only a start in thinking through how to address attrition in experiments.

We argue that a complementary lens through which to consider attrition is to focus on patterns of nonresponse and the estimands we can recover – and assumptions we can still plausibly make – in the face of different patterns of attrition. In a simple example, we think it is more helpful to consider what estimands are plausible following pretreatment attrition – the ATE among those that remain, "always reporters" in the typology below – rather than playing down the consequences by simply giving ground on the "generalizability" of the study.[7]

To focus attention on estimands and assumptions – and drawing on the framework in Gerber and Green 2012, chapter 7 – we refer to four (latent) *types* of respondents, defined with respect to specific post-treatment outcomes (some observed, others not). Let R be an indicator for whether respondents *respond* for the outcome (and therefore do not *attrite* on the Y) and values in parentheses denote treatment status (treatment (1) or control (0) arms in a simplified two-arm setting):

(a) *Never-reporters*: $R(1) = 0; R(0) = 0$ or individuals who, regardless of treatment status, always attrite on the outcome;
(b) *If-treated reporters*: $R(1) = 1; R(0) = 0$ or individuals who only report answers to the outcome if given treatment, but otherwise attrite;
(c) *If-untreated reporters*: $R(1) = 0; R(0) = 1$ or respondents who report answers to the outcome if given control, but attrite under the treatment regime. These folks are sometimes ruled out by monotonicity assumptions, $R_i(1)R_i(0)$;
(d) *Always-reporters*: $R(1) = 1; R(0) = 1$ or individuals who always report on the outcome regardless of treatment status;

Critically, the shares of each type listed above influence the estimands we can recover and the assumptions we must make to do so. The problem that confronts applied researchers is that, because of the fundamental problem of causal inference, we cannot necessarily know whether individuals are one type or another. Below, we provide an overview of how to diagnose and address (as well as preempt when possible) different patterns of nonresponse, focusing on whether the attrition is pre- or posttreatment. A visual check of the experiment's timeline of attrition can help

[7]Note that this would also correct what we see as a common mistake, which is to equate external validity with generalizing to different *pools* of respondents rather than *any respondents other than those that remain in the study* (even those in the same "pool"). E.g., Scholars fielding MTurk studies with significant pre-treatment attrition are not just limited from generalizing to nationally representative samples, but also to the broader MTurk population as well.

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press

diagnose whether a study suffers from one or both types of attrition. Figure 3(b), for instance, suggests attrition is occurring primarily pretreatment, while (c) and (d) both present as posttreatment attrition cases and (c) displays evidence of imbalanced attrition across arms.

Pre-treatment attrition

Attrition that occurs exclusively pre-treatment can be addressed in several ways depending on the availability of resources and the timeline of the research process. The first step is to diagnose *who* attrited pre-treatment, paying close attention to whether there was nonrandom selection out of the study. Respondents who drop out here can be seen as "never-responders" and, while recovering the true ATE among those remaining is still possible, scholars typically want to know if remainers are a substantial majority of the starting sample (which can be done via visualizations like plot_attrition to find the proportion attrited) and still reflective of the larger sample population. For the latter inquiry, we can verify how *similar* remainers look to the sample population: if they look similar on a host of measured demographics, we can more persuasively argue that the ATE recovered extends to the sample population.

Researchers can use statistical tests of differences such as t-tests against a null of the population value (an option in balance_cov function).[8] Note that this does not reveal *which* stage of selection – into the study in the first place or attriting out of the study – is responsible for any differences, but it does provide helpful information about the extent to which inferences might generalize outside of the sample (and are useful to the extent that researchers wish to makes those inferences). Researchers can also rely on a host of weighting options, including post-stratification, IPW or raking (see Mercer, Lau, and Kennedy 2018 for a summary), with their choices depending on contextual constraints such as available population information and the within-correlation of characteristics respondents are weighted on (see Franco et al. 2017 for a careful take on reweighting in experiments).

For researchers who can field extra studies, or are analyzing pilot data, the full value of the temporal approach becomes apparent. Pinpointing specific moments where attrition is occurring allows researchers to revise their instruments as necessary to reduce troublesome attrition. Common causes might include overly-long instruments – researchers can shorten the instrument, prepare respondents better or increase compensation – or questions that are aversive, either to most or to a particular subgroup.[9]

Post-treatment attrition

Attrition that occurs at or post-treatment is, for experimentalists, potentially more worrisome as it threatens the core assumption that underlies their causal inferences:

[8]A similar exercise, depending on available information on drop-outs, can be conducted to compare remainers and drop-outs.

[9]Our purpose is to attend to loss of respondents that is unintentional and/or based on choices of the respondents. In some cases, researchers might utilize sampling quotas or attention filters that effectively screen out respondents; these are scenarios where the researcher has chosen to focus on specific types of respondents and – as a result – selectively removes or keeps observations based on whatever criteria they've decided upon.

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press

specifically that treatment assignment is independent of potential outcomes. If that assumption is not satisfied – if, for example, in a study about the effect of sadness on political beliefs there is high attrition in only one treatment arm because it asks respondents to imagine a negative event that is aversive – then treatment status could be correlated with potential outcomes, threatening internal validity in a general sense and more specifically our ability to recover the estimands that we desire (e.g., $ATE_{\text{sample population}}$). Put differently, the concern is that the potential outcomes of attriting respondents might be different from those of the "remainers." In these situations, the difference-in-means estimator no longer recovers an unbiased estimate for the ATE (and does not give us the ATE for any meaningful subgroup; see Gerber and Green 2012, 219). If treatment is correlated with potential outcomes, there are a number of steps we can take (detailed below) to either account for the attrition, reduce attrition in the future, bound our estimates, or revise our interpretations.

The first priority is diagnostic and begins with assessing the extent of the damage through visual examination of attrition in the study using our timeline visualization (plot_attrition). In Figure 3(c), there appears to be differential attrition (across treatment arms) around treatment-delivery at Q5. If attrition appears around treatment, the next step is examining whether treatment is correlated with attrition using both the visualization timeline and t-tests of differences at multiple points in the study (using the balance_attrite function, paired with p_adjust to account for multiple tests).

Here, a distinction can be made between the content of the treatment and its delivery. If treatment status is correlated with attrition, one possibility is that something about the *delivery* of the treatment has caused attrition. For a certain class of experiments – e.g., GOTV studies comparing modes of communication, Gerber and Green 2000 – it is possible that something has gone awry with the delivery (e.g., postage was applied carelessly on some mailers) that does not implicate potential outcomes, leaving the MCAR (missing completely at random) assumption intact. In example (c) we might explore the missingness around the treatment in more detail, visualizing it at the respondent level faceted by treatment arm (Figure 4). Emphasis in this plot is placed on individual (rows, ordered here by respondent ID, assigned at study start) missingness throughout the experiment (columns) and across intervention arms (number of faceted plots). Treatment delivery is represented by the red vertical line (Q5), and within-question and intervention-arm percent missing are calculated. Figure 4 shows that the Treatment group in example (c) suffers from 20.3% missing, while the Control group features nearly half (10.2%) that amount of attrition. On closer inspection, our visualization focuses attentions on problem spots, such as in respondents numbering 50-125, the block of whom all attrite in the treatment, which may point to ad hoc glitches in the delivery of the intervention that occurred during that time.

For *most* experiments in which attrition occurs at or following treatment, we must assume that our assumption of MCAR is in jeopardy and proceed accordingly.[10] If missingness is correlated with potential outcomes, core assumptions (MCAR) and

[10]High attrition that is convincingly unrelated to potential outcomes is possible, but rare in our experience. This may occur, for instance, if during a lab experiment a bomb threat or smoke alarm occurs that requires quick evacuation of many study participants (experienced in Woon 2014).

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press

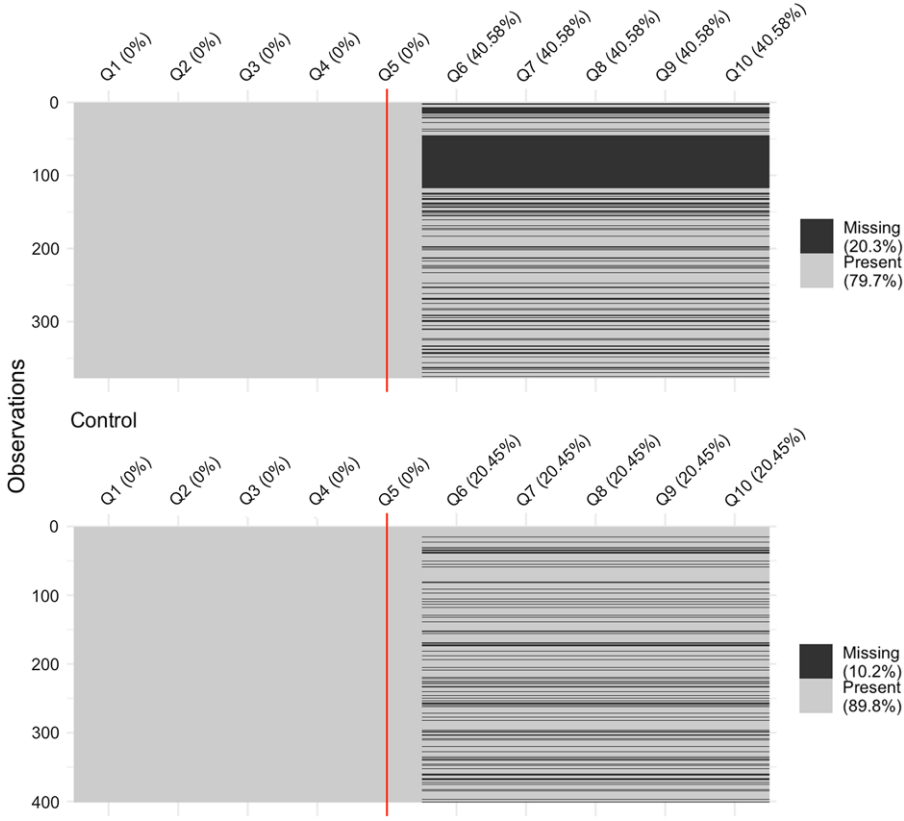


Figure 4. Visualizing missingness by treatment and control group plot produced using the vis_miss_treat function; the function allows users to facet by conditions to present respondent-level visualization of missingness. Red vertical line marks treatment delivery. This figure demonstrates visualization of a toy example with immediate posttreatment attrition, where treatment caused attrition.

estimands $\left(\text{ATE}_{\text{sample population}}\right)$ are at risk. There are a number of steps that might follow, depending on where you are in the research process, level of resources, and what information you have collected. Below, we discuss those steps in rough order of preferability.

If research is still in progress, attrition has occurred in a pilot or re-fielding is possible, we suggest researchers utilize debriefing information and focus groups to understand why the attrition is occurring and, if possible, reduce it in the future. Eliciting this kind of information can help with revising treatments to be less aversive, ascertaining whether there are glitches or technical issues in the delivery of the treatment and even figuring out what covariates one might collect to statistically control for the propensity to attrite.[11]

[11]Sometimes, researchers may not have to field again to collect the extra covariates they find themselves needing, if for example it is attainable from a survey company or, as in the case of some elite studies, is information that is publicly available (Kertzer and Renshon, 2022, 539).

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press

The next practical step is to investigate whether and what covariates predict "selection into attrition" – using balance.cov – to achieve MCAR conditional on X (missing independent of potential outcomes, conditional on a set of observed covariates, X).[12] As opposed to revising the treatment or the study – a "design-based" approach to reducing attrition – we consider this to be a modeling approach to account for attrition. If researchers can determine which respondent demographics (X's) can be conditioned on to achieve $MCAR|X$, reweighting the sample can recover the ATE for the sample population (SATE) (Wooldridge 2007). The weights are based on the likelihood of being observed if respondents have certain values of the conditioned covariates and are required because the difference-in-means estimator alone is no longer sufficient to recover the SATE.

Following this, a practical step one might take is to utilize the common robustness check of framing the treatment effect in the context of the attrition by estimating *bounds* around it (in our package under the bounds function[13]). The rationale behind a bounds exercise is to impute missing values associated with best- and worst-case scenarios for the ATE; this will not estimate the ATE for the sample, but does let you know the range your sample ATE could fall in, which can be useful (even if it will in practice often include zero if attrition is significant or your treatment effect is small).

Other approaches to addressing post-treatment attrition are significantly more resource intensive and involve further sampling. Some of these approaches (e.g., Coppock et al. 2017) require planning for double-sampling and are useful if you anticipate high attrition in advance of fielding. Other approaches do not require planning to double-sample – one simply re-contacts attriters and increases the incentives in an effort to convince them to participate – but may present worries about statistical power (Gomila and Clark 2020). Both of these approaches can present as more practical in the contexts of panel studies.

The solutions noted above are not mutually exclusive. For example, one might use balance.cov to find covariates that predict attrition and then use inverse probability weighting to get closer to MCAR conditional on X and then complement that with bounds as a hedge or robustness check. Those two approaches are particularly suitable to use in concert since one – IPW – relies more on model specifications and the other – bounds – is nonparametric. Another example of combining solutions is to use double sampling along with bounds (Coppock et al. 2017).

A final option is to consider what estimands are still possible if you cannot plausibly achieve $MCAR|X$. Even with significant imbalanced attrition (correlated with potential outcomes), one can technically estimate the ATE among always-responders (never-attriters). This may be useful if we simply want to have predictions of real-world interventions (and we just care about what the forecasted outcome might look like across the sample pool) or because we care only about outcomes among people who are measured for it. However, a word of caution is in order since if, like the vast majority of experimentalists, researchers are interested in estimating a treatment effect in a survey or lab or field experiment, the ATE among

[12]Practically speaking, whether $MCAR|X$ is satisfied is by assumption; researchers will have to rely on expertise and context to determine if they have found a set of covariates sufficient to achieve this.

[13]The bounds function draws from the attrition package Coppock (2022).

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press

always-responders in the presence of significant attrition is likely to be extremely misleading.

Conclusion

Attrition is an "Achilles heel" (Shadish et al. 1998, 3) of randomized experiments, threatening if and how researchers can generalize from their data and even their ability to identify causal effects in the first place. A number of helpful methods have been developed over the years to address the problems in the analysis stage, notably re-weighting approaches, bounds and re-contacting attritors. Yet, in a quantitative literature search, we found that researchers typically ignore the issue, failing to account for or even report attrition rates in most published studies.

We argued that at least part of the mismatch between the scope of the problem and the attention paid to it results from lack of usable tools for researchers to diagnose the types of attrition in their studies. Our contribution was to offer a suite of tools in an open-source **R** package (attritevis) that allow experimentalists to visualize when attrition is occurring. The primary outputs of the package are study-level and unit-level "attrition timelines" of the experiment, enabling researchers to easily answer the critical questions of whether/where attrition is occurring across treatment arms.

To supplement the tools in attritevis, we provided a set of guidelines for applied researchers, focusing on where and among whom the attrition occurs and the implications of the missingness for the estimands we typically desire and the assumptions required to recover them. Our guide balances post-hoc analytical tools with advice for revising designs to ameliorate problematic attrition and a battery of visualization techniques to transparently report attrition. We emphasized the utility of pilot testing and debriefing – used in tandem with the tools provided in our package – to diagnose and then preempt attrition: as is always the case with experiments, the cheapest and most efficient way to fix things is by doing so in the pilot stage. However, regardless of whether researchers are able to lower attrition rates or simply account for them statistically *ex-post*, we hope and expect the tools we provide to increase the level of transparency surrounding attrition in experiments.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/XPS.2023.22

Data availability statement. The data, code, and any additional materials required to replicate all analyses in this article are available at the Journal of Experimental Political Science Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/O2IAWW. Lo, Adeline Y.; Renshon, Jonathan; Bassan-Nygate, Lotem, 2023, "Replication Data for: A Practical Guide to Dealing with Attrition in Political Science Experiments," Harvard Dataverse.

Competing interests. None.

Ethics statement. The research was approved by the University of Wisconsin Madison Institutional Review Board Office of Human Research Ethics, study 2020-0843-CP002.

The research adheres to APSA's Principles and Guidance for Human Subjects Research.

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press

References

Boas, Taylor C. 2016. "Pastors for Pinochet: Authoritarian Stereotypes and Voting for Evangelicals in Chile." *Journal of Experimental Political Science 3*(2): 197–205.

Brutger, Ryan, Josh Kertzer, Jonathan Renshon, Dustin Tingley and Chagai Weiss. 2022. "Abstraction and Detail in Experimental Design." *American Journal of Political Science*.

Callen, Michael, Clark C. Gibson, Danielle F. Jung and James D. Long. 2016. "Improving Electoral Integrity with Information and Communications Technology." *Journal of Experimental Political Science 3*(1): 4–17.

Castiglioni, Laura, Klaus Pforr and Ulrich Krieger. 2008. "The Effect of Incentives on Response Rates and Panel Attrition: Results of a Controlled Experiment." *Survey Research Methods 2*(3): 151–8. https://ojs.ub.uni-konstanz.de/srm/article/view/599

Coppock, Alexander. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods 7*(3): 613–28.

Coppock, Alexander. 2021. "Visualize as You Randomize." In Advances in Experimental Political Science, eds. James N. Druckman and Donald P. Green. New York, NY: Cambridge University Press, 320–337.

Coppock, Alexander. 2022. *attrition: What the Package Does (one line, title case)*. R package version 0.0.0.9000.

Coppock, Alexander, Alan S. Gerber, Donald P. Green and Holger L. Kern. 2017. "Combining Double Sampling and Bounds to Address Nonignorable Missing Outcomes in Randomized Experiments." Political Analysis 25(2): 188–206.

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

Deng, Yiting, D. Sunshine Hillygus, Jerome P. Reiter, Yajuan Si and Siyu Zheng. 2013. "Handling Attrition in Longitudinal Studies: The Case for Refreshment Samples." Statistical Science 28(2): 238–56.

Druckman, James N., Donald P. Green, James H. Kuklinski and Arthur Lupia. 2011. "An Introduction to Core Concepts." In *The Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Green, James H. Kuklinski and Arthur Lupia. New York, NY: Cambridge University Press, 15–26.

Franco, Annie, Neil Malhotra, Gabor Simonovits and L. J. Zigerell. 2017. "Developing Standards for Post-Hoc Weighting in Population-Based Survey Experiments." *Journal of Experimental Political Science 4*(2): 161–72. https://www.cambridge.org/core/journals/journal-of-experimental-political-science/article/abs/developing-standards-for-posthoc-weighting-in-populationbased-survey-experiments/A01077D9A2CFB29FA02CE7963B04BB64

Gerber, Alan S. and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." American Political Science Review 94(3): 653–63.

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. WW Norton.

Gerber, Alan S., Kevin Arceneaux, Cheryl Boudreau, Conor Dowling, Sunshine D. 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.

Gomila, Robin and Chelsey S. Clark. 2020. "Missing Data in Experiments: Challenges and Solutions." Psychological Methods.

Göritz, Anja S. 2014. "Determinants of the Starting Rate and the Completion Rate in Online Panel Studies." In *Online Panel Research: A Data Quality Perspective*, eds. Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick and Paul J. Lavrakas. Wiley Online Library, 154–70.

Green, Donald P. and Adam Zelizer. 2017. "How Much GOTV Mail is Too Much? Results from a Large-Scale Field Experiment." *Journal of Experimental Political Science 4*(2): 107–18.

Hansen, Morris H. and William N. Hurwitz. 1946. "The Problem of Non-Response in Sample Surveys." Journal of the American Statistical Association 41(236): 517–29.

Horton, John J., David G. Rand and Richard J. Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics 14*(3): 399–425.

Jerit, Jennifer, Jason Barabas and Scott Clifford. 2013. "Comparing Contemporaneous Laboratory and Field Experiments on Media Effects." *Public Opinion Quarterly 77*(1): 256–82.

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press

Kaufman, Gordon M. and Benjamin King. 1973. "A Bayesian Analysis of Nonresponse in Dichotomous Processes." *Journal of the American Statistical Association 68*(343): 670–78.

Kertzer, Joshua D. 2020. "Re-assessing Elite-Public Gaps in Political Behavior." *American Journal of Political Science*, Forthcoming.

Kertzer, Joshua D. and Jonathan Renshon. 2022. "Experiments and Surveys on Political Elites." *Annual Review of Political Science 25*: 529–50.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *The Review of Economic Studies 76*(3): 1071–102.

Lynn, Peter. 2018. "Tackling Panel Attrition." In *The Palgrave Handbook of Survey Research*. Springer, 143–53.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Manski, Charles F. 2009. *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press.

McCambridge, Jim, Eleftheria Kalaitzaki, Ian R. White, Zarnie Khadjesari, Elizabeth Murray, Stuart Linke, Simon G. Thompson, Christine Godfrey and Paul Wallace. 2011. "Impact of Length or Relevance of Questionnaires on Attrition in Online Trials: Randomized Controlled Trial." *Journal of Medical Internet Research 13*(4): e1733. https://www.jmir.org/2011/4/e96

McDermott, Rose. 2002. "Experimental Methods in Political Science." *Annual Review of Political Science 5*(1): 31–61.

Mercer, Andrew, Arnold Lau and Courtney Kennedy. 2018. "1. How different weighting methods work." https://www.pewresearch.org/methods/2018/01/26/how-different-weighting-methods-work/

Morrison, Theodore C., Dennis R. Wahlgren, Melbourne F. Hovell, Joy Zakarian, Susan Burkham-Kreitner, C. Richard Hofstetter, Donald J. Slymen, Kristen Keating, Stergios Russos and Jennifer A. Jones. 1997. "Tracking and Follow-Up of 16,915 Adolescents: Minimizing Attrition Bias." Controlled Clinical Trials 18(5): 383–96.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science 2*(2): 109–38.

Mummolo, Jonathan and Erik Peterson. 2019. "Demand Effects in Survey Experiments: An Empirical Assessment." *American Political Science Review 113*(2): 517–29.

Musch, Jochen and Ulf-Dietrich Reips. 2000. "A Brief History of Web Experimenting." In Psychological Experiments on the Internet, ed. Michael H. Birnbaum. Elsevier, 61–87.

Mutz, Diana C. and Robin Pemantle. 2015. "Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods." *Journal of Experimental Political Science 2*(2): 192–215.

Reips, Ulf-Dietrich. 2000. "The Web Experiment Method: Advantages, Disadvantages, and Solutions." In *Psychological Experiments on the Internet*. Elsevier, 89–117.

Shadish, William R, Xiangen Hu, Renita R. Glaser, Richard Kownacki and Seok Wong. 1998. "A Method for Exploring the Effects of Attrition in Randomized Experiments with Dichotomous Outcomes." *Psychological Methods 3*(1): 3.

Wooldridge, Jeffrey M. 2007. "Inverse Probability Weighted Estimation for General Missing Data Problems." *Journal of Econometrics 141*(2): 1281–301.

Woon, Jonathan. 2014. "An Experimental Study of Electoral Incentives and Institutional Choice." *Journal of Experimental Political Science 1*(2): 181–200.

Zhou, Haotian and Ayelet Fishbach. 2016. "The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (Yet False) Research Conclusions." *Journal of Personality and Social Psychology 111*(4): 493–504.

Cite this article: Lo A, Renshon J, and Bassan-Nygate L (2024). A Practical Guide to Dealing with Attrition in Political Science Experiments. *Journal of Experimental Political Science* **11**, 147–161. https://doi.org/10.1017/XPS.2023.22

https://doi.org/10.1017/XPS.2023.22 Published online by Cambridge University Press