When and Why is Attrition a Problem in Randomized Controlled Experiments and How to Diagnose It

Fernando Martel García*
Harvard School of Public Health
124 Mt. Auburn St., Suite 410 South
Cambridge, MA 02138
fmartelg[at]hsph.harvard.edu

First draft October 1, 2012
Current draft January 20, 2013

Abstract

Attrition is the Achilles' Heel of the randomized experiment: it is fairly common, and it can unravel the benefits of randomization. This study considers when and why attrition is a problem, and how it can be diagnosed. The extant literature remains ambiguous because it relies on the language of probability, whereas problematic attrition depends on the underlying causal relations. This ambiguity arises because causation implies correlation but not *vice versa*. Using the structural causal language of directed acyclic graphs I show attrition is a problem when it is an active collider between the treatment and the outcome, or when the latent outcome is a mediator between the treatment and the attrition. Moreover, whether observed outcomes are representative of all outcomes, or only comparable across experimental arms, depends on two d-separation conditions. One of these is directly testable from the data.

*I would like to thank .... All errors are my own.

1 Introduction

Attrition, or missing data on the outcome of interest, is very common in social experiments, impact evaluations, clinical trials, and studies where outcomes are measured using survey instruments, or well after the intervention. Attrition is problematic because it can severely compromise the comparability of units across experimental arms; limit the representativeness of the observed sample with regards to the experimental population of interest; and increase the variance of the estimated effects (Cook and Campbell 1979). Because attrition is so common, and because it has the potential to completely unravel the benefits of randomization, it has been labelled the "Achilles' Heel of the randomized experiment" (Shadish et al. 1998, 3). This study asks when and why attrition is a problem for causal inference in randomized controlled experiments (RCEs), and how problematic forms of attrition can be reliably diagnosed.

The study finds that existing answers to these questions are at best ambiguous, and often misleading; and that this ambiguity can be avoided by relying on the structural causal language of directed acyclic graphs. The study also demonstrates that whenever attrition is an active collider between the treatment and the latent outcome, average treatment effects (ATE) are never identified. And it shows how the possibility of attrition being an active collider can be ruled out by testing whether the treatment causes attrition. Under a weak stability assumption, if attrition is not caused by the treatment then it cannot be a collider, and the ATE for units with observed outcomes is always identified. Furthermore, the ATE for *all* experimental units is always identified under the assumption that the outcome itself is not a cause of the attrition and it has no causes in common with the attrition (other than the treatment). This

assumption is not directly testable from the data.

This study contributes a formal proof that problematic attrition depends on just two *causal* assumptions about the missingness mechanism. It proposes a diagnostic test to license causal inference using the incomplete outcome data – even if no covariates are available. It debunks myths and superstitions, like the belief that attrition is problematic whenever the outcome and attrition are not independent of each other. And it demonstrates how directed acyclic graphs can be used to *infer*, *justify*, and *test* identification strategies including which covariates to control for, what assumptions are plausible, which are testable, and what variables are good instruments (Pearl 2009*a*). Finally, this study contributes a checklist researchers can use to improve the process of causal inference in experiments subject to attrition.

Almost all RCEs experience some degree of attrition. Attrition rates of 30 to 40 percent are common in social experiments (Hausman and Wise 1979; Ashenfelter and Plant 1990; Heckman and Smith 1995; Krueger 1999). In political science, a field experiment studying the effect of media on political behaviour reported attrition rates close to 70 percent for survey data, which is fairly common, and 23 percent for administrative records (Gerber, Karlan and Bergan 2009, 41–43). And a survey of clinical trials found that out of 71 trials 63 (89 percent) reported missing outcome data, with more than 20 percent of patients having missing outcomes in 13 trials (Wood, White and Thompson 2004). As a general rule attrition tends to be higher in longitudinal studies, and in field studies that use survey instruments to measure outcome data.

To answer when and why attrition is a problem for causal inference in RCEs, and how problematic attrition can be diagnosed, this study makes use of the language of directed acyclic graphs. The choice of language is justified by the

following logic: (i) Whether attrition is problematic depends on the underlying causal structure; (ii) probability statements cannot pin down this causal structure; therefore (iii) discussion of attrition using probability statements is necessarily ambiguous.[1] By contrast DAGS can: (i) Encode and communicate researchers' private causal knowledge in unambiguous terms; (ii) highlight assumptions implicit in that knowledge; and (iii) provide justifications for tautological statements like "Missing at Random" (Little and Rubin 2002).

2 Literature review

The methodological literature on attrition in experiments has mostly focused on improving measurement protocols to prevent attrition or minimize its impact, and on analytical techniques for dealing with attrition after it has occurred (see Shadish (2002, 5-6) for a quick overview). Much less has been written about the focus of this study: Defining when and why attrition is a problem in RCEs, and how problematic forms of attrition can be reliably diagnosed.

Jurs and Glass (1971) provide and early and interesting discussion of attrition (which they refer to as mortality). In their view attrition may be problematic depending on whether it is random or non-random within and between comparison groups. These four conditions determine a fourfold table, and for each cell they state whether internal, external, or both types of validity are

[1] For example, the outcome and attrition might be correlated if they cause each other; or if they have causes in common; or if they have an effect in common that is being controlled for (a collider). As this study demonstrates, each of these underlying causal structures has different implications for identification, which joint probability statements cannot disambiguate.

threatened. They then propose using covariates in a MANOVA design to test various restriction and determine in which cell the particular application lies.

As noted by West, Biesanz and Pitts (2000, 53), the tests proposed by Jurs and Glass (1971) may yield erroneous inferences when differential attrition is a function of unobserved variables, or when there is an effect but it is not statistically significant, a point to which I shall return. In addition, because Jurs and Glass (1971) rely on the ambiguous language of probability, their fourfold typology is inherently ambiguous. Consider their universal statement: "if the mortality is systematically related to the treatments, the inferences possible from the reduced sample are not the same as those that could be drawn from the original sample." (Jurs and Glass 1971, 63). Now lets show an exception. Suppose the only causes of attrition are the randomized treatment, and its interaction with a covariate. Suppose no causal relation connects that covariate to the outcome of interest. And suppose treatment also has an effect on outcomes. This implies: (i) more missing outcomes in one group; (ii) missingness that is systematically related to a covariate; and (iii) outcomes that are correlated with the attrition. No matter, the *outcomes* remain both comparable across groups, and representative of the full experimental population. Why? Because the treatment is randomized; the covariate and the outcome do not *cause* each other nor have *causes* in common; and the outcome and the attrition do not *cause* each other and their only *cause* in common is the observed treatment.[2]

Cook and Campbell (1979) advanced the now common belief that problematic

[2]Despite its ambiguity the Jurs and Glass (1971) method remains relatively popular. For example, Kalichman et al. (2001) used it to analyse attrition in a randomized field trial studying the effect of a behavioural intervention on HIV transmission.

attrition can be diagnosed *only* when pre-test measures related to the outcome are available (West, Biesanz and Pitts 2000, 52). Though pre-test measures are obviously useful, this study demonstrates that a key diagnostic is whether treatment has an effect on attrition, which is testable in the absence of any covariates. To their credit Cook and Campbell (1979) do note, correctly, that when attrition is unrelated to the treatment internal validity is not compromised for the population of units whose outcomes are observed. They also point out, correctly, that attrition associated with the treatment *may* compromise internal validity, even for units with observed outcomes. But lacking systematic proof of when and why attrition is a problem their analysis remains ambiguous.

Ambiguous language is the fundamental reason why diagnosing problematic attrition remains mired in confusion. Consider the following statement: "[selective attrition bias] occurs when the underrepresentation of some groups in the longitudinal sample leads to correlations between variables that are different than the true correlations in the original sample" (Miller and Wright 1995, 922). Not only is this statement wrong – such differences are neither necessary nor sufficient for problematic attrition – it is also ambiguous. The reason is Miller and Wright (1995) define attrition using the probability language of symptoms rather than graphical language of causes. Yet, as noted in the Introduction, it is the casual model that defines when attrition is problematic not the covariance structure. The former implies the latter but the reverse is not true.

Not surprisingly field researchers remain unsure when and why attrition is a problem, and how problematic forms of attrition can be reliably diagnosed. Appendix A lists some of the reasons prominent experimenters have given for

why attrition is a problem, and how problematic attrition might be diagnosed. Notably, these studies tend to assert what makes attrition a problem, and then propose diagnostic tests *ex nihilo*: the methodological literature justifying the diagnosis and procedures is seldom cited. This is not to say the explanations and procedures of field researchers are necessarily wrong, often they are sensible, but to highlight how ambiguous they are. As this study will demonstrate, many of the reasons provided for problematic attrition are neither necessary nor sufficient.

The distinction between the language of probability and the language of causality is of fundamental importance (Pearl 2009*b*), and for this reason this study takes a structural approach to causality (e.g. Heckman and Vytlacil 2007; Pearl 2009*a*; Spirtes, Glymour and Scheines 2000). Indeed, because problematic attrition is fundamentally a causal question, so attrition should be discussed exclusively in terms of causes and effects. Within this structural school, Hernán, Hernández-Díaz and Robins (2004) present an analysis of selection bias and confounding that is very close in spirit to the analysis herein. I build on their insights by giving formal proof of when and why attrition is problematic in RCEs; by providing a full partition of mutually exclusive but collectively exhaustive attrition cases where the ATE is identified; and by proposing a test to reliably diagnose whether realized attrition is a problem.

3 Introduction to Directed Acyclic Graphs

3.1 What are Directed Acyclic Graphs (DAGs)

A graph is a collection $\mathcal{G} = \langle \boldsymbol{V}, \boldsymbol{E} \rangle$ of nodes $\boldsymbol{V} = (V_1, \ldots, V_N)^T$ and edges $\boldsymbol{E} = (E_1, \ldots, E_M)^T$ where the nodes correspond to variables and the edges denote the relation between pairs of variables.[3] A DAG is a graph that only admits: (i) *directed* edges with one arrowhead (e.g. $\rightarrow$); (ii) *bi-directed* edges with two arrowheads (e.g. $\leftarrow\!\!-\!\!-\!\!-\!\!-\!\!\rightarrow$); and (iii) no directed cycles (e.g. $X \rightarrow Y \rightarrow X$), thereby ruling out mutual or self causation. Put simply $X \leftarrow\!\!-\!\!-\!\!-\!\!-\!\!\rightarrow Y$ reads "X and Y have unknown causes in common or are confounded" and $X \rightarrow Y$ reads "X causes Y". In addition to directed and bi-directed edges a third type of association between variables in a DAG can be induced by holding constant the effect of two causes, a so-called *collider*. For example, controlling for C in $X \rightarrow C \leftarrow Y$ induces an association between X and Y. Selection bias often results from selecting units on a collider (Hernán, Hernández-Díaz and Robins 2004). A *path* in a DAG is any unbroken route traced along the edges of a graph – irrespective of how the arrows are pointing. A *directed* path, however, is a path composed of directed edges where all edges point in the direction of the path (e.g. $X \rightarrow M \rightarrow Y$). Any two nodes are *connected* if there exists a path between them, else they are *disconnected*.

3.2 What are DAGs useful for?

DAGs provide researchers with a formal language to encode and communicate their private causal knowledge in unambiguous terms. A well specified

[3] I follow closely the presentation in Pearl (2009*b*, §1.2)

DAG embodies a (reduced form) theory of causation. DAGs are specially useful as a means of deriving identification conditions independent of statistical estimation. Relying on the the formal theorems of the graphical language identification conditions can be derived algorithmically. Quite literally, one can often feed a DAG into a computer, ask whether the effect of say X on Y is identifiable conditional on the information encoded in the DAG, and get an automated reply, including the set of conditioning variables that need to be controlled for, or of instruments that are available.[4] In this way DAGs can be used to *infer*, *justify*, and *test* identification strategies including which covariates to control for, what assumptions are plausible, which are testable, and what variables are good instruments (Pearl 2009*a*). DAGs are non-linear and completely non-parametric, so they are silent about the modeling of outcomes once (non-parametric) identification is achieved.[5] Even so, quantities of interest are often estimable non-parametrically.

3.3 Econometric and graphical identification

In the econometrics literature the key identification conditions for causal inference about the effect of Z on Y are unconfoundedness and overlap (Angrist and Pischke 2010). The latter is easy to check form the data but the former

[4]See the package `graph` (Gentleman et al. 2012) for `R` (R Development Core Team 2012). The package allows the researcher to draw a DAG that captures all her causal knowledge and assumptions. It can then identify the identification conditions, if any, for any estimand of interest conditional on the given DAG. This can be used prospectively, to determine what measurements need to be collected, or retrospectively, to determine whether a target estimand can be identified given a multivariate distribution P compatible with $\mathcal{G}$.

[5]If non-parametric identification is not possible, parametric assumptions can be made to yield identification. Typically parametric identification is not accorded as much credibility.

relies on assumptions. These assumptions are typically the subject of great contention, yet they are seldom communicated formally and explicitly.

In the causal language of DAGs the uncounfoundedness assumption can be unbundled into two separate assumptions: (i) that Z and Y have no (uncontrolled) causes in common, and (ii) that common effects of treatment (or the causes of treatment) and the outcome (or the causes of the outcome) are *not* controlled for (Hernán, Hernández-Díaz and Robins 2004). The first is often referred to as no confounding, and the latter as no selection bias. When these two assumptions hold conditional on a given DAG, Z and Y are independently distributed under the null hypothesis of no effect.

In well-conducted randomized controlled trials the first assumption – that treatment Z and outcome Y have no causes in common – is built into the design: randomization rules out any *back-door* paths between between them.[6]. The reason is Z has no causes other than a randomization mechanism under the experimenter's control. However, in the presence of attrition the second assumption – that common effects of treatment (or the causes of treatment) and the outcome (or the causes of the outcome) are *not* controlled for – may not hold. The reason is some units are missing data on the outcome of interest, and discarding these units for analytical purposes can induce selection bias whenever attrition is a collider between the treatment and the outcome. Section 5 provides a formal proof.

[6]Back-door paths between an ordered set of variables (Z, Y) are paths starting from arrows pointing into Z rather than out of Z (e.g. $Z \leftarrow U \rightarrow Y$)

3.4 Useful theorems and definitions

Formally, causal identification requires *d-Separation*:

Definition 1 (d-Separation, Pearl (2009, pp 16-17))**.** *A path p is said to be d-separated (or blocked) by a set of nodes S iff:*

1. *p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in S, or*
2. *p contains an inverted fork (or collider) $i \to m \leftarrow j$ such that the middle node m is not in S and such that no descendant of m is in S.*

A set S is said to d-separate X from Y iff S blocks every path from a node in X to a node in Y.

Following Pearl (2009*b*, p 18) let $(X \perp Y|S)_{\mathrm{P}}$ capture the probabilistic notion of conditional independence and $(X \perp Y|S)_{\mathcal{G}}$ the graphical notion of d-separation in Definition 1. Their connection is established by the following theorem:

Theorem 1 (Probabilistic implications of d-Separation, Pearl (2009, p 18))**.** *For any three disjoint subset of nodes (X, Y, S) in DAG $\mathcal{G}$ and for all probability functions P, we have:*

1. *$(X \perp Y|S)_{\mathcal{G}} \Rightarrow (X \perp Y|S)_P$ whenever G and P are compatible;*[7] *and*

[7] By compatibility Pearl means that the probability distribution admits the factorization $P(x_1, \ldots, x_n) = \prod_i \mathrm{P}(x_i|pa_i)$, where pa_i are the parents of x_i, that is the nodes with directed edges pointing into x_i. For example, $X \leftarrow U \to Y$ factorizes as $\mathrm{P}(X, U, Y) = \mathrm{P}(U)\,\mathrm{P}(X|U)\,\mathrm{P}(Y|U)$.

2. *if* $(X \perp Y|S)_P$ *holds in all distributions compatible with* G*, it follows that* $(X \perp Y|S)_{\mathcal{G}}$.

Theorem 1 allows us to determine whether two variables, or sets of variables, are independent on the basis of d-separation alone.

Definition 2 (Back-door criterion, Pearl (2009, p 79))**.** *A set of variables* S *satisfies the back-door criterion relative to an ordered pair of variables* (X_i, X_j) *in a DAG if:*

1. *no node in* S *is a descendant of* X_i*; and*
2. S *d-separates (or blocks) every path between* X_i *and* X_j *that contains and arrow into* X_i.

Similarly, if X *and* Y *are two disjoint subsets of nodes in* G*, then* S *is said to satisfy the back-door criterion relative to* (X, Y) *if it satisfies the criterion relative to any pair* (X_i, X_j) *s.t.* $X_i \in X$ *and* $X_j \in Y$.

The back-door criterion essentially restricts d-separation to common causes of X and Y. For example, suppose $X \rightarrow M \rightarrow Y$ and $X \leftarrow U \rightarrow Y$. Set $S = \{M, U\}$ d-separates X and Y but only set $S' = \{U\}$ meets the back-door criterion. Intuitively conditioning on M is not useful if our goal is to identify the total (mediated) effect of X on Y, whereas conditioning on U blocks a confounder. Formally:

Theorem 2 (Back-door adjustment, Pearl (2009, p 79))**.** *If a set of variables* S *satisfies the back-door criterion relative to* (X, Y)*, then the causal effect of* X *on* Y *is identifiable and is given by the formula:*

$$P(y|do(x)) = \sum_s P(y|x, s)\, P(s).$$

.

By do(x) Pearl means delete all the arrows pointing into x and set x to any desired level. For example, consider a structural equation system. All DAGs have an equivalent representation as a nonlinear, nonparametric structural equation model with a set of equations of the form $x_i = f_i(pa_i, \varepsilon_i), \quad i = 1, \ldots, N$ (Pearl 2009*b*, p 27), where pa_i denotes the parents of x_i while ε_i captures independently distributed unknown causes. Trivially $x \to y$ is written $y = f_1(x, \varepsilon_y)$. In this context do($x$) involves replacing $x_i = f_2(pa_i, \varepsilon_x)$ with $x_i = x'$, where x' is some realization of X. Consequently, the general formula in Theorem 2 allows us to compute $E[Y|do(x')] - E[Y|do(x'')]$ for any two distinct realizations x' and x'' of X (Pearl 2009*b*, p 70).

The above discussion was simply to convince the reader that d-separation gives us conditional independence and unconfoundedness, and that the back-door criterion is a sufficient though not necessary condition for identification of causal effects. Other identification criteria are available, like the front-door criterion, that will not be considered here (see Pearl (2009*b*); Morgan and Winship (2007)). Hopefully this brief discussion has established how DAGs provide a language for identification that can help researchers infer, justify, and test the exact same identification assumptions called for in econometrics.

4 Examples and intuition

The following three examples provide the intuition behind the formal results in the next section.

4.1 Example 1: Attrition and the latent outcome have causes in common other than the treatment

Consider the DAG in Figure 1. It illustrates the case where attrition R and the latent outcome of interest Y^* are related through unknown causes in common (the bi-directed edge $R \dashleftarrow\dashrightarrow Y^*$). Even so, I will show that conditioning on R allows unbiased estimation of the ATE for those units with observed outcomes only. First, by virtue of randomization, itself an assumption about the chance mechanism allocating experimental units to experimental arms, treatment Z may have effects but no causes for a given experimental population, so there can be no edges pointing into Z.[8]

Second, to focus attention on attrition I assume a specific parametric form for $Y^* \rightarrow Y \leftarrow R$, namely:

$$Y = \begin{cases} Y^*, & \text{if } R = 0 \\ \text{Missing}, & \text{if } R = 1. \end{cases} \tag{1}$$

This rules out measurement error, or the possibility of Z having an effect on the measurement Y but not the latent outcome Y^* (e.g. it rules out the edge $Z \rightarrow Y$). This assumptions help separate attrition from measurement issues. Specifically, Equation 1 ensures $Y^* \equiv Y$ after selection on observables (e.g. after dropping all units with outcomes labelled missing).

[8]Some populations self-select into experiments. For example, administrators may invite randomized evaluations whenever they expect the experimental population to respond positively to a treatment but not otherwise, in which case experimental evaluations are endogenous to the outcome. Even so, conditional on the selected experiment treatment is d-separated from outcomes for this specific population so absent any other threats to inference the local ATE can be estimated consistently.

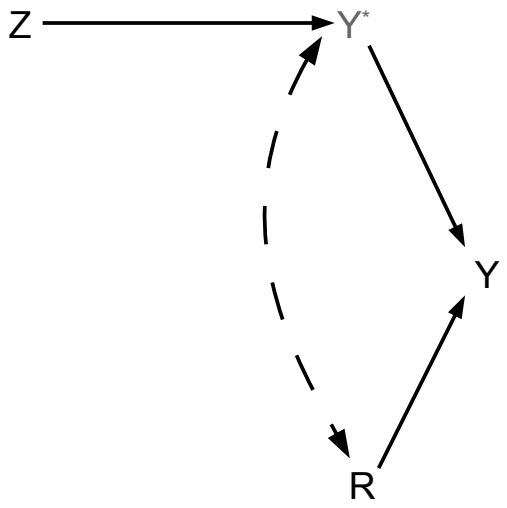

Figure 1: Directed acyclic graph of attrition mechanism $\mathcal{G}_1$. Variable Z is the randomized treatment. By virtue of randomization it has no arrows pointing into it: Z may have effects but not causes. Y^* denotes latent outcomes observed by experimental units but not by the researchers. Attrition R is a binary variable indicating whether Y^* is visible to the researcher ($R = 0$) or not ($R = 1$). By assumption the observed outcome Y equals Y^* if $R = 0$ and is missing otherwise. This assumption rules out measurement error, misreporting caused by treatment (e.g. $Z \rightarrow Y$), and so on to restrict attention to attrition. The bi-directed edge $R \dashleftarrow\dashrightarrow Y^*$ indicates unknown causes in common. These induce a correlation between the latent outcomes and the attrition irrespective of Z. Even so, conditioning on R allows unbiased estimation of the ATE (e.g. the effect $Z \rightarrow Y^*$) for those units with observed outcomes (e.g. $ATE_{R=0}$).

Third, in principle estimating the ATE of Z on Y^* is straightforward. By virtue of randomization there is no backdoor path connecting these two nodes so the set $S = \{\emptyset\}$ d-separates the ordered pair (Z, Y^*). By Theorem 2 the causal effect of Z on Y^* is identifiable and given by the formula:

$$P(y^*|\text{do}(z)) = \text{P}(y^*|z), \tag{2}$$

However, in practice the ATE cannot be estimated using Equation 2 for the simple reason that y^* is not observed, and y is labelled as missing whenever

$r = 1$. For estimation purposes we need to condition on R. In this case the correct formula is:

$$P(y^*|\text{do}(z)) = \text{P}(y^*|z, r = 0)\,\text{P}(r = 0) + \text{P}(y^*|z, r = 1)\,\text{P}(r = 1) \qquad (3)$$

By Equation 1 $\text{P}(y^*|z, r = 0) \equiv \text{P}(y|z, r = 0)$ but $\text{P}(y|z, r = 1)$ is labelled as missing so the ATE for the full experimental group cannot be calculated. What is estimable is $E[y^*|z', r = 0] - E[y^*|z'', r = 0]$ for any two distinct realizations z' and z'' of Z, or the effect of treatment Z on the units with observed outcomes. Note that conditioning on set $S = \{R\}$ still satisfies the back-door criterion (see Definition 2): (i) R is not a descendant of Z and (ii) by virtue of randomization there are no paths between Z and any other node in $\mathcal{G}_1$ that contain an arrow into Z. Interestingly criterion (i) would be violated if the attrition involved self-selection and treatment is effective. In this case set $S = \{R\}$ would be conditioning on a descendant of Z in the directed edge $Z \rightarrow Y^* \rightarrow R$, biasing the estimated effect.

In sum, Figure 1 illustrates a simple case where the outcome of interest Y^* and attrition R are correlated yet conditioning on observables allows an unbiased estimation of $ATE_{R=0}$.

4.2 Example 2: The only cause attrition and the latent outcome have in common is the treatment

Mechanism $\mathcal{G}_1$ above (Figure 1) included unknown common causes of attrition and the latent outcome, as depicted by the bi-directed edge $R \dashleftarrow\dashrightarrow Y^*$. Now consider mechanism $\mathcal{G}_2$ (Figure 2) where treatment is the only cause of

attrition and the latent outcome. As in the previous example, and by virtue of randomization, there is no backdoor path connecting the ordered pair (Z, Y^*) so the set $S = \{\emptyset\}$ d-separates them and causal effects can be calculated using Equation 2. However, for estimation we again need to condition on R, as observed outcomes are labelled missing whenever $R = 1$. But in this case set $S = \{R\}$ does not meet the back-door criterion: R is a descendant of Z. No matter, to estimate the ATE we also need to condition on Z, and conditioning on Z d-separates attrition R from the latent outcome Y^*. This implies $(R \perp Y^*|Z)_P$ and also $P(Y^*|Z, R) = P(Y^*|Z)$. The latter allows unbiased estimation of the ATE for the full population of units in the experiment. By substitution:

$$\begin{aligned} P(y^*|\mathrm{do}(z)) &= \mathrm{P}(y^*|z, r = 0)\,\mathrm{P}(r = 0) + \mathrm{P}(y^*|z, r = 1)\,\mathrm{P}(r = 1) \\ &= \mathrm{P}(y^*|z)\,\mathrm{P}(r = 0) + \mathrm{P}(y^*|z)\,\mathrm{P}(r = 1) \\ &= \mathrm{P}(y^*|z); \end{aligned}$$

and because $\mathrm{P}(y^*|z, r = 0) \equiv \mathrm{P}(y^*|z)$ the ATE for the experimental population can be estimated as $E[Y|do(z'), r = 0] - E[Y|do(z''), r = 0]$ for any two distinct realizations z' and z'' of Z.

This example shows how unbiased estimation of the ATE is possible in some mechanisms exhibiting unbalanced attrition that is directly affected by treatment. And it illustrates how the back-door criterion is sufficient but not necessary for causal identification. One caveat to this example is that estimation is not possible if treatment causes all outcomes to be missing in the treatment group.

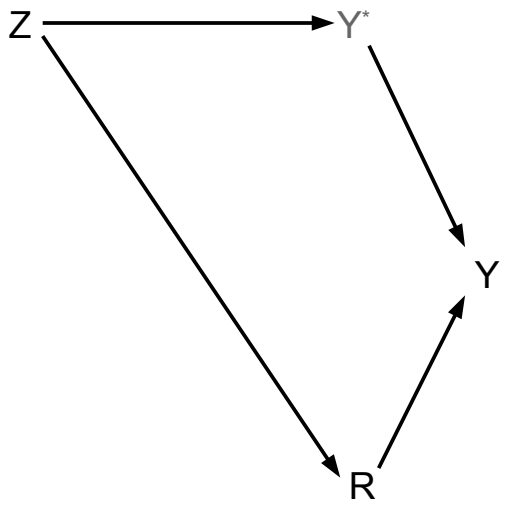

Figure 2: Directed acyclic graph of attrition mechanism $\mathcal{G}_2$. The treatment Z is the only common cause of attrition R and the latent outcome Y^*. The ATE is identified.

4.3 Example 3: Treatment causes attrition, and attrition and the latent outcome have other causes in common

Missingness mechanism $\mathcal{G}_3$ in Figure 3 is a combination of mechanisms $\mathcal{G}_1$ and $\mathcal{G}_2$. As in the previous two examples the problem is not so much identification (set $S = \{\emptyset\}$ satisfies the back-door criterion) but estimation, as some observable outcomes are labelled as missing. However, unlike the previous two examples, where we could estimate the ATE for the units with observed outcomes or for the full experimental population, in the present case neither estimand can be estimated without bias. The intuition is simple: R is now a collider (e.g. it has two edges pointing into it) and controlling for it activates the path $Z \rightarrow R \dashleftarrow\dashrightarrow Y^*$. As a result conditioning on R introduces a dependence between the treatment Z and the latent outcome Y^*, even if the null

is true (e.g. there is no directed edge $Z \to Y^*$). Once again the problem with attrition is not identification so much as estimation of the quantities of interest. The ATE is unconditionally identified but is not identified conditional on R, and for estimation we need to condition on R.[9]

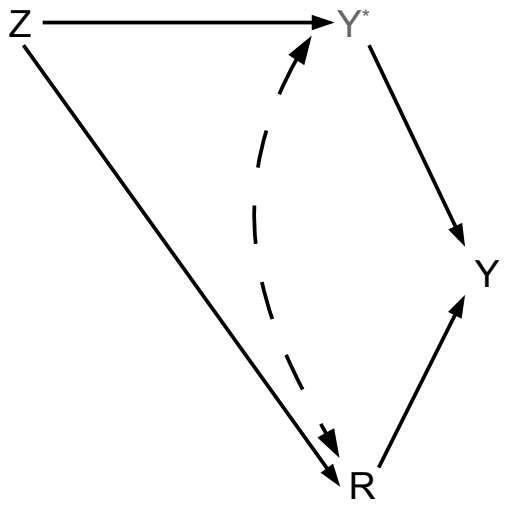

Figure 3: Directed acyclic graph of attrition mechanism $\mathcal{G}_3$. ATE is not identified because R is a collider (has more than one arrow pointing into it) and for estimation purposes we need to control for it (e.g. set R=0). This activates the path $Z \to R \dashleftarrow\dashrightarrow Y^*$ and introduces an association between treatment Z and the latent outcome Y^* even under the null of not direct effect (e.g. deleting the path $Z \to Y^*$)

[9]Jumping ahead, this provides some rationale for imputing the missing outcomes, as it avoids the need to condition on R. Of course the inference will only be as good as the imputation. Since Y^* has unobserved causes this may appear an impossible task. But for imputation all we need is good predictions, so any proxies or correlates of the unmeasured variable may help yield reasonable imputations. This appears more feasible.

$$
\begin{aligned}
Z &= z; \\
Y^* &= f_{Y^*}(Z, \varepsilon_{Y^*}); \\
R &= f_R(Y^*, \varepsilon_R).
\end{aligned}
$$

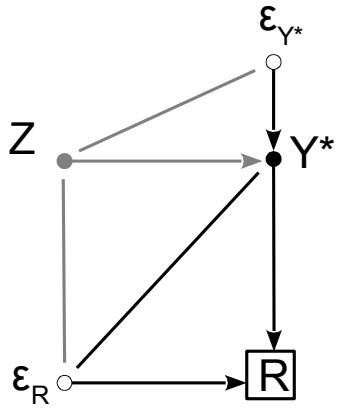

Figure 4: Empty nodes (∘) denote unobserved or unmeasured variables, while full nodes (•) denote measured variables. Under the null of no effect $Y^* = f_{Y^*}(\varepsilon_{Y^*})$ and the DAG reduces to the black directed and non-directed edges. IN this case the ATE is identified and estimable, and ATE=0. Controlling for R introduces the non-directed edge ε_R — Y^* depicting a non-causal association. Under the alternative hypothesis $Z \to Y^*$. If Z causes Y^* then controlling for R introduces two additional non-causal correlations (depicted in gray). In general, controlling for a collider connects *all* parents of the collider.

4.4 Example 4: Treatment causes the outcome which in turn causes the attrition

Since this example is useful for later theorems, I state it as a remark:

Remark 1 (On the implications of controlling for R if $Z \to Y^* \to R$). Let $Z \to Y^* \to R$ be a complete specification of a causal graph. In the context of an experimental manipulation of $Z = z$, the system can be represented graphically as a DAG, and, equivalently, as a non-parametric and non-linear structural equation representation as shown in Figure 4. If Z does not cause Y^* then the ATE is identified and ATE=0, but if Z causes Y^* then conditioning on $R = 0$ activates the back-door paths Z — $\varepsilon_{y^*} \to Y^*$; Z — ε_R — Y^*; and Z — $\varepsilon_R \to R \leftarrow Y^*$. Consequently the direct effect $Z \to Y^*$ cannot be identified separate from these alternative associations.

The point of Remark 1 is to disambiguate. For any intervention Z and outcome Y it is commonly stated that a simple comparison of means across treatment groups identifies the ATE whenever $(Y \perp Z|S)_P$, where set S is empty or includes some "control variables". Implicitly this is a statement about the dependency between Y and Z under the null of no effect. If under the null of no effect Y and Z are assumed independent, then observing a correlation is evidence against the null (implies causation). However in Figure 4 the statement $(Y \perp Z|R = 0)_P$ is true under the null, and the ATE is both identified and estimable (and ATE=0), yet the effect is not identified under the alternative hypothesis.

Taken together these examples give some intuition on the use of DAGs for causal identification in randomized controlled trials subject to attrition. Rather than think of the identification and estimation problem as one of censoring or truncation, DAGs invite us to think of attrition in terms of the underlying causal process and associated quantities of interest.

5 When and why is attrition a problem and how to diagnose it

Attrition can distort the representativeness of the observed sample with regards to the experimental population; the comparability across samples of observed outcomes; and the balance (and power) of observed outcomes across experimental arms. To focus attention on attrition I limit the discussion to the class of Simple Attrition DAGs defined below. Next I prove conditions under which the (i) ATE is identified and estimable despite the attrition; (ii) the

ATE is identified but only $\text{ATE}_{R=0}$ can be estimated; and (iii) the ATE is not identified. Finally I propose some diagnostics tests to license the estimation of unbiased causal effects despite the attrition.

Definition 3 (Simple Attrition DAGs (SADAGs))**.** *A simple attrition DAG is one where: (i) observed outcomes Y that include missing indicators are determined by Equation 1 (an exclusion restriction); (ii) latent outcomes Y^* for any unit i are independent of treatment assigned to any other unit j, $j \neq i$ (non-interference); (iii) a single treatment Z taking two or more values is applied in a randomized fashion (no arrows can point into Z); and (iv) attrition R does not cause Y^*.*

The definition of SADAGs restricts attention to attrition in experimental settings separate from errors in variables, interference, or effects mediated by R. Clearly if the only effect of Z on Y^* is via R, as in $Z \rightarrow R \rightarrow Y^*$, the ATE cannot be estimated: the effect $R \rightarrow Y^*$ is not identified because for $R = 1$ all Y are labelled missing. These restrictions notwithstanding, the intellectual apparatus presented here can be easily generalized to situations where only some or none of these assumptions hold.

Theorem 3 (Identification of ATE conditional on R=0)**.** *For all SADAG defined by the collection $\mathcal{G} = \langle \boldsymbol{V}, \boldsymbol{E} \rangle$, the ATE is not identifiable conditional on R=0 if and only if:*

1. *Z is a parent of R; and*
2. *Y^* and R have causes in common; and*
3. *$\nexists S$ s.t. $(R \perp Y^* | S)_G$ or $(R \perp Z | S)_G$;*

where S is any measured subset of the nodes $\boldsymbol{V}$ not including Y^*.

Proof. ($\Rightarrow$): For any SADAG $\mathcal{G}$, suppose conditions 1–3 above are all true and by contradiction the ATE is identified. Now consider two always mutually exclusive and exhaustive possibilities on $\mathcal{G}$: (i) Y^* is a mediator between Z and R (e.g. $Z \to \ldots \to Y^* \to \ldots \to R$); or (ii) Y^* is not a mediator between Z and R. In case (i) if conditions 1–3 are all true then by Remark 1 the ATE is not identified $\to\leftarrow$. In case (ii) it must be true that R is a collider (by conditions 1 and 2, and by Y^* not a mediator). Whenever R is a collider conditioning on $R = 0$ will result in an association between Z and Y^*, even under the null of no effect, unless $\exists S$ s.t. $(R \perp Y^*|S)_G$ or $(R \perp Z|S)_G$ $\to\leftarrow$. This implies $\neg(Z \perp Y^*|R = 0)_G$ and, by Theorem 1 $\neg(Z \perp Y^*|R = 0)_P$. Therefore, for any SADAG, if conditions 1–3 are all true then the ATE can never be identified.
($\Leftarrow$): For any SADAG $\mathcal{G}$, suppose conditions 1–3 are not all true and by contradiction the ATE is not identified. If conditions 1–3 are not all true, it mus be the case that (i) Z is not a parent of R (and so Y^* cannot be a mediator between Z and R); or (ii) Y^* and R have no causes in common (and so Y^* cannot be a mediator between Z and R); or (iii) $\exists S$ s.t. $(R \perp Y^*|S)_G$ and $(R \perp Z|S)_G$. Although the latter by itself does not rule out the possibility of Y^* being a mediator between Z and R (S could block the mediation if, for example, $Y^* \to S \to R$), for Y^* to be a mediator (i) and (ii) must both be true, and so (iii) cannot also be true. Therefore, whenever conditions 1–3 are not all true Y^* can never be a mediator between Z and R, nor can it be the case that that R is a collider (which implies (i) and (ii) are true) and at the same time $\nexists S$ s.t. $(R \perp Y^*|S)_G$ or $(R \perp Z|S)_G$ (e.g. (iii) also true). Also, by randomization, Z and R can have no causes in common. Together these imply $(Z \perp Y^*|R = 0, S)_G$ and, by Theorem 1 $(Z \perp Y^*|R = 0, S)_P$, which in turn implies the ATE is identified $\to\leftarrow$. Therefore, for any SADAG, if conditions 1–3 are not all true, the ATE is always identified.

□

To say that the ATE is identified does not imply it is estimable, as shown by the next theorem:

Theorem 4 (Estimation of ATE when ATE is identified conditional on R=0)**.** *In any SADAG defined by the collection* $\mathcal{G} = \langle \boldsymbol{V}, \boldsymbol{E} \rangle$*, and assuming the ATE is identified conditional on* $R = 0$*, the ATE is estimable if and only if:*

1. Y^* *is not a parent of* R*, or* Y^* *is a parent of* R *and* $\exists S$ *s.t.* $(Y^* \perp R|S)_G$*; and*

2. Y^* *and* R *have no causes in common, or* Y^* *and* R *have causes in common and* $\exists S$ *s.t.* $(R \perp Y^*|S)_G$*;*

where S *is any (measured) subset of the nodes* $\boldsymbol{V}$*. Otherwise, given that ATE is identified, all we can estimate is the ATE for units with observed outcomes not labelled as missing (*$ATE_{R=0}$*).*

Proof. ($\Rightarrow$): If (1) and (2) are true then it follows from previous arguments that $(Y^* \perp R|Z, S)_G$ (where S includes the empty set). By Theorem 1 this implies $(Y^* \perp R|Z, S)_P$ which, given that the ATE is identified, implies $P(Y^*|do(z)) = \sum_S \sum_R P(Y^*|Z, S, R) = \sum_S (Y^*|Z, S)$ or, equivalently, $\text{ATE}_{R=0} \equiv \text{ATE}$.
($\Leftarrow$): Suppose both (1) and (2) cannot be true. Start with the negation of (1). This implies that Y^* is a parent of R and $\nexists S$ s.t. $(Y^* \perp R|S)_G$. By Theorem 3 this implies that ATE is not identified $\rightarrow\leftarrow$. Consequently (2) must be false. If so Y^* and R have causes in common and $\nexists S$ s.t. $(R \perp Y^*|S)_G$. (Because the ATE is identified, it cannot be the case that Z is a parent of R.) In turn this implies $(Y^* \not\perp R)$ s.t. $P(Y^*|Z, R) \neq P(Y^*|Z)$. Because outcome data

are only observed when $R = 0$ all we can estimate is $\sum_S P(Y^*|Z, R = 0)$ or $ATE_{R=0}$. □

Definition 4 (Representativeness and comparability of observed outcomes not labelled as missing)**.** *The observed outcomes not labelled as missing (Y_{obs}) are said to be:*

Representative *iff the ATE is identified and estimable;*

Comparable *iff the ATE is identified but only $ATE_{R=0}$ is estimable.*

Deriving representativeness or comparability from the causal structure of a DAG is very different from assuming MAR, say, and justifying it willy-nilly. With DAGs researchers work from the ground up. First they write down their explicit causal knowledge in a DAG, then they query the DAG to derive feasible identification and estimation conditions.[10] Ideally the hypothesized attrition mechanism would have been specified in the study protocol and registered in a public registry prior to the experimental intervention. This helps differentiate *post hoc* specification searches from *ex ante* expectations. The implications of Theorems 3, 4, and Definition 4 are summarized in Table 1.

5.1 Diagnosis

In applications subject to attrition researches will want to know to which cell in Table 1 their application belongs to. Knowing this information suffices to

[10]Economists often proceed by writing down a system of structural equations to derive identification strategies. As noted DAGs have an isomorphic representation as a non-linear and non-parametric structural equation system so in principle the two approaches are equivalent. This is not so when the system is written down using specific functional forms, in which case more is being assumed than is strictly needed for the present purposes.

$(Z \perp R\|S)_G$	$(Y^* \perp R\|S)_G$	
	True	False
True	Comparable Representative	Comparable Not representative
False	Comparable Representative	Not comparable Not representative

Table 1: This table summarizes the main implications of Theorems 3, 4, and Definition 4. Given any SADAG $\mathcal{G}$, the table shows how representativeness and comparability of the estimate depend on just two d-separation conditions $(Z \perp R|S)_G$ and $(Y^* \perp R|S)_G$. These can be checked algorithmically for any $\mathcal{G}$.

know which of ATE or $\text{ATE}_{R=0}$ are estimable, if any. Since the focus here is on experiments one possibility is to test whether Z causes R: because Z is exogenous by experimental design, so Z and R can only be dependent if Z causes R (or if we control for a collider in common). Rejecting the null of no effect of Z on R implies Z causes R, and so at least we know we are in the bottom row of Table 1. Otherwise we should consider the top row. This test is unproblematic because R can be treated just like any other outcome of interest (as indeed it is when attrition is caused by death). There is one caveat, however, and to understand it we need to define stability:

Definition 5 (Stability, Pearl (2009, p 48))**.** *Let $I(P)$ denote the set of all conditional independence relationships embodied in P. A causal model $M = \langle D, \Theta_D \rangle$ generates a stable distribution if and only if $P(\langle D, \Theta_D \rangle)$ contains no extraneous independencies – that is, if and only if $I\left(P(\langle D, \Theta_D \rangle)\right) \subseteq I\left(P(\langle D, \Theta'_D \rangle)\right)$ for any set of parameters Θ'_D.*

When we parametrize a DAG structure D (denoted $\mathcal{G}$ in the notation we've been using) with parameters Θ we have a causal model M (e.g. trivially $X \to Y$ becomes a model as in $Y = \alpha + \beta X + \varepsilon_y,\ \varepsilon_y \sim \mathcal{N}(0, \sigma)$ with param-

eters $\Theta = (\alpha, \beta, \sigma)^T$). That model generates distributions $P(\langle D, \Theta_D \rangle)$, and these are likely to vary as we change the parameters. A model is stable by Definition 5 if perturbing the parameters does not destroy any independencies. For example, consider any two pathways by which the treatment Z can affect attrition R: (i) $Z \to R$; and (ii) $Z \to Y^* \to R$. When two or more paths connect Z and R a precise tuning of parameters could negate the effect via one path with a countervailing effect via another path. Observationally Z and R appear independent but this apparent independence is not robust to small perturbations of the parameters. These implications are best illustrated by the truth table in Table 2.

Stability is important because by ruling out such precise tuning of parameters then, and by Theorem 1, it follows that $(Z \perp R|S)_P \Leftrightarrow (Z \perp R|S)_G$ (where S includes the empty set). If this is true then failing to reject the null of $(Z \perp R|S)_P$ implies failure to reject $(Z \perp R|S)_G$, which should focus the researcher's attention on the first row of Table 1. If so $ATE_{R=0}$ is identified and estimable.[11] Moreover, this works for any SADAG, so we do not have to make any more assumptions than those in Definitions 3 and 5. In the context of experiments these are relatively uncontroversial.

6 Discussion

In randomized control trials the ATE is always unconditionally identified: problematic attrition arises, if at all, when estimating the ATE. In general

[11] This is one instance where testing the null of exactly zero effect really makes sense: A small effect is not always evidence of a small problem if big effects are almost offsetting one another.

P	Q	W	¬P	∧	¬Q	∨	W	$(Z \perp R\|S)_G$	$(Z \perp R\|S)_P$
F	F	F	T	T	T	T	F	T	T
F	T	F	T	F	F	F	F	F	F
T	F	F	F	F	T	F	F	F	F
T	T	T	F	F	F	T	T	**F**	**T**

Table 2: Truth table depicting implications of failure to reject the null hypothesis that treatment Z does not cause attrition R using information from the observed distribution, where $P =$"$Z \rightarrow R$"; $Q =$"$Z \rightarrow Y^* \rightarrow R$"; and $W =$"$P \wedge Q \wedge$ they exactly offset". Assuming stability (Definition 5) rules out the last row of the table where independence of Z and R in the statistical sense does not imply their independence in the underlying causal structure.

there are two ways in which the problem comes up. First, if researchers drop the units with missing outcome data they are fixing $R = 0$. This is problematic whenever R is an active collider between the treatment Z and the outcome Y^*. Second, in principle the problem can be avoided if we impute the missing data (e.g. Rubin 1976, 2004); or do so implicitly using inverse probability weights (e.g. Robins, Rotnitzky and Zhao 1995); or impute bounds (e.g. Horowitz and Manski 2000), as these remove the need to condition on R. But in all these cases the point or interval estimates will only be as good as the imputation.

That said, imputation only requires good predictions. Hence, even if causes of both attrition and the outcome are not observed their proxies or correlates might be, and these might provide a decent imputation. By contrast, parametric selection models work very differently (see Greene (2008, §24.5) for an overview). Rather than predict the missing outcomes they impute the unobserved variable (λ) that is supposedly causing both attrition and the outcome, such that $(Y^* \perp R|\lambda)_G$. This obviates condition 3 in Theorem 3 and so the ATE is identified so long as attrition is not a function of the latent outcome. But in this case the imputation is made on the basis of parametric assumptions and a structural model for the attrition. Arguably these are more restrictive procedures than the multiple imputation approach. But these judgements can

only be made conditional on the DAG (e.g. what the researcher claims she knows about the causal structure) and the available measures.

Because attrition is typically problematic whenever it acts as a collider between the treatment and the latent outcome, researchers should think hard about what variables might cause both attrition and the outcome, and consider ways of measuring them as part of their measurement protocol. If problematic attrition is expected on the basis of existing causal knowledge, then researchers might consider two-stage sampling, whereby a random sample of attriters in the first round of surveys is surveyed intensively in the second round with the objective of collecting responses from all of them (see Lohr (1999, §8.3)). Ideally this would be done at baseline, so problems can be detected early. Such data can be used to test whether outcomes and attrition are correlated, which, together with a test of whether treatment causes attrition, can pinpoint which cell of Table 1 this particular application falls into. These data can also improve and check multiple imputations, as it provides an estimated mean and variance for non-respondents.

7 Conclusion

Attrition has been labelled the Achilles' Heel of the randomized experiment: it is very common, and it can completely unravel the benefits of randomization. Much of the extant literature has approached attrition using the language of probability. This language is inherently ambiguous. Whereas causation implies correlation the reverse is not true, and yet it is underlying causal relations that determine when and why attrition is a problem. Not surprisingly researchers' discussion of attrition remains ambiguous (see Appendix A). This

study recommends attrition be always discussed in terms of causes and effect. The structural causal language of DAGs is ideally suited to this purpose, and makes the discussion simpler by virtue of being non-parametric and applicable to linear and non-linear problems.

Attrition is a problem when the attrition variable R is an active (e.g. unblocked) collider between treatment Z and the latent outcome Y^*, or whenever the latent outcome is an unblocked mediator between the treatment and the attrition. In the first case attrition is a problem because conditioning on $R = 0$ activates the collider path connecting the treatment and the latent outcome, even under the null that treatment has no effect on the outcome. In the second case attrition is a problem because conditioning on attrition introduces an association between unknown causes of R and Y^*, and Z. Hence the estimated effect includes the true effect plus a measure of these correlations (see Figure 4). In terms of diagnosis this study has shown that whether observed outcomes are representative of all experimental outcomes (observed and unobserved), and whether they are comparable across experimental arms depends on two d-separation conditions: $(Z \perp R|S)_G$ and $(Y^* \perp R|S)_G$. Whenever the latter is true observed outcomes are always comparable and representative. Otherwise they are comparable but not representative if the former is true, while the ATE is not even identified otherwise. Condition $(Z \perp R|S)_G$ is testable form the data, and tests for $(Y^* \perp R|S)_G$ can be had under the right measurement protocol (e.g. two-stage sampling).

This study has focused mostly on answering when and why attrition is a problem in randomized controlled experiments, and how problematic attrition might be reliably diagnosed. Less attention has been paid to how to deal with problematic attrition. Moreover, attention has been focused on conditioning

strategies for identification. This ignores other possibilities, like the use of instruments for attrition (see (Huber 2012)). Even so, the framework presented here provides an unambiguous language to propel these efforts forward.

References

Angrist, Joshua D. and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24(2):3–30.

Ashenfelter, Orley and Mark W. Plant. 1990. "Nonparametric Estimates of the Labor-Supply Effects of Negative Income Tax Programs." *Journal of Labor Economics* 8(1):pp. S396–S415.

Cook, T.D. and D.T. Campbell. 1979. *Quasi-experimentation: design & analysis issues for field settings*. Rand McNally College.

Duflo, Esther, Pascaline Dupas and Michael Kremera. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *The American Economic Review* 101(5):1739–1774.

Duflo, Esther, Rachel Glennerster and Michael Kremer. 2007. Chapter 61 Using Randomization in Development Economics Research: A Toolkit. Vol. 4 of *Handbook of Development Economics* Elsevier pp. 3895 – 3962.

Gentleman, R., Elizabeth Whalen, W. Huber and S. Falcon. 2012. *graph: A package to handle graph data structures*. R package version 1.34.0.

Gerber, Alan S., Dean Karlan and Daniel Bergan. 2009. "Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting

Behavior and Political Opinions." *American Economic Journal: Applied Economics* 1(2):35–52.

Gerber, alan S., gregory A. Huber and ebonya Washington. 2010. "Party Affiliation, Partisanship, and Political Beliefs: A Field Experiment." *American Political Science Review* 104:720–744.

Greene, William H. 2008. *Econometric Analysis*. Pearson International Edition.

Hausman, Jerry A. and David A. Wise. 1979. "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment." *Econometrica* 47(2):455–473.

Heckman, James J. and Edward J. Vytlacil. 2007. Chapter 70 Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation. Vol. 6, Part B of *Handbook of Econometrics* Elsevier pp. 4779 – 4874.

Heckman, James J. and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *The Journal of Economic Perspectives* 9(2):85–110.

Heckman, James J., Robert J. Lalonde and Jeffrey A. Smith. 1999. The economics and econometrics of active labor market programs. In *Handbook of Labor Economics*, ed. Orley C. Ashenfelter and David Card. Vol. Volume 3, Part 1 Elsevier pp. 1865–2097.

Hernán, Miguel A., Sonia Hernández-Díaz and James M. Robins. 2004. "A Structural Approach to Selection Bias." *Epidemiology* 15(5):615–625.

Horowitz, Joel L. and Charles F. Manski. 2000. "Nonparametric Analysis

of Randomized Experiments with Missing Covariate and Outcome Data." *Journal of the American Statistical Association* 95(449):77–84.

Huber, Martin. 2012. "Identification of Average Treatment Effects in Social Experiments Under Alternative Forms of Attrition." *Journal of Educational and Behavioral Statistics* 37(3):443–474.

Jurs, Stephen G. and Gene V. Glass. 1971. "The Effect of Experimental Mortality on the Internal and External Validity of the Randomized Comparative Experiment." *The Journal of Experimental Education* 40(1):62–66.

Kalichman, Seth C, David Rompa, Marjorie Cage, Kari DiFonzo, Dolores Simpson, James Austin, Webster Luke, Jeff Buckles, Florence Kyomugisha, Eric Benotsch, Steven Pinkerton and Jeff Graham. 2001. "Effectiveness of an intervention to reduce HIV transmission risks in HIV-positive people." *American Journal of Preventive Medicine* 21(2):84 – 92.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics* 114(2):497–532.

Little, Roderick .J.A. and Donald B. Rubin. 2002. *Statistical analysis with missing data*. Wiley New York.

Lohr, S.L. 1999. *Sampling: Design And Analysis*. Advanced Series Brooks/Cole.

Miller, Richard B. and David W. Wright. 1995. "Detecting and Correcting Attrition Bias in Longitudinal Family Research." *Journal of Marriage and Family* 57(4):pp. 921–929.

Morgan, Stephen L. and Christopher Winship. 2007. *Couterfactuals and*

Causal Inference: Methods and principles of Social Research. Cambridge Univ. Press.

Pearl, J. 2009*a*. "Causal inference in statistics: An overview." *Statistics Surveys* 3:96–146.

Pearl, Judea. 2009*b*. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

R Development Core Team. 2012. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

Robins, James M., Andrea Rotnitzky and Lue Ping Zhao. 1995. "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data." *Journal of the American Statistical Association* 90(429):106–121.

Rubin, D.B. 2004. *Multiple Imputation for Nonresponse in Surveys*. Wiley Classics Library John Wiley & Sons.

Rubin, Donald B. 1976. "Inference and missing data." *Biometrika* 63(3):581–592.

Shadish, William R. 2002. "Revisiting Field Experimentation: Field Notes for the Future." *Psychological Methods* 7(1):3–18.

Shadish, W.R., X. Hu, R.R. Glaser, R. Kownacki and S. Wong. 1998. "A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes." *Psychological Methods* 3(1):3.

Spirtes, P., C.N. Glymour and R. Scheines. 2000. *Causation, Prediction, and Search*. Adaptive Computation and Machine Learning Mit Press.

West, S.G., J.C. Biesanz and S.C. Pitts. 2000. Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In *Handbook of research methods in social and personality psychology*. Cambridge University Press, Cambridge, UK chapter 3, pp. 40–84.

Wood, Angela M, Ian R White and Simon G Thompson. 2004. "Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals." *Clinical Trials* 1(4):368–376.

A Field researchers' explanations of when attrition is a problem and diagnostic tests

Without pretending to be systematic the following are common rationales for why attrition is a problem. These are quoted not because they may be inappropriate in the context in which they are stated, so much as to illustrate the variety of ways in which attrition is deemed to be a problem. This study tries to unify and simplify these criteria, disambiguate the definition of problematic attrition, and provide a diagnostic test consistent with the definition.

A.1 When or why attrition is a problem:

- "If the probability of attrition is correlated with experimental response, then traditional statistical techniques will lead to biased and inconsistent estimates of the experimental effect." (Hausman and Wise 1979, 456)

- "While the attrition in these surveys has typically not been as severe as in social experiments, the same problems of potential bias arises, if attrition is not random"(Hausman and Wise 1979, 456)

- "If there is attrition based on unobserved variables that are correlated with the outcome measures but not predicted by the observables, our results may be biased." (Gerber, Karlan and Bergan 2009, 43)

- "Sample attrition poses a problem for experimental evaluations when it is correlated with individual characteristics or with the impact of treatment conditional on characteristics. In practice, persons with poorer labor market characteristics tend to have higher attrition rates (see, e.g., Brown, 1979). Even if attrition affects both experimental and control groups in the same way, the experiment estimates the mean impact of the program only for those who remain in the sample. Usually, attrition rates are both non-random and larger for controls than for treatments. In this case, the experimental estimate of training is biased because individuals' experimental status, R, is correlated with their likelihood of being in the sample. In this setting, experimental evaluations become non-experimental evaluations because evaluators must make some assumption to deal with selection bias." (Heckman, Lalonde and Smith 1999, x).

- "attrition may be a serious problem if it is not random" (Ashenfelter and Plant 1990, S402); "If it is not random, then the reliance on the assumption of random assignment is unjustified, and, in fact, nonparametric estimation is impossible. When there is attrition, some assumption must be made regarding the behavior of the families who left the experiment and how it differs from those remaining."(Ashenfelter and Plant

1990, S408); "The fact of attrition necessitates some sort of "parametric" assumption about the sample."(Ashenfelter and Plant 1990, S410)

- "In the JTPA evaluation, the parameter of interest was the mean impact of JTPA training on those receiving it. Given this parameter of interest, random assignment should be placed so as to minimize attrition from the program within the treatment group, because the experimental mean-difference estimate corresponds exactly to the impact of training on the trained only in the case where there is no attrition."(Heckman and Smith 1995, 102); "In the presence of this level of attrition, the only way to obtain a credible estimate of the mean impact of the program is to model the attrition process using the very non-experimental methods eschewed by proponents of randomized social experiments. "(Heckman and Smith 1995, 103)

- "At heart, adjusting for possible nonrandom attrition is a matter of imputing test scores for students who exited the sample. With longitudinal data, this can be done crudely by assigning the student's most recent test percentile to that student in years when the student was absent from the sample." (Krueger 1999, 515).

- "Random attrition will only reduce a study's statistical power; however, attrition that is correlated with the treatment being evaluated may bias estimates. [...] Even if attrition rates are similar in treatment and comparison groups, it remains possible that the attritors were selected differently in the treatment and comparison groups." (Duflo, Glennerster and Kremer 2007, 3943).

- "Whenever substantial attrition occurs this is cause for major concern because there is a possibility that the attrition is nonrandom and may

lead to bias." (Gerber, Huber and Washington 2010, 727).

A.2 Diagnostics and tests for analyzing attrition

In terms of diagnostics for determining whether attrition is a problem:

- "The key question is whether the experimental treatments are changing the rate of attrition of [units in the experimental group]"(Ashenfelter and Plant 1990, S408)

- "There was no difference between tracking and nontracking schools in overall attrition rates. The characteristics of those who attrited are similar across groups, except that girls in tracking schools were less likely to attrit in the endline test [...] The proportion of attritors and their characteristics do not differ between the two treatment arms"(Duflo, Dupas and Kremera 2011, 1752)

- "A first step in the analysis of an evaluation must always be to report attrition levels in the treatment and comparison groups and to compare attritors with non-attritors using baseline data (when available) to see if they differ systematically, at least along observable dimensions. If attrition remains a problem, statistical techniques are available to identify and adjust for the bias. These techniques can be parametric (see Hausman and Wise, 1979; Wooldridge, 2002 or Grasdal, 2001) or nonparametric. We will focus on non-parametric techniques here because parametric methods are more well known. Moreover, non-parametric sample correction methods are interesting for randomized evaluation, because they do not require the functional form and distribution assumptions

characteristic of parametric approaches. Important studies discussing non-parametric bounds include Manski (1989) and Lee (2002)" (Duflo, Glennerster and Kremer 2007, 3944).

- "To test for differential attrition across conditions, a 2 attrition (lost vs retained) × 2 condition (risk-reduction skills intervention vs comparison) contingency table, chi-square test was performed; 44 of 187 (24%) participants in the experimental condition and 33 of the 145 (23%) comparison-intervention participants were lost at follow-up, a nonsignificant difference. We also conducted attrition analyses for differences on baseline measures using 2 (attrition) × 2 (condition) analyses of variance,23 where: (1) an attrition effect signals differences between participants lost and retained, (2) an intervention effect indicates a breakdown in randomization, and (3) an attrition x condition interaction indicates differential loss between conditions. Results indicated a main effect for attrition on participant age: participants lost to follow-up were younger ($M = 37.9$) than those retained ($M = 40.7$). No other differences between lost and retained participants were significant, there were no significant differences between intervention conditions on baseline variables, and the attrition by intervention-group interactions were not significant." (Kalichman et al. 2001, 87).