# Overcoming Attrition in Experimental Research

Andrew Bertoli

16 March 2015

Abstract. Experiments, both real and natural, can be powerful tools for causal inference, but the standard difference-in-means estimator may be biased when the outcomes of certain units are undefined or unobserved. This problem arises in studies across a wide range of fields, including epidemiology, economics, political science, and psychology. Put simply, the initial randomization of treatment does not guarantee that the treated and control units with observed outcomes will be comparable. In this paper, I formally derive an expression for attrition bias that takes into account different types of attrition, and I clarify under what conditions this bias will arise. I then analyze several major studies that face different versions of the attrition problem and outline some strategies that the researchers could use to address it. These examples are primarily from natural experiments in international relations, although they illustrate similar problems faced by many other studies in different fields. The examples show that the best solution usually depends on the question of interest and the version of the attrition problem that researchers face.

Attrition can be a major threat to causal inference in experimental research. Even when treatment assignment is random, the normal estimates may be biased if some subjects die, migrate, or fail to respond for other reasons. This problem arises for medical experiments where some patients pass away or move to other hospitals before the outcome is measured. Similarly, it also faces studies that compare citizens who were randomly drafted into military service to those who were not.[1] Even if there was perfect compliance, the survivors in the treatment and control groups are not comparable, since whether a person survives is not random. Attrition can even cause bias when there is no missing outcomes in the dataset. For instance, a study that tests how indiscriminate

[1] Henderson 2014

bombing affects the nationalism of citizens might not appear to have missing outcomes if it was collected after the war, but there would likely be a substantial number of bombed individuals who never appeared in the dataset. If these citizens tended to be older or poorer than the citizens who survived, then this problem could be a major source of bias.

As these examples suggest, attrition problems come in a variety of forms. Sometimes units die or are destroyed, and other times they survive but fail to respond. Researchers may be able to distinguish between these different types of non-response, or they may have no information about the units with missing outcomes. Moreover, researchers sometimes know all of the units that were originally in the sample, but are just missing outcomes for some of them. Other times, researchers never even observe certain units and might have little information about how serious the attrition problem is.

The researcher's question of interest can also affect the extent to which attrition undermines inference. For example, a policymaker who wants to know if indiscriminate bombing during war breaks the enemy's resolve might be indifferent to whether the bombing changes the victims beliefs or just kills the more nationalistic citizens at higher rates. On the other hand, a political psychologist might care deeply about which of these two explanations is driving the results, and thus have more reason to worry about attrition.

In this study, I attempt to address the attrition problem in a more comprehensive way than past research. Previous studies have contributed to solving the attrition problem by defining the bias in one context and offering a single solution, such as bounding the estimated effect or using imputation to recover lost outcomes. Some of these techniques are very innovative and can be highly useful in certain contexts. In contrast, this study categorizes different versions of the attrition problem, formally derives a more general expression for attrition bias, and recommends which techniques are most appropriate in different situations.

No doubt, the best strategy for managing attrition will vary from case to case. In some situations, researchers can estimate meaningful causal parameters by simply listwise deleting the units with missing outcomes. In others, the solution is not so simple, but researchers can make progress by using some of the methods proposed here or developed in past research. There are also cases where it is very difficult to overcome attrition, and researchers might consider using an alternative research design like regression or matching. In this paper, I illustrate these key distinctions between different versions of the attrition problem through examples from international relations and security studies. My goal is to provide a framework for thinking about attrition in different contexts and finding the best possible way to resolve it in each case.

This paper proceeds as follows. In Section 1, I lay the groundwork by discussing different types of attrition in the context of the potential outcomes framework and formally derive the bias caused by attrition. Section 2 outlines some of the techniques that previous studies have suggested to solve attrition problems, along with presenting some new ones. In Section 3, I classify the different versions of the attrition problem. Section 4 explains how these versions can be addressed with different methods and provides some examples to illustrate each strategy. The final section concludes.

## Section 1: The Potential Outcomes Framework and the Attrition Problem

**Basic Set-up.** Rubin (1974) proposed a clear framework for understanding how experiments can be used for causal inference. Assume that there are $n$ units, each with a potential outcome under treatment ($Y_{it}$) and a potential outcome under control ($Y_{ic}$). The Average Treatment Effect ($\bar{\tau}$) is defined as the average difference in these potential outcomes across all the units: $\bar{\tau} = \frac{1}{n}\sum_i (Y_{it} - Y_{ic})$. We cannot compute $\bar{\tau}$ directly, since we only observe $Y_{it}$ or $Y_{ic}$ for each unit, depending on whether that unit was assigned to treatment or control. Thus, causal inference is a missing data problem. Randomized treatment assignment allows researchers to estimate $\bar{\tau}$ without bias by providing random samples from the $Y_{it}$'s and $Y_{ic}$'s. It also makes it very straightforward to calculate the probability of seeing a $\hat{\tau}$ as extreme or more extreme than the observed one under the assumption that $\bar{\tau} = 0$.

An attrition problem arises when some units do not have observed or defined outcomes under both treatment and control. For instance, a draftee who died in Vietnam would have no defined income (U) in 1980. It would make little sense to code this person's income at $0, since being dead is very different than being unemployed. Units can also have unobserved outcomes, whether they are defined or undefined. For example, if researchers knew that a draftee was alive, but he was unwilling to report his income, then his outcome would be defined but unobserved (M). On the other hand, he could have an undefined and unobserved outcome if he died but the researchers were not aware of it. I refer to these units as lost (L), as researchers do not even know whether they are defined or defined. However, researchers do know that lost units were originally in the sample.

Now most researchers label all missing or undefined outcomes as NA, but doing so discards important information. There are major substantive differences between people who die and people who fail to answer their phones, and there are also key differences in terms of how you handle these different types of attrition statistically, as I will discuss in

**Figure 1: Potential Scenarios**

**Case 1**

| Unit | $Y_{it}$ | $Y_{ic}$ |
|---|---|---|
| 1 | 2 | 4 |
| 2 | 1 | 1 |
| 3 | 4 | 3 |
| 4 | 0 | 2 |
| 5 | 3 | 2 |

**Case 2**

| Unit | $Y_{it}$ | $Y_{ic}$ |
|---|---|---|
| 1 | 3 | 1 |
| 2 | 1 | 2 |
| 3 | 2 | 3 |
| 4 | 1 | 2 |
| 5 | 0 | 3 |

**Case 3**

| Unit | $Y_{it}$ | $Y_{ic}$ |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 3 | U |
| 3 | U | 2 |
| 4 | U | U |
| 5 | 0 | 1 |

**Case 4**

| Unit | $Y_{it}$ | $Y_{ic}$ |
|---|---|---|
| 1 | 2 | 3 |
| 2 | U | 3 |
| 3 | 1 | U |
| 4 | 3 | 2 |
| 5 | 0 | U |

**Case 5**

| Unit | $Y_{it}$ | $Y_{ic}$ |
|---|---|---|
| 1 | 5 | 2 |
| 2 | 1 | 1 |
| 3 | 2 | 0 |
| 4 | 3 | 1 |
| 5 | 1 | 0 |

**Case 6**

| Unit | $Y_{it}$ | $Y_{ic}$ |
|---|---|---|
| 1 | 3 | 1 |
| 2 | 2 | 1 |
| 3 | 2 | 2 |
| 4 | 3 | ❶ |
| 5 | ❶ | 0 |

Notes: Black numbers denote observed outcomes, gray numbers denote unobserved outcomes, “U” stands for undefined, and the white numbers in dark circles represent potential outcomes for units that are missing from the dataset under treatment or control.

the next two sections. Better practice would be to label missing data as U when known to be undefined, M when known to be defined but missing, and L when lost, as well as provide specific information about the sources of missingness when possible.

Figure 1 shows the potential outcomes for several scenarios that researchers might face. Case 1 has no attrition problem, since all units have defined and observed potential outcomes under treatment and control. In Case 2, all units have defined potential outcomes, but some are unobserved. Case 3 is a scenario where all units have observed potential outcomes, but some are undefined. Case 4 features both unobserved and undefined outcomes. Case 5 has a hidden attrition problem, in the sense that whenever Unit

1 is assigned to control, researchers will be unaware of attrition because all units will have defined and observed outcomes. However, the normal difference in means estimator will be biased, since $E[\hat{\tau}] \neq \bar{\tau}$ prior to randomization. Specifically, $\bar{\tau} = \frac{7}{5}$, whereas $E[\hat{\tau}] = \frac{3}{5}$ if we ignore missingness when it arises. In Case 6, Unit 4 is never observed by researchers when it is assigned to control, and Unit 5 is never observed when it is assigned to treatment. For instance, if researchers in 1980 compared men who were drafted to go to Vietnam to men who were not, Unit 4 would only be found by researchers if he was drafted, and Unit 5 would only be found if he was not drafted. Otherwise, these men would not be in the dataset. I refer to such subjects as "phantom units", and they can cause an attrition problem even when no units in the dataset have missing outcomes.

**Average Treatment Effect Under Attrition.** Assume that there are $n$ units in the sample at the time of randomization, and $m$ of them are assigned the treatment group. Each unit's treatment status is denoted as $T_i \in \{0, 1\}$. Furthermore, each units potential outcomes can be defined or undefined, $D_{it}, D_{ic} \in \{0, 1\}$, and observed or unobserved, $O_{it}, O_{ic} \in \{0, 1\}$.

The the usual difference in means estimator may be biased if there are unobserved outcomes, as I will show shortly, but a more fundamental problem arises when potential outcomes are undefined. Note that $\tau_i = Y_{it} - Y_{ic}$ only exists if Unit $i$ has defined potential outcomes under both treatment and control. Thus, if any unit has an undefined potential outcome, then the Average Treatment Effect ($\bar{\tau} = \frac{1}{n} \sum_i (Y_{it} - Y_{ic})$) is undefined. In other words, the standard parameter of interest in a normal experiment simply does not exist if any units have undefined potential outcomes, and researchers will have to choose something else to estimate.

**Alternative Parameters.** One parameter that still exists when potential outcomes are undefined is the Difference in World Averages (DIWA): $E[Y_{it}|D_{it} = 1] - E[Y_{ic}|D_{ic} = 1]$. DIWA is the mean difference between the defined outcomes in the world where all units are assigned to treatment and the defined outcomes in the world where all units are assigned to control. Suppose that a dictator ran an experiment to see how participation in war affected support for his regime. The DIWA would be the average level of support for the survivors when the entire sample went to war minus the the average level of support for the survivors when the entire sample stayed home.

Researchers can also estimate the Restricted Average Treatment Effect (RATE), which is the treatment effect for units that would survive under both treatment and control. Specifically, it is written as RATE $= E[Y_{it} - Y_{ic}|D_{it}, D_{ic} = 1]$. Zhang and Rubin (2003) call this parameter the Surviver Average Causal Effect (SACE). I break from this terminology because units may have undefined potential outcomes for reasons other than death.

When all units have defined potential outcomes, the ATE equals the RATE and the DIWA. However, when there are any undefined potential outcomes, the ATE is undefined and the RATE and DIWA will not necessarily be equal.

**Deriving the Bias.** We can now calculate the bias of the normal difference in means estimator when attrition is present with respect to the DIWA and the RATE. Recall that the ATE ($\bar{\tau}$) is only defined when all units have defined potential outcomes, in which case it is simply equivalent to the DIWA and the RATE. Thus, I will not derive the bias specifically for the ATE, since it is just a special case of the DIWA and the RATE.

The normal difference-in-means estimator is written as

$\hat{\tau} = Avg(Y_{it}|T_i = 1, D_{it} = 1, O_{it} = 1) - Avg(Y_{ic}|T_i = 0, D_{ic} = 1, O_{ic} = 1)$

The expected value of $\hat{\tau}$ is easily calculated after noting that

$E[Avg(Y_{it}|T_i = 1, D_{it} = 1, O_{it} = 1)] = \frac{1}{\sum_{i=1}^{n} D_{it}O_{it}} \sum_{i=1}^{n} Y_{it}D_{it}O_{it}$

$E[Avg(Y_{ic}|T_i = 0, D_{it} = 1, O_{ic} = 1)] = \frac{1}{\sum_{i=1}^{n} D_{ic}O_{ic}} \sum_{i=1}^{n} Y_{ic}D_{ic}O_{ic}$

So

$E[\hat{\tau}] = \frac{1}{\sum_{i=1}^{n} D_{it}O_{it}} \sum_{i=1}^{n} Y_{it}D_{it}O_{it} - \frac{1}{\sum_{i=1}^{n} D_{ic}O_{ic}} \sum_{i=1}^{n} Y_{ic}D_{ic}O_{ic}$

In comparison, the DIWA can be written as

DIWA$= E[Y_{it}|D_{it} = 1] - E[Y_{ic}|D_{ic} = 1]$

DIWA$= \frac{1}{\sum_{i=1}^{n} D_{it}} \sum_{i=1}^{n} Y_{it}D_{it} - \frac{1}{\sum_{i=1}^{n} D_{ic}} \sum_{i=1}^{n} Y_{ic}D_{ic}$

and the RATE can be written as

RATE $= E[Y_{it} - Y_{ic}|D_{it}, D_{ic} = 1]$

RATE $= E[Y_{it}|D_{it}, D_{ic} = 1] - E[Y_{ic}|D_{it}, D_{ic} = 1]$

RATE $= \frac{1}{\sum_{i=1}^{n} D_{it}D_{ic}} \sum_{i=1}^{n} Y_{it}D_{it}D_{ic} - \frac{1}{\sum_{i=1}^{n} D_{it}D_{ic}} \sum_{i=1}^{n} Y_{ic}D_{it}D_{ic}$

So the bias of the normal difference in means estimator with respect to the DIWA is

$Bias_{DIWA}(\hat{\tau}) = E[\hat{\tau}]$ - DIWA

$Bias_{DIWA}(\hat{\tau}) = \frac{1}{\sum_{i=1}^{n} D_{it}O_{it}} \sum_{i=1}^{n} Y_{it}D_{it}O_{it} - \frac{1}{\sum_{i=1}^{n} D_{it}O_{ic}} \sum_{i=1}^{n} Y_{ic}D_{ic}O_{ic} -$

$[\frac{1}{\sum_{i=1}^{n} D_{it}} \sum_{i=1}^{n} Y_{it}D_{it} - \frac{1}{\sum_{i=1}^{n} D_{ic}} \sum_{i=1}^{n} Y_{ic}D_{ic}]$

$Bias_{DIWA}(\hat{\tau}) = \frac{1}{\sum_{i=1}^{n} D_{it}O_{it}} \sum_{i=1}^{n} Y_{it}D_{it}O_{it} - \frac{1}{\sum_{i=1}^{n} D_{it}} \sum_{i=1}^{n} Y_{it}D_{it} -$

$[\frac{1}{\sum_{i=1}^{n} D_{it}O_{ic}} \sum_{i=1}^{n} Y_{ic}D_{ic}O_{ic} - \frac{1}{\sum_{i=1}^{n} D_{ic}} \sum_{i=1}^{n} Y_{ic}D_{ic}]$

This equation can be rewritten as (proof in the online appendix)

$Bias_{DIWA}(\hat{\tau}) = P(O_{it} = 0|D_{it} = 1)[E[Y_{it}|D_{it}, O_{it} = 1] - E[Y_{it}|D_{it} = 1, O_{it} = 0] -$

$(P(O_{ic} = 0|D_{ic} = 1)[E[Y_{ic}|D_{ic}, O_{ic} = 1] - E[Y_{ic}|D_{ic} = 1, O_{ic} = 0])$

It is clear to see from either of the two equations above that $\hat{\tau}$ will be unbiased if (1) $P(O_{it} = 0|D_{it} = 1)$ and $P(O_{ic} = 0|D_{ic} = 1)$ or (2) $E[Y_{it}|D_{it}, O_{it} = 1] = E[Y_{it}|D_{it} = 1]$ and $E[Y_{ic}|D_{ic}, O_{ic} = 1] = E[Y_{ic}|D_{ic} = 1]$. In other words, $\hat{\tau}$ will be unbiased if (1) all defined outcomes can be observed or (2) the observed potential outcomes under treatment are representative of all the defined potential outcomes under treatment, and the observed potential outcomes under control are representative of all the defined potential outcomes under control. Otherwise, there will be bias in the estimator except in the rare case where the values in the first and second lines offset. In short, estimation of the DIWA is threatened by defined but missing potential outcomes, where the missingness is "non-random" (meaning that it is uncorrelated with the potential outcomes, or $Y_{it} \perp O_{it}$ and $Y_{ic} \perp O_{ic}$).

The bias of the normal difference in means estimator with respect to the RATE is

$Bias(\hat{\tau}) = E[\hat{\tau}]$ - RATE

$Bias_{RATE}(\hat{\tau}) = \frac{1}{\sum_{i=1}^{n} D_{it}O_{it}} \sum_{i=1}^{n} Y_{it}D_{it}O_{it} - \frac{1}{\sum_{i=1}^{n} D_{ic}O_{ic}} \sum_{i=1}^{n} Y_{ic}D_{it}O_{ic} -$

$[\frac{1}{\sum_{i=1}^{n} D_{it}D_{ic}} \sum_{i=1}^{n} Y_{it}D_{it}D_{ic} - \frac{1}{\sum_{i=1}^{n} D_{it}D_{ic}} \sum_{i=1}^{n} Y_{ic}D_{it}D_{ic}]$

$Bias_{RATE}(\hat{\tau}) = \frac{1}{\sum_{i=1}^{n} D_{it}O_{it}} \sum_{i=1}^{n} Y_{it}D_{it}O_{it} - \frac{1}{\sum_{i=1}^{n} D_{it}D_{ic}} \sum_{i=1}^{n} Y_{it}D_{it}D_{ic} -$

$[\frac{1}{\sum_{i=1}^{n} D_{it}O_{ic}} \sum_{i=1}^{n} Y_{ic}D_{it}O_{ic} - \frac{1}{\sum_{i=1}^{n} D_{it}D_{ic}} \sum_{i=1}^{n} Y_{ic}D_{it}D_{ic}]$

So, in general, there will be no bias when the average of the defined and observable $Y_{it}$'s equals the average of the $Y_{it}$'s for units with defined potential outcomes under both treatment and control, and when the same holds true for the $Y_{ic}$'s.

Specifically, the condition above will hold if, for all $i$ such that $D_{it} = 1$, $Y_{it} \perp \{O_{it}, D_{ic}\}$, and for all $i$ such that $D_{ic} = 1$, $Y_{ic} \perp \{O_{ic}, D_{it}\}$. This will be guaranteed if attrition is orthogonal to the potential outcomes, although complete orthogonality is not necessary for unbiasedness. A special case of the condition above is when the treatment does not affect $D$ or $O$ for any unit (the treatment does not cause any outcomes to be missing or undefined) and the units with observed and defined outcomes do not have a different average treatment effect than all the units with defined outcomes. In this case, $D_{it} = D_{ic}$ and $O_{it} = O_{ic}$ for all units, and the proof of unbiasedness is trivial from the equation above. If the units with observed and defined outcomes have different average treatment effect, then the average treatment effect for units with observed potential outcomes can still be estimated unbiasedly, but the average treatment effect for all of the units with defined potential outcomes is no longer identified.

Thus, while intuition might suggest that missingness must be random to estimate important causal parameters, this is not actually the case. If it is very unlikely that the treatment affected either $D$ or $O$, then it is possible to estimate the treatment effect for the sub-sample of the data with observed and defined potential outcomes without bias. As mentioned earlier, it is also possible to estimate the DIWA provided that O=1 for all defined potential outcomes. However, there are many scenarios where these conditions will not be enough to allow researchers to estimate what they want. In these cases, they must look to the strategies for dealing with attrition problems, which I will outline in the next section.

## Section 2: Techniques for Overcoming Attrition

There are a diverse group of methods that researchers have developed to resolve attrition problems. In this section, I outline six of the most common general approaches used in past research, and I propose a some new ones. For several of these approaches, there are a variety of specific techniques that can be used. The purpose of this paper is not to explain and evaluate the specific techniques, but to investigate which general approaches are most useful in different contexts. Therefore, I will focus on the key aspects and assumptions of the general approaches while briefly describing some of the specific techniques throughout this discussion.

**Imputation**. Imputation involves predicting outcomes by using some model that is based on the covariates. There are a number of possible ways to impute outcomes.

**Principal-Stratification**. Like imputation, principal-stratification relies on the covariates, although it is essentially the opposite of imputation. Whereas imputation involves estimating outcomes for the units that do not have them, principal stratification reduces the sample to the units that researchers believe would have outcomes under both treatment and control.[2] The goal is to estimate the RATE for some subset of the data. For instance, if researchers noticed that the treatment killed a high percentage of men, but all of the women survived, they might focus their analysis only on the women. Imputation,

[2] Rubin 2006

on the other hand, is typically used to estimate the ATE. In general, imputation is more appropriate when dealing with missing but defined outcomes (M), whereas principal-stratification is better suited for undefined outcomes (U).

A limitation of this approach is that it may not be clear which units would have outcomes under both treatment and control. When this is the case, a common approach is to predict the presence of outcomes based on some model of the covariates.[3] Researchers can then reweigh the observed outcomes by the inverse their estimated probability of being observed.

As with imputation, it is very difficult to assess the bias that could be induced by this method. Whether it is preferable to regression or matching on the observed and observed and defined data may be hard to determine and will vary from case to case.

**Upward Bounding.** Upward bounding involves calculating the largest and smallest treatment effects for any possible values of the missing outcomes.[4] This is straightforward when outcomes are restricted to a range of values. When outcomes are unbounded, researchers must assume that they are restricted to some reasonable set of values. They can also use covariates to specify a range of plausible outcomes for each unit. Like imputation, upward bounding involves recovering the dataset and computing the ATE for all units.

**Downward Bounding.** Downward bounding is useful when researchers are interested in the RATE. Since the RATE is the treatment effect for units with defined potential outcomes under both treatment and control, researchers can drop the units with undefined outcomes, since they are clearly not part of this group. Next, researchers just

[3] ; Gelman and Hill 2007; Jo and Stuart 2009

[4] Lee 2002

need to compute the highest and lowest treatment effects after removing units with defined outcomes, since these are the only units that might have potential outcomes under treatment and control. Zhang, Rubin, and Mealli derive a formula for estimating these bounds based on large-sample approximations.[5] I propose an alternative method here that does not depend on these assumptions.

The primary issue is determining how many units should be dropped. In theory, all of the treated and control units with defined outcomes might have undefined counterfactuals, meaning that there would be no units with defined outcomes under both treatment and control. However, this problem will be very unlikely in most cases. In fact, it is usually possible to estimate a reasonable upper limit on the possible number of missing units. If 5% of the treated units have undefined outcomes, then we can assume that about 5% of the control units would have had undefined outcomes if they were assigned to treatment, and put a confidence interval on this estimate. This estimation is possible as long as we are not missing units from our dataset, since treatment was random and we would know whether or not we had defined outcomes for each unit.

Assuming that all outcomes our observed (with some undefined), the 95% confidence interval for the number of treated units that have undefined control potential outcomes is easy to estimate with permutation inference. Specifically, the confidence interval for total number of units $U_c$ with undefined potential outcomes under control can be estimated using the hypergeometric distribution. Let $N$ be the total number of units, $n_c$ be the number of control units, and $u_c$ be the number of control units with undefined outcomes. Then the one-sided 95% confidence sets the upper bound as

$$U_c \leq \underset{K \geq u_c}{argmin} \{ \textstyle\sum_{i=1}^{u_c} \frac{\binom{K}{i}\binom{N-K}{n_c-i}}{\binom{N}{n_c}} \leq 0.05 \}$$

[5] Zhang, Rubin, and Mealli 2008

Similarly, the one-sided 95% confidence interval for total number of units $U_t$ with undefined potential outcomes under treatment is

$$U_t \leq \underset{K\geq u_t}{argmin}\{\textstyle\sum_{i=1}^{u_t}\frac{\binom{K}{i}\binom{N-K}{n_t-i}}{\binom{N}{n_t}} \leq 0.05\}$$

The number of units that we should drop from the treatment group is the upper limit of $U_c$ minus the number of control units with undefined outcomes $u_c$ , and the number of units that we should drop from the control group is the upper limit of $U_t$ minus $u_t$. If we have 100 units in both the treatment and control groups, and five from the treatment group have missing outcomes, then we would estimate that at most $15$ of the $200$ $Y_{it}$'s are undefined. Thus, we would subtract a maximum of $15 - 5 = 10$ units from the control group.

Of course, this formula will cause computational problems for sample sizes above about 300, as computers have trouble dealing with very large factorials. However, it is easy to get around this issue by using logarithms. Specifically, the fractions above can be rewritten as

$$\frac{\binom{K}{i}\binom{N-K}{n_c-i}}{\binom{N}{n_c}} = exp\{(\textstyle\sum_{j=1}^{K} ln(j) - \sum_{j=1}^{i} ln(j) - \sum_{j=1}^{K-i} ln(j) + \sum_{j=1}^{N-K} ln(j) -$$
$$\textstyle\sum j = 1^{n_c-i} lnj - \sum_{j=1}^{N-K-(n_c-i)} -(\sum_{j=1}^{N} ln(j) - \sum_{j=1}^{n_c} ln(j) - \sum_{j=1}^{N-n_c} ln(j)))\}$$

and

$$\frac{\binom{K}{i}\binom{N-K}{n_t-i}}{\binom{N}{n_t}} = exp\{(\textstyle\sum_{j=1}^{K} ln(j) - \sum_{j=1}^{i} ln(j) - \sum_{j=1}^{K-i} ln(j) + \sum_{j=1}^{N-K} ln(j) -$$
$$\textstyle\sum j = 1^{n_t-i} lnj - \sum_{j=1}^{N-K-(n_t-i)} -(\sum_{j=1}^{N} ln(j) - \sum_{j=1}^{n_t} ln(j) - \sum_{j=1}^{N-n_t} ln(j)))\}$$

These relationships can be derived by writing the left-hand equations out in factorial form and then using the fact that $x = exp\{ln(x)\}$. Unlike the original fractions, the right-hand formulas can be calculated easily with a computer for very large samples. My website provides R code for calculating the upper limits of $U_t$ and $U_c$ for any sample size, as well as computing the bounds for the RATE.

Like upward bounding, the feasibility of downward bounding depends on the situation. One of its major advantages is that it does not rely on the covariates. Another is that researchers do not need to worry about unrestricted outcomes, which is sometimes the case with upward bounding. However, in cases where more than about 10% of units have undefined outcomes, only very strong treatment effects will survive downward bounding. Another option that researchers can use is these cases is sensitivity analysis, which I develop in the next section.

**Sensitivity Analysis**. In cases where a significant number of units have missing or undefined outcomes, researchers can report the sensitivity of their results to different levels of attrition bias. There are a number of ways that researchers could test the sensitivity of the results. For now, I will propose two straightforward techniques that are easy to interpret. First, for studies with missing outcomes, researchers can do upward sensitivity analysis by examining what values the missing units would need to have to make the results insignificant. For undefined outcomes, researchers can use downward sensitivity analysis, which involves determining what percentage of the units with the highest or lowest outcomes in treatment or control groups would need to have undefined counterfactual outcomes for the results to be insignificant.

The steps for upward sensitivity analysis are as follows. First, find the least extreme value that will make the results insignificant if all missing treatment outcomes are set at that value. Second, do the same for the missing control outcomes. Third, find the standardized values for these least extreme imputed outcomes with respect to the treatment and control groups. Researchers can report the standardized values. They can also try different pairs of imputed values, the first for the treated units and the second the control units, and graph the region where the results remain significant. I construct such a graph for a study about ... on page X.

For downward sensitivity analysis, researchers can simply drop units with the highest or lowest outcomes from the treatment group until the results become insignificant, and then do the same for the control group. They can report these numbers as percentages of treatment or control groups. They could also graph the set of these values were the results will remain significant. A number of interest here is the lowest sum of the pair of values. This number is the minimum amount of units that would need to have unobserved counterfactual outcomes for the results to be insignificant.

My website provides R code for all of the procedures described in this section. Except in cases where there is low attrition or very high treatment effects, sensitivity analysis will usually be a more viable option for researchers than bounding.

**Transforming the Outcomes.** The simplest solution to attrition is sometimes to redefine the outcomes so that all units have outcomes. This is usually easiest when all outcomes are observed but some are undefined. For instance, political scientists are sometimes interested in using regression discontinuity to estimate how winning an election at time $t$ affects a candidates vote share at time $t+1$. However, some winning and losing candidates do not run at time $t+1$, making their outcomes undefined. An easy solution to this problem is to change the outcome to a dummy variable indicating whether the candidate was elected to office at time $t+1$, which gives all candidates an outcome, even if they

**Table 1. Different Scenarios for Studies Facing Attrition**

| Degree of Information | Type of Attrition | Parameter of Interest |
|---|---|---|
| 1. Known Units | 1. Missing Outcomes (M) | 1. Restricted Average Treatment Effect |
| 2. Phantom Units | 2. Undefined Outcomes (U) | 2. Difference in World Averages |
|  | 3. Known Mixture (M, U) |  |
|  | 4. Unknown Mixture (M, U, L) |  |

did not run. The feasibility of redefining the outcome usually depends on the question of interest and the data available, but it can be an adequate solution in some cases.

## Section 3: Versions of the Attrition Problem

Table 1 shows how attrition problems can vary across studies along three different dimensions. The first column lists the degrees of the attrition problem. The most manageable cases are ones where researchers know all of the units that were originally in the sample. In other words, there are no phantom units. This information makes it much easier to use techniques like imputing, bounding, or sensitivity analysis. In other cases, researchers may know that some of the units that were initially assigned to treatment or control are missing. If it is reasonable to assume that treatment status did not affect whether these units were missing from the dataset, then researchers can simply redefine the treatment effect to the effect for the units in their sample. However, if treatment assignment might have affected whether these units were in the sample, then the options moving forward will be limited. Some progress may be made if it is possible to determine (or estimate) the original size of the sample and recreate the missing units, setting their outcomes as lost (L).

The second column shows the possible types attrition for units in the dataset. In the first case, all units are known to have defined outcomes, but some are missing. This scenario allows for imputation, upward bounding, and upward sensitivity analysis. In the second case, all units have observed outcomes, but some are undefined. Principal stratification, downward bounding, and downward sensitivity analysis are possible in this scenario. The third case is a known mixture of the first two, meaning that all missing outcomes are known to be defined. Lastly, there may be an unknown mixture of missing and unknown outcomes. This type of attrition is the most difficult to manage.

The last column shows two main parameters of interest. Recall that both of these parameters are equivalent to the Average Treatment Effect whenever all potential outcomes are defined. On the other hand, when any potential outcomes are undefined, so is the ATE. The Restricted Average Treatment Effect and Difference in World Averages are not only different substantively, but sometimes require different estimation strategies. In general, the DIWA is easier to estimate the RATE. The main difference is that when estimating the DIWA, it is possible to ignore to ignore phantom units, unless researchers want to include the outcomes of units that may not be in the dataset under treatment or control.

These categories are key to determining the strategies that researchers have to address attrition. There can be very promising options, or very limited ones, depending on what

## Section 4: Strategies and Examples

Table 2 shows the viable strategies for each of the general scenarios that could arise under the framework I outline in the previous section. I do not have space here to go through scenario in detail, but there are some general patterns that are worth mentioning before moving on to the examples.

## Table 2. Solutions for Different Scenarios

| Scenario | Listwise Deletion | Imputation | Principal Stratification | Upward Bounding | Downward Bounding | Sensitivity Analysis | Transform the Outcome |
|---|---|---|---|---|---|---|---|
| 1. Known Units/Missing Outcomes/RATE | | ✓ | | ✓ | | ✓ | |
| 2. Missing Units/Missing Outcomes/RATE | | ✓ | | ✓ | | ✓ | |
| 3. Known Units/Undefined Outcomes/RATE | | | ✓ | | ✓ | ✓ | ✓ |
| 4. Missing Units/Undefined Outcomes/RATE | | | ✓ | | ✓ | ✓ | ✓ |
| 5. Known Units/Known Mixture/RATE | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 6. Missing Units/Known Mixture/RATE | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 7. Known Units/Unknown Mixture/RATE | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 8. Missing Units/Unknown Mixture/RATE | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 9. Known Units/Missing Outcomes/DIWA | | ✓ | | ✓ | | ✓ | |
| 10. Missing Units/Missing Outcomes/DIWA | | ✓ | | ✓ | | ✓ | |
| 11. Known Units/Undefined Outcomes/DIWA | ✓ | | | | | | |
| 12. Missing Units/Undefined Outcomes/DIWA | ✓ | | | | | | |
| 13. Known Units/Known Mixture/DIWA | | ✓ | | ✓ | | ✓ | |
| 14. Missing Units/Known Mixture/DIWA | | ✓ | | ✓ | | ✓ | |
| 15. Known Units/Unknown Mixture/DIWA | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 16. Missing Units/Unknown Mixture/DIWA | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Note: The dark checkmarks indicate that the attrition problem can be entirely resolved with the available methods. I use light checkmarks for situations that cannot be entirely resolved, but where the methods could provide useful robustness checks.

The first is that it is very difficult for researchers to use any of these methods if they do not know whether missing outcomes are defined or undefined. Without this information, it is unclear to know whether to impute for certain units or use principal stratification, or upward or downward bound. In this cases, researchers might consider using a variety of these methods to determine the robustness of their results to different techniques, since no single method will be adequate on its own.

The second relates to the extent tow which missing units pose a problem to inference. When datasets are missing units, researchers sometimes spend time and resources hunting down these units and determining what happened to them. In some cases, these efforts can greatly improve a researchers ability to draw inferences. In others, however, there is surprisingly little to be gained from this additional information. For instance, they provide no help in estimating the DIWA, and they do not provide leverage for estimating the RATE when it unclear whether many missing units are defined or undefined. Thus, researches should verify that getting this additional information will be helpful before they invest resources in acquiring it.

Third, the RATE is never easier to estimate than the DIWA. The DIWA is either identified or can be bounded in every case except (15) and (16), which are the cases where researchers do not know whether missing outcomes are defined or undefined. However, researchers are usually more interested in the RATE than the DIWA, because they care about how certain interventions affect individuals. The DIWA is just the difference between the defined treated outcomes and defined control outcomes, which could be partly explained by attrition. Nevertheless, estimating the DIWA could be important in some contexts, and researchers might consider focusing on it when the RATE is difficult to estimate.

I will now move on to discuss three examples of studies that face attrition, starting with a 2009 article by Blattman that tests whether quasi-random conscription into the military made males in Uganda more likely to participate in politics later in life.[6] Blattman uses the fact that boys were as-if randomly abducted and forced into military service at young ages to test whether exposure to war affected their willingness to vote later in life. This study is notable because although there were missing units, Blattman hunted down the names of everyone who was initially in the sample so that no one would be missing from the dataset.

[Note: I will finish this section once I have a better idea about the available methods.]

## Conclusion

This paper investigates one of the major threats to causal inference in experimental research. My intention is not to deter researchers from using real and natural experiments, even when there is substantial attrition. Rather, I simply want to encourage them to think carefully about attrition problems when they arise. There are situations where units with missing outcomes can be dropped, but only can justify doing so. I provide a mathematical framework here that reveals when dropping units may be permissible, even when missingness is not as-if random. When units with missing outcomes cannot be dropped, researchers should select the best available strategy for their study, which will largely depend on the issues that I discuss in this paper.

[6] Blattman 2009

# References

Barnard, John, Constantine E. Frangakis, Jennifer L. Hill, and Donald B. Rubin. 2003. Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City. *Journal of the American Statistical Association* 98(462), 299-323.

Blattman, Christopher. 2009. From violence to voting: War and Political Participation in Uganda. *American Political Science Review* 103(02), 231-247.

Henderson, John. 2012. Demobilizing a Generation: The Behavioral Effects of the Vietnam Draft Lottery. Working Paper.

Zhang, Junni L., and Donald B. Rubin. 2003. Estimation of Causal Effects via Principal Stratification When Some Outcomes Are Truncated by "Death". *Journal of Educational and Behavioral Statistics* 28(4), 353-368.

Rubin, Donald B. 2006. Causal Inference Through Potential Outcomes and Principal Stratification: Application to Studies with" Censoring "Due to Death." *Statistical Science*, 299-309.

Zhang, Junni L., Donald B. Rubin, and Fabrizia Mealli. 2008. Evaluating the Effects of Job Training Programs on Wages Through Principal Stratification. *Advances in Econometrics* 21, 117-145.