Testing Attrition Bias in Field Experiments*

Dalia Ghanem
UC Davis

Sarojini Hirshleifer
UC Riverside

Karen Ortiz-Becerra
UC Davis

March 27, 2019

Latest Version here.

Abstract

Attrition is a common threat to the internal validity of field experiments in economics. We conduct a systematic review of the field experiment literature in economics and find no consensus on how to test for attrition bias. Bringing insights from the panel identification literature, we formalize this question in the presence of baseline outcome data. We differentiate between two populations of interest, the respondent subpopulation and the study population. We establish the identifying assumptions for treatment effects for each population and their sharp testable restrictions on the baseline outcome distribution. These restrictions consist of joint hypotheses of distributional equality. We propose randomization procedures to obtain p-values for the test statistics of these hypotheses. We further demonstrate that the most commonly used test in the literature, the differential attrition rate test, is not an appropriate test of internal validity in general. Finally, we illustrate our results numerically using simulations.

Keywords: attrition, randomized experiments, randomized controlled trials, internal validity, Kolmogorov-Smirnov, Cramer-von-Mises, randomization tests
JEL Codes: C12, C21, C33, C93

*Working Draft. Comments are welcome: dghanem@ucdavis.edu, sarojini.hirshleifer@ucr.edu, kaortizb@ucdavis.edu.
We thank Alberto Abadie, Josh Angrist, Federico Bugni, Jia Li, Aprajit Mahajan, Matthew Masten, Craig McIntosh, David McKenzie, Ismael Mourifié, Adam Rosen and Monica Singhal for helpful discussions. We also thank participants at the Duke econometrics seminar and the UC Davis development workshop for valuable comments.

1 Introduction

Randomized control trials (RCTs) are an increasingly important tool of applied economics since, when properly designed and implemented, they can produce internally valid estimates of causal impact.[1] Non-response on outcome measures at endline, however, is an unavoidable threat to the internal validity of many carefully implemented trials. Long-distance migration can make it prohibitively expensive to follow members of an evaluation sample. Conflict, intimidation or natural disasters sometimes make it unsafe to collect complete response data. The recent, increased focus on the long-term impacts of interventions has also made non-response especially relevant. Thus, researchers often face the question: How much of a threat is attrition to the internal validity of a given study?

In this paper, we approach attrition in field experiments with baseline outcome data as an identification problem in a nonseparable panel model. We focus on two identification questions generated by attrition in field experiments. First, does the difference in mean outcomes between treatment and control respondents identify the average treatment effect for the respondent subpopulation? Second, is this estimand equal to the average treatment effect for the study population?[2] To answer these questions, we examine the testable implications of the relevant identifying assumptions and propose statistics to test them. Our results provide insights on current empirical practice in testing attrition bias in field experiments.

We first conduct a systematic review of 88 recent journal articles that report the results of field experiments with baseline data in order to understand both the extent of attrition as well as to document how authors test for attrition bias.[3] First, we note that attrition is common in published field experiments. Specifically, 43% of the reviewed field experiments have an attrition rate that is higher than 15%, and 66% have an attrition rate higher than 5%.[4] We also find that tests of attrition bias are widely used. The majority of papers (89%) with an attrition rate higher than 1% conduct such a test. We find that there is no consensus in the literature on how to test for response bias, however. We identify two main types of tests: (i) a *differential attrition rate test* that determines if attrition rates are different across treatment and control groups, and (ii) a *selective attrition test* that determines if the mean of baseline

[1] Since in the economics literature the term "field experiment" generally refers to a randomized controlled trial, we use the two terms interchangeably in this paper. We do not consider "artefactual" field experiments, also known as "lab experiments in the field," since attrition is often not relevant to such experiments.

[2] Following Andrews and Oster (2017), the study population is the population that was selected for the evaluation. Identification of treatment effects for this population may or may not have implications for external validity depending on the specific context of the study (Andrews and Oster, 2017; Azzam et al., 2018).

[3] Our review covers papers published in nine respected economics journals over the years 2009 to 2015. We note that findings are relevant to a much larger group of papers since a number of other major economics journals also commonly publish field experiments, and the number of published field experiments with baseline data grew by a factor of eight over the course of the years that we looked at (and continues to grow).

[4] We only consider attrition rates relevant to results reported in the abstract.

outcomes differs across the treatment and/or control groups conditional on response status. The use of attrition tests and their implementation vary widely across papers. For instance, while authors report a differential attrition rate test for 80% of field experiments, authors report a selective attrition test only 60% of the time.[5]

Next, we present a formal treatment of attrition in field experiments with baseline outcome data. Specifically, we establish the identifying assumptions in the presence of attrition for two cases that are likely to be of interest to the researcher. For the first case, in which the researcher's objective is internal validity for the respondent subpopulation, the identifying assumption is random assignment conditional on response status. In the second case, where internal validity for the study population is of interest, the identifying assumption is "missing-at-random" as defined in Manski (2005) in addition to the initial random assignment of the treatment.[6] This second case is especially relevant in settings where the study population is representative of a larger population.

We then derive testable restrictions for each of the above identifying assumptions. The assumption required for internal validity for respondents implies a joint hypothesis of two equalities on the baseline outcome distribution; specifically, for treatment and control respondents as well as treatment and control attritors. Meanwhile, the assumption required for internal validity for the study population implies a joint hypothesis of equality on the baseline outcome distribution across all four treatment/response subgroups. The approach presented in this paper highlights that a test of attrition bias is a test of an identifying assumption, which (like other identifying assumptions) can only be tested by implication in general.[7] Hence, we prove that the aforementioned testable restrictions are sharp, meaning that they are the strongest implications that we can test given our data.[8]

Since the assumptions required for identification are random-assignment-type restrictions, randomization tests are a natural choice in this context.[9] We therefore propose "subgroup"-randomization procedures in the spirit of Lehmann and Romano (2005, Chapter 5.11) to approximate exact p-values for Kolmogorov-Smirnov (KS) and Cramer-von-Mises (CM) statistics of the sharp testable restrictions mentioned above. We further extend this approach to testing

[5]We chose broad definitions for all type of attrition tests. The categorization of estimation strategies is described in Appendix A. All analysis of attrition tests is restricted to papers with greater than 1% attrition.

[6]A sufficient condition for missing-at-random in our context is that the unobservables that affect response and outcome are independent.

[7]As a result, a rejection of a testable implication of an identifying assumption implies that the identifying assumption does not hold. The converse is not true, however.

[8]Sharp testable restrictions are the restrictions for which there are the smallest possible set of cases such that the testable restriction holds even though the identifying assumption does not. The concept of sharpness of testable restrictions was previously developed by Kitagawa (2015), Hsu et al. (2016), and Mourifié and Wan (2017).

[9]The mean versions of our sharp testable restrictions for both internal validity for respondents and the population can be implemented using simple regression tests which we outline in Appendix D of this paper.

for attrition bias given stratified randomization and to identify heterogeneous treatment effects.

We also provide a formal treatment of the differential attrition rate test, since it is the most frequently used in the field experiment literature. In order to do so, we apply the framework of partial compliance from the local average treatment effect (LATE) literature to potential response.[10] We demonstrate that even though equal attrition rates are sufficient for internal validity for the respondent subpopulation under additional assumptions, they are not a necessary condition for internal validity in general. Hence, the differential attrition rate test is not an appropriate test of internal validity in the absence of *a priori* information.

Finally, we conduct a simulation experiment to illustrate our analytical results. In our designs, we consider cases in which there are either differential or equal attrition rates. We also vary whether there is internal validity for the study population, for respondents only, or no internal validity at all. We find that the mean and distributional tests of internal validity for respondents (the study population) only reject at a higher-than-nominal level when internal validity for respondents (the study population) is violated. In contrast, the differential attrition rate test: (i) does not control size in some cases when internal validity holds,[11] and (ii) has trivial power in some cases when internal validity is violated.

Our results have important empirical and policy implications. The theoretical results shed light on the testing procedures used in the field experiment literature. In particular, we find that the mean implication of the sharp testable restriction of internal validity for respondents is equivalent to the null hypothesis of the joint selective attrition test on respondents and attritors. This version of the test, however, is only performed on 12% of reviewed field experiments. In addition, the differential attrition rate test, which is more widely implemented, is not based on a necessary condition of internal validity for respondents in general and hence may lead to a false rejection of internal validity in practice. Attrition in a given study is often used as a metric to evaluate the study's reliability to inform policy. For instance, *What Works Clearinghouse*, an initiative of the U.S. Department of Education, has specific (differential) attrition rates standards that a study must meet in order to be deemed reliable (WWC, 2017). The results in this paper suggest potential revisions to these standards.

This paper contributes to a growing literature that considers methodological questions relevant to field experiments.[12] Given the wide use of attrition tests and their policy relevance,

[10] See the foundational work in the LATE literature (Imbens and Angrist, 1994; Angrist et al., 1996) as well as Vytlacil (2002) who shows the relationship between sample selection models and the LATE framework.

[11] That is, it can overreject when the null hypothesis of internal validity for respondents holds.

[12] Bruhn and McKenzie (2009) compare the performance of different randomization methods; McKenzie (2012) discusses the power trade-offs of the number of follow-up samples in the experimental design; Baird et al. (2018) propose an optimal method to design field experiments in the presence of interference; de Chaisemartin and Behagel (2017) present how to estimate treatment effects in the context of randomized wait lists; Abadie et al. (2018) propose alternative estimators that reduce the bias resulting from endogenous stratification in field experiments.

this paper is focused on the testing problem. There is a thread in this literature however that outlines various approaches to correcting attrition bias in field experiments (e.g. Behagel et al., 2015; Millán and Macours, 2017). This paper also relates to recent work that examines the potential uses of randomization tests in analyzing field experiment data (Young, 2018; Athey and Imbens, 2017; Athey et al., 2017; Bugni et al., 2017).

The attrition corrections in the field experiments literature build on the larger sample selection literature in econometrics going back to Heckman (1976, 1979). Nonparametric Heckman-style corrections have been proposed for linear and nonparametric outcome models (e.g. Ahn and Powell, 1993; Das et al., 2003).[13] Inverse probability weighting is another important category of corrections for sample selection bias (e.g. Angrist, 1997; Wooldridge, 2007).[14] The aforementioned approaches assume unconfoundedness or selection on observables. Nonparametric bounds, on the other hand, is an alternative approach which requires weaker assumptions. Horowitz and Manski (2000), Manski (2005) and Kline and Santos (2013) propose bounds on the conditional outcome distribution and related objects of interest. This sample selection literature is concerned with objects that pertain to the population, i.e. study population in our framework. Lee (2009) bounds the average treatment effect for a subpopulation of respondents, the always-responders, assuming monotonicity of selection in the context of a randomized experiment. This paper provides tests of identifying assumptions for both the (study) population and the respondent subpopulation.

We also build on other strands of the econometrics literature. Recent work on nonparametric identification in nonseparable panel data models informs our approach (Altonji and Matzkin, 2005; Bester and Hansen, 2009; Chernozhukov et al., 2013; Hoderlein and White, 2012; Ghanem, 2017). Specifically, the identifying assumptions relevant here fall under the nonparamteric correlated random effects category (Altonji and Matzkin, 2005; Bester and Hansen, 2009). Furthermore, we build on the literature on randomization tests for distributional statistics (Dufour, 2006; Dufour et al., 2008).

The paper proceeds as follows. Section 2 presents the review of the field experiments literature. Section 3 formally presents the identifying assumptions and their sharp testable restrictions. In Section 4, we propose a subgroup-randomization procedure to obtain p-values for the distributional test statistics. Section 5 presents simulation experiments to illustrate the theoretical results. Section 6 concludes.

[13] Vella (1998) provides a detailed review of this category of sample selection models. See Brownstone (1998) for some interesting discussions on several sample selection corrections and the trade-offs between them.

[14] Angrist (1997) examines the connection between Heckman-style corrections and the inverse probability weighting approach.

2 Attrition in the Field Experiment Literature

We systematically reviewed recent papers published in economics journals that report the results of field experiments in order to understand both the extent to which attrition is observed and how authors test for attrition bias. We identify two main types of tests that aim to determine the impact of attrition on internal validity: (i) a *differential attrition rate test*, and (ii) a *selective attrition test*. A *differential attrition rate test* simply determines whether the rates of attrition are statistically significantly different across treatment and control groups. In contrast, a *selective attrition test* determines whether, conditional on being a respondent and/or attritor, the mean of observable characteristics is the same across treatment and control groups. We further categorize selective attrition tests as *simple* or *joint* tests. A simple test determines whether treatment and control groups are the same for respondents *or* attritors, while a joint test determines whether treatment and control samples are the same for respondents and attritors jointly. Finally, we consider whether authors conduct a test of whether baseline outcomes and covariates are identically distributed across respondents and attritors. We call this a *determinants of attrition test*. This categorization of tests for attrition bias imposes some structure on the variety of different estimation strategies use to test for attrition bias in the literature.[15]

We focused our review on field experiments for which baseline outcome data is available. Without baseline outcome data, selective attrition tests are less likely to be informative. To select our sample, we first identified papers that report the results of field experiments, and that were published in nine highly regarded (applied and general interest) economics journals between 2009 and 2015.[16] We restrict our analysis here to 88 articles in which the authors had baseline data on at least one main outcome variable.[17] Although some articles report a wide range of attrition rates for various sub-samples, we focus here on attrition rates that are relevant to outcomes reported in the abstract (we call these "abstract results").[18]

In order to understand the extent of attrition in field experiments, we reviewed reported overall and differential attrition rates. We find that attrition is common in field experiment

[15]In our review, we identified three estimation strategies that were used to conduct a differential attrition rate test, nine estimation strategies used to conduct a selective attrition test, and four approaches to conducting a determinants of attrition test. We present details on these approaches to estimation in Appendix A.

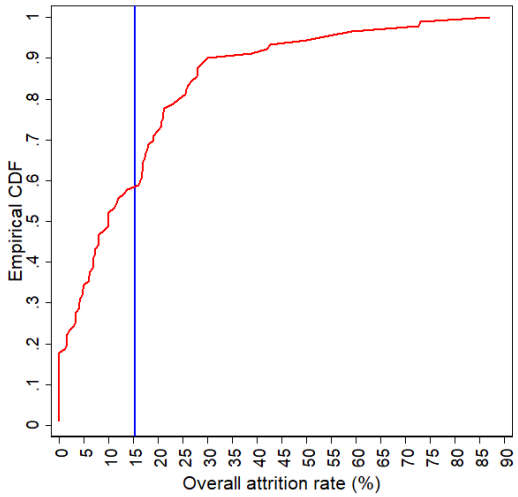
[16]In our review, we included the top five journals in economics, as well as four highly regarded applied economics journals that commonly publish field experiments: *American Economic Review*, *American Economic Journal: Applied Economics*, *Econometrica*, *Economic Journal*, *Journal of Development Economics*, *Journal of Political Economy*, *Review of Economics and Statistics*, *Review of Economic Studies*, and *Quarterly Journal of Economics*.

[17]Since some papers report the results of more than one intervention, our analysis includes 91 field experiments reported in those 88 articles.

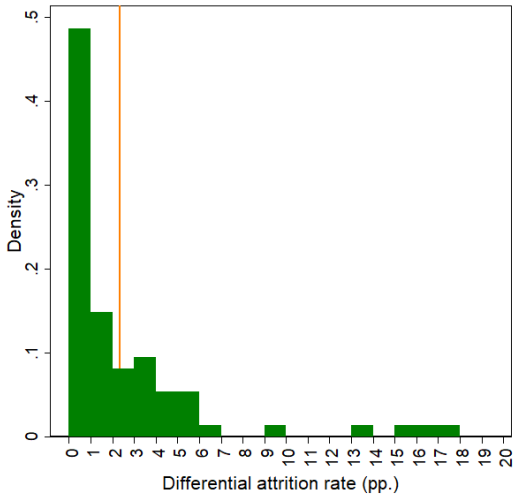
[18]See Section A in the supplementary appendix for additional details on the selection of papers and relevant attrition rates in our review. Section D in the supplementary appendix contains a list of all the papers included in our review.

Figure 1: Attrition Rates Relevant to Main Outcomes in Field Experiments

Panel A. Overall Attrition Rate



Panel B. Differential Attrition Rate



Notes: We report one observation per field experiment. Specifically, the highest attrition rate relevant to an abstract result. The *Overall* rate is the attrition rate for both the treatment and control groups. The *Differential* rate is the absolute value of the difference in attrition rates across treatment and control groups. The blue (yellow) line depicts the average overall (differential) attrition rate in our sample of field experiments. Of the 91 field experiments reviewed, 1 did not report overall attrition rate and 17 did not report differential attrition rate. Hence, the total number of field experiments considered in Panel A and Panel B are 90 and 74, respectively. The 91 field experiments reviewed correspond to 88 papers.

papers. As depicted in Panel A in Figure 1, the average attrition rate in the full evaluation sample is 15%. Furthermore, even though 22% of field experiments have less than 2% attrition overall, the distribution of attrition rates has a long right tail. Specifically, 43% of reviewed field experiments have an attrition rate higher than 15%.[19] Of the experiments that report a differential attrition rate, Panel B in Figure 1 illustrates that a majority have little differential attrition for the abstract results: 66% have a differential rate that is less than 2 percentage points, and only 12% have a differential attrition rate that is greater than 5 percentage points. It is possible, however, that these numbers reflect authors' exclusion of results with higher differential attrition rates than those that were reported or published.[20]

Table 1: Distribution of Field Experiments by Attrition Test

Panel A: Distribution of field experiments by attrition test

Proportion of field experiments that conduct:		Selective attrition test		
		No	*Yes*	*Total*
Differential attrition rate test	*No*	11%	9%	20%
	Yes	29%	51%	80%
	Total	40%	60%	100%

Panel B: Distribution of field experiments by type of selective attrition test

Proportion of field experiments that conduct:	
Joint test: report at least one test using both samples	20%
Simple test: only using sample of respondents	69%
Simple test: only using sample of attritors	4%
Simple test: one using respondents & one using attritors	7%
Total	100%

Notes: The total number of field experiments (papers) considered in Panel A is 75 (73). We exclude 16 field experiments that have attrition rates below 1%. The total number of field experiments (papers) considered in Panel B is 45 (43). The nine field experiments that test using the sample of both respondents and attritors include i) one field experiment that provides a description of the test in a footnote but does not report any table for it, and ii) one field experiment that also conducts a selective attrition test using the sample of respondents only.

We then study how authors test for attrition bias (conditional on attrition rates higher than 1%). First, we note that such tests are widely used in the literature: 89% of the field experiments in our sample conduct at least one test for attrition bias.[21] Next, we determine

[19]The full evaluation sample includes the treatment and control groups. For simplicity, we select one attrition rate relevant to the abstract results for each paper. For many papers, all the abstract results are drawn from one sample and thus there is only one relevant attrition rate. In some papers, however, different abstract results are drawn from different samples. Hence, we focus on the highest attrition rate both for consistency and so that we can understand the extent of attrition that is relevant to abstract results.

[20]These distributions of overall and differential attrition rates inform our simulations in Section 5.

[21]Of the field experiments in our sample, 18% have an attrition rate of less than 1%. Examples of attrition

whether authors conduct a differential attrition rate test or a selective attrition test (Panel A in Table 1). We find that there is no consensus on whether to conduct a differential attrition rate test or a selective attrition test. Differential attrition rate tests are substantially more common than selective attrition tests, however. Of the field experiments that we reviewed, 80% conducted a differential attrition rate test, yet only 60% conducted a selective attrition test. In fact, 29% of the articles that conducted a differential attrition rate do not conduct a selective attrition test. Panel B in Table 1 illustrates the proportion of papers that conduct the simple and joint selective attrition tests. Conditional on having conducted any type of selective attrition test, authors attempt a joint test on only 20% of those field experiments. Instead, authors conduct a simple test of selective attrition on the sample of respondents in most cases (69%).[22]

Table 2: Distribution of Field Experiments by Determinants of Attrition Test

Proportion of field experiments that conduct:	Determinants of attrition test		
	Yes	*No*	*Total*
Differential attrition rate test only	11%	19%	29%
Selective attrition test only	1%	8%	9%
Differential & selective attrition tests	21%	29%	51%
No differential & no selective attrition test	0%	11%	11%
Total	33%	67%	100%

Notes: The total number of field experiments (papers) considered in this table is 75 (73). We exclude 16 field experiments that have attrition rates below 1%.

We also examine whether authors test for differences across respondents and attritors. We find that for approximately one-third of field experiments (33%), the authors conduct a determinants of attrition test (Table 2). It is clear from Table 2 that conducting such a test does not have a one-to-one relationship with either conducting a differential attrition rate test or conducting a selective attrition test.[23] The use of determinants of attrition tests is suggestive evidence that a large minority of authors have considered whether they are reporting the average treatment effect for respondents or the average treatment effect for the study population.

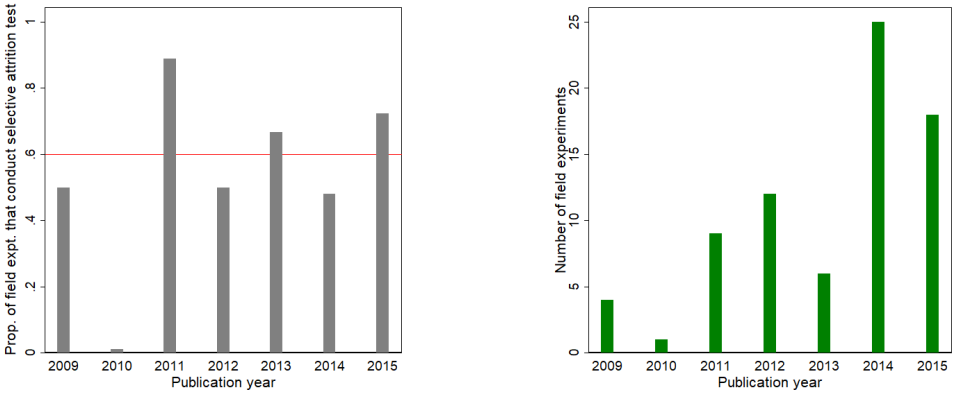
rate lower than 1% include: studies whose unit of observation is the village, studies that use administrative data and are able to match all individuals, and studies that find all individuals at follow-up.

[22]A small minority of field experiments that conduct a selective attrition test conduct two simple tests (7%): one on the respondent subpopulation and one on the attritors. In this case, a multiple testing correction is required to control family-wise error rate, otherwise authors are likely to overreject internal validity based on these tests.

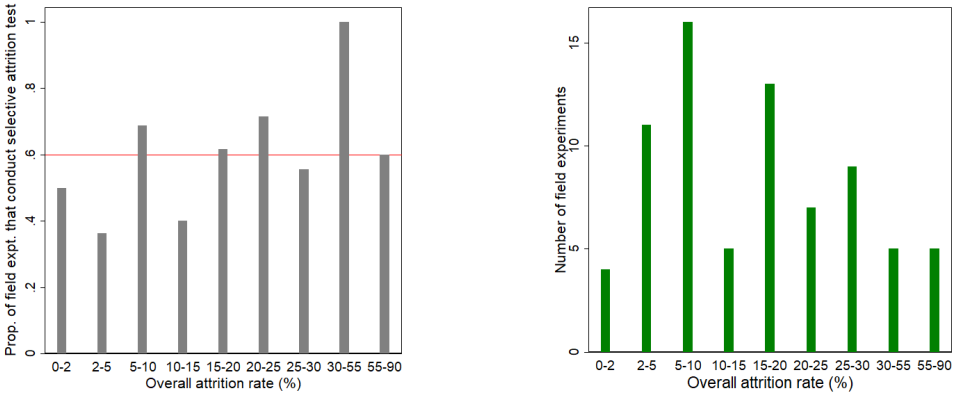
[23]There are twelve field experiments that conduct a differential attrition rate test in the same regression as the determinants of attrition test. This comprises about half the determinants of attrition tests reported and less than a quarter of the differential attrition rate tests. We categorize this strategy as both a differential attrition rate test and a determinants of attrition test since authors typically interpret both the coefficients on treatment and the baseline covariates (see Sections A.1 and A.3)

Figure 2: Attrition Rates, Publication Year & Selective Attrition Tests

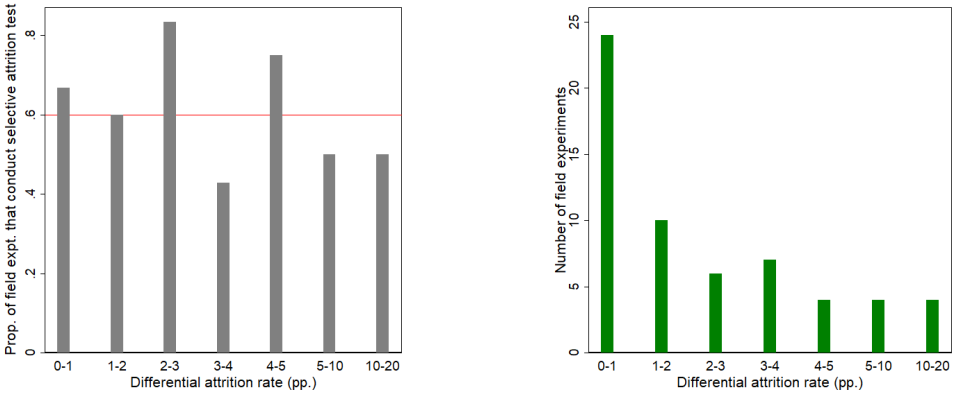
Panel A. Selective Attrition Test & Publication Year



Panel B. Selective Attrition Test & Overall Attrition Rate



Panel C. Selective Attrition Test & Differential Attrition Rate



Notes: The total number of field experiments (papers) included in these figures is 75 (73). We exclude 16 field experiments that have attrition rates below 1%. The gray bars represent the proportion of field experiments in each category displayed on the *x-axis* that conduct a selective attrition test. The red line depicts the proportion of field experiments that conduct the selective attrition test for the sample of 75 experiments. The green bars depict the number of field experiments in each category displayed on the *x-axis*. Of the 75 field experiments with attrition rates above 1%, 16 did not report differential attrition rate. Hence, the number of field experiments considered in Panel C is 59. Each interval in Panels B and C includes (excludes) the upper (lower) bound.

Finally, Figure 2 examines the relationship between conducting a selective attrition test and year of publication as well as attrition rates. With the caveat that our sample is small when split across a number of categories, we do not find any clear trends in the probability of conducting the selective attrition test by year of publication. Furthermore, the overall and differential attrition rates appear to be only weakly correlated with the probability of conducting selective attrition tests.

3 Identifying Treatment Effects in the Presence of Attrition

In this section, we present a formal treatment of attrition in field experiments with baseline outcome data. First, we present identifying assumptions for counterfactual distributions in the presence of non-response and show their sharp testable implications when baseline outcome variables are available. Second, we formally illustrate that equal attrition rates in general are not a necessary condition of internal validity for respondents. Finally, we extend our results to stratified randomized experiments and heterogeneous treatment effects. In this section, we focus on the classical binary treatment setup to simplify presentation. For the extensions of the main results in this section to the multiple treatment case, see Section C in the supplementary appendix.

3.1 Internal Validity and its Testable Restrictions

In a field experiment with baseline outcome data, we observe individuals $i = 1, \ldots, n$ over two time periods, $t = 0, 1$. We will refer to $t = 0$ as the baseline period, and $t = 1$ as the follow-up period. Individuals are randomly assigned in the baseline period to the treatment and control groups. We use D_{it} to denote treatment status for individual i in period t. Hence, the treatment and control groups can be characterized by $D_i \equiv (D_{i0}, D_{i1}) = (0, 1)$ and $D_i = (0, 0)$, respectively. For notational brevity, we let an indicator variable T_i denote the group membership. Specifically, $T_i = 1$ if the individual i belongs to the treatment group and $T_i = 0$ if the individual i belongs to the control group.

For each period $t = 0, 1$, we observe an outcome Y_{it}, which is determined by the treatment status and a vector of time-invariant and time-varying unobservables, U_{it},

$$Y_{it} = \mu_t(D_{it}, U_{it}). \tag{1}$$

Given this structural function, we can define the potential outcomes $Y_{it}(d) = \mu_t(d, U_{it})$ for $d = 0, 1$.[24]

[24]We choose to use the structural notation here since it is more common in the panel literature. This notation also allows us to refer to the unobservables that affect the outcome, which play an important role in

Consider a properly designed and implemented RCT such that by random assignment the treatment and control groups have the same distribution of unobservables. That is, $(U_{i0}, U_{i1}) \perp T_i$, which can be expressed as $(Y_{i0}(0), Y_{i0}(1), Y_{i1}(0), Y_{i1}(1)) \perp T_i$ using the potential outcomes notation. This implies that the control group provides the appropriate counterfactual outcome distribution for the treatment group, i.e. $Y_{i1}(0)|T_i = 1 \overset{d}{=} Y_{i1}|T_i = 0$, where $\overset{d}{=}$ denotes the equality in distribution. In this case, any difference in the outcome distribution between treatment and control groups can be attributed to the treatment. Thus, the average treatment effect (ATE) can be identified as the difference in mean outcomes between the treatment and control group in the follow-up period, i.e.

$$\underbrace{E[Y_{i1}(1) - Y_{i1}(0)]}_{ATE} = E[Y_{i1}|T_i = 1] - E[Y_{i1}|T_i = 0]. \tag{2}$$

We now introduce the possibility of attrition in our setting. We assume that all individuals respond in the baseline period ($t = 0$), but there is possibility of non-response in the follow-up period ($t = 1$). Response status in the follow-up period is determined by the following equation,[25]

$$R_i = \xi(T_i, V_i), \tag{3}$$

where V_i denotes a vector of unobservables that determine response status, and $R_i = 1$ if individual i responds, otherwise it is zero. We can also define potential response for individual i as $R_i(\tau) = \xi(\tau, V_i)$ for $\tau = 0, 1$. Following Lee (2009), random assignment in the context of attrition is given by $(U_{i0}, U_{i1}, V_i) \perp T_i$, which implies $(Y_{i0}(0), Y_{i0}(1), Y_{i1}(0), Y_{i1}(1), R_i(0), R_i(1)) \perp T_i$ using potential outcomes and response notation as in Assumption 1 in Lee (2009). Hence, instead of observing the outcome for all individuals in the treatment and control groups, we can only observe the outcome for respondents in both groups.

Two questions arise in this setting. First, do the control respondents provide an appropriate counterfactual for the treatment respondents; specifically, does $Y_{i1}|T_i = 0, R_i = 1 \overset{d}{=} Y_{i1}(0)|T_i = 1, R_i = 1$? This would imply that we can obtain internally valid estimands for the respondent subpopulation, such as the average treatment effect for respondents (ATE-R), i.e. $E[Y_{i1}(1) - Y_{i1}(0)|R_i = 1]$. Second, do the outcome distributions of treatment and control respondents in period 1 identify the potential outcome distribution of the study population with and without the treatment; specifically, $Y_{i1}|T_i = \tau, R_i = 1 \overset{d}{=} Y_{i1}(\tau)$ for $\tau = 0, 1$? This would imply that we can obtain internally valid estimands for the study population, such as the average treatment

understanding internal validity questions in our problem.

[25]Since non-response is only allowed in the follow-up period, we omit time subscripts from the response equation for notational convenience.

effect (ATE).

Proposition 1 provides sufficient conditions to obtain each of the aforementioned equalities as well as their respective sharp testable restrictions. Part a (b) of the following proposition refers to the case where we can obtain valid estimands for the respondent subpopulation (study population).

Proposition 1. *Assume* $(U_{i0}, U_{i1}, V_i) \perp T_i$.

(a) If $(U_{i0}, U_{i1}) \perp T_i | R_i$ *holds, then*

(i) (Identification) $Y_{i1}|T_i = 0, R_i = 1 \stackrel{d}{=} Y_{i1}(0)|T_i = 1, R_i = 1$

(ii) (Sharp Testable Restriction) $Y_{i0}|T_i = 0, R_i = r \stackrel{d}{=} Y_{i0}|T_i = 1, R_i = r$ *for* $r = 0, 1$.

(b) If $(U_{i0}, U_{i1}) \perp R_i | T_i$ *holds, then*

(i) (Identification) $Y_{i1}|T_i = \tau, R_i = 1 \stackrel{d}{=} Y_{i1}(\tau)$ *for* $\tau = 0, 1$.

(ii) (Sharp Testable Restriction) $Y_{i0}|T_i = \tau, R_i = r \stackrel{d}{=} Y_{i0}$ *for* $\tau = 0, 1$, $r = 0, 1$.

The proof of the proposition is given in Section B of this paper. The assumption in (a) is random assignment conditional on response status.[26] The equality in (a.i) implies the identification of the ATE-R, i.e. $E[Y_{i1}|T_i = 1, R_i = 1] - E[Y_{i1}|T_i = 0, R_i = 1] = E[Y_{i1}(1) - Y_{i1}(0)|R_i = 1]$, as well as the identification of quantile and other distributional treatment effects for the respondent subpopulation. We will refer to this case as *internal validity for the respondent subpopulation*. The restriction in (a.ii) implies that the appropriate test of the implication of internal validity for the respondent subpopulation is a *joint test* of the equality of the baseline outcome distribution between treatment and control respondents as well as treatment and control attritors.

The assumption in (b) implies "missing-at-random" (MAR) as defined in Manski (2005).[27] Together with random assignment, it implies that both treatment and response status are jointly independent of the unobservables in the outcome equation. We will refer to this case as *internal validity for the study population*. The equality in (b.i) implies identification of the ATE as the difference in mean outcomes between treatment and control respondents, i.e. $E[Y_{i1}|T_i = 1, R_i = 1] - E[Y_{i1}|T_i = 0, R_i = 1] = E[Y_{i1}(1) - Y_{i1}(0)]$, as well as the identification of quantile and other distributional treatment effects for the study population.[28] The restriction

[26] We state our assumptions in terms of the joint distribution of (U_{i0}, U_{i1}) to be consistent with the statement of random assignment. Our results also follow if we replace the assumptions on the joint distribution by their counterparts on the marginal distribution of U_{it} for $t = 0, 1$.

[27] In the cross-sectional setup, the MAR assumption is given by $Y_i|T_i, R_i \stackrel{d}{=} Y_i|T_i$. Note that Manski (2005) establishes that the MAR assumption is not testable. We obtain the testable implications here by exploiting the panel structure.

[28] If the linear model holds, the ATE-R equals the ATE, specifically if $Y_{i1} = \alpha + \beta D_{i1} + U_{i1}$, then $\beta = E[Y_{i1}(1) - Y_{i1}(0)|R_i = 1] = E[Y_{i1}(1) - Y_{i1}(0)]$. This relates to Proposition 1 in Angrist (1997), which establishes that random assignment conditional on response status and missing at random are equivalent in the context of the linear model if treatment is randomly assigned.

in (b.ii) is the testable implication of random assignment together with MAR. The resulting null hypothesis in this case is the equality of the baseline outcome distribution regardless of both treatment and response status.

A subtle, yet important implication of the above proposition is that testing internal validity in the presence of attrition inherits the well-known problem in testing identifying assumptions. Specifically, we can only test an implication of an identifying assumption in general.[29] Hence, as illustrated in the following example, while rejection of such a test allows us to refute the identifying assumption in question, it is possible not to reject the test even when identification fails.

Example. *Suppose that there are two unobservables that enter the outcome equation,* $U_{it} = (U_{it}^1, U_{it}^2)'$ *for* $t = 0, 1$*, such that* $(U_{i0}^1, U_{i1}^1) \perp T_i | R_i$ *whereas* $(U_{i0}^2, U_{i1}^2) \not\perp T_i | R_i$*. Let the outcome at baseline be a trivial function of* U_{i0}^2*, whereas the outcome in the follow-up period is a non-trivial function of both* U_{i0}^1 *and* U_{i0}^2*, e.g.*

$$
\begin{aligned}
Y_{i0} &= U_{i0}^1 \\
Y_{i1} &= U_{i1}^1 + U_{i1}^2 + T_i(\beta_1 U_{i1}^1 + \beta_2 U_{i1}^2)
\end{aligned}
$$

As a result, even though $Y_{i0}|T_i = 1, R_i \overset{d}{=} Y_{i0}|T_i = 0, R_i$ *holds,* $Y_{i1}(0)|T_i = 1, R_i = 1 \overset{d}{\neq} Y_{i1}|T_i = 0, R_i = 1$*. In other words, the control respondents do not provide a valid counterfactual for the treatment respondents in the follow-up period despite the identity of the baseline outcome distribution for treatment and control groups conditional on response status. We can illustrate this by looking at the average treatment effect for the treatment respondents,*

$$
\begin{aligned}
&E[Y_{i1}(1) - Y_{i1}(0) | T_i = 1, R_i = 1] \\
&= \underbrace{E[U_{i1}^1 + U_{i1}^2 + \beta_1 U_{i1}^1 + \beta_2 U_{i1}^2 | T_i = 1, R_i = 1]}_{E[Y_{i1}|T_i=1,R_i=1]} - \underbrace{E[U_{i1}^1 + U_{i1}^2 | T_i = 1, R_i = 1]}_{\neq E[Y_{i1}|T_i=0,R_i=1]}.
\end{aligned}
$$

Hence, $E[Y_{i1}|T_i = 1, R_i = 1] - E[Y_{i1}|T_i = 0, R_i = 1] \neq \beta_1 E[U_{i1}^1|T_i = 1, R_i = 1] + \beta_2 E[U_{i1}^2|T_i = 1, R_i = 1]$*, i.e. the difference in mean outcomes between treatment and control respondents does not identify an average treatment effect for the treatment respondents.*[30]

[29] If we impose time stationarity on the structural function and unobservable distribution (Chernozhukov et al., 2013), specifically $\mu_0 = \mu_1$ and $U_{i0}|T_i, R_i \overset{d}{=} U_{i1}|T_i, R_i$, then the testable restriction in (a.ii) holds iff identification (a.i) holds.

[30] We could however have a case in which the control respondents provide a valid conuterfactual for the treatment respondents even though the treatment effect for individual i depends on an unobservable that is not independent of treatment conditional on response, i.e. U_{it}^2. Specifically, let $Y_{it} = U_{it}^1 + T_i(\beta_1 U_{it}^1 + \beta_2 U_{it}^2)$ and consider the identification of the average treatment effect, $E[Y_{i1}(1) - Y_{i0}(0)|T_i = 1, R_i = 1] = E[U_{i1}^1 + \beta_1 U_{i1}^1 + \beta_2 U_{i1}^2 | T_i = 1, R_i = 1] - E[U_{i1}^1 | T_i = 1, R_i = 1] = E[Y_{i1}|T_i = 1, R_i = 1] - E[Y_{i1}|T_i = 0, R_i = 1]$, since $E[U_{i1}^1|T_i = 1, R_i = 1] = E[U_{i1}^1|T_i = 0, R_i = 1]$. Note however that in this case what we identify is no

The above example illustrates why we cannot test identification "directly", since it would require us to observe the counterfactual of the treatment respondents. This brings us to the importance of using sharp testable restrictions. Since we can only test an identifying assumption by implication, it is crucial to ensure that we test the strongest possible implication of the identifying assumption in question on our data.

An important question that arises in empirical practice is whether covariates should be used in testing the identifying assumptions in question. Suppose that the researcher has *a priori* information that establishes that there are covariates determined by the same unobservables as the outcome Y_{it}, i.e. $W_{it} = \nu_t(U_{it})$ for $t = 0, 1$. Then, both sharp testable restrictions of the identifying assumptions in Proposition 1 would be on the joint distribution of $Z_{i0} = (Y_{i0}, W'_{i0})'$ and not solely on the marginal distribution of Y_{i0}. However, if this *a priori* information is false and W_{it} also depends on unobservables that affect response, V_i, then testing restrictions on the joint distribution of Z_{i0} may lead to false rejection of the identifying assumption in question.

Proposition 1(a) provides two main insights on the selective attrition tests conducted in the field experiment literature. First, the selective attrition tests, which are based on the equality of the *means* of the baseline outcome variable, are an implication of the joint *distributional* restriction in (a.ii). Thus, they test an implication of internal validity for the respondent subpopulation and not internal validity for the study population. Second, the *joint* test of selective attrition, as opposed to a simple test using respondents or attritors, is based on the sharp testable restriction of $(U_{i0}, U_{i1}) \perp T_i | R_i$, i.e. random assignment conditional on response status.[31]

3.2 Differential Attrition Rates and Internal Validity for Respondents

When attrition rates across treatment and control groups are not equal, specifically $P(R_i = 0|T_i = 1) \neq P(R_i = 0|T_i = 0)$, we call this a differential attrition rate as in Section 2. Our review of the field experiment literature indicates that many researchers put a lot of weight on the differential attrition rate test when testing for attrition bias. Thus, in this section we examine the relationship between internal validity and equal attrition rates. We focus on internal validity for respondents since this is the first-order problem.

In order to understand the role of differential attrition rates in the context of internal validity

longer internally valid for the entire respondent subpopulation, but for the smaller subpopulation of treatment respondents.

[31] If internal validity for respondents is of interest, a natural question is whether one should simply test the implication of $(U_{i0}, U_{i1}) \perp T_i | R_i = 1$. This is of interest if it is plausible that $(U_{i0}, U_{i1}) \perp T_i | R_i = 1$ holds while $(U_{i0}, U_{i1}) \perp T_i | R_i = 0$ is violated. Using the subgroups defined by potential response status in Figure 3, we note that a primitive condition for such case to hold is $(U_{i0}, U_{i1}) | R_i(0), R_i(1) \stackrel{d}{=} (U_{i0}, U_{i1}) | \max\{R_i(0), R_i(1)\}$. This primitive condition is a correlated random effects assumption that implies that the unobservable distribution (U_{i0}, U_{i1}) is the same for always responders, treatment-only and control-only responders, but different for the never-responders.

for respondents, we use potential response to characterize different subpopulations that will differ in terms of their distribution of unobservables. Here we adapt the terminology of never-takers, always-takers, compliers and defiers from the LATE literature (Imbens and Angrist, 1994; Angrist et al., 1996) to our setting. Specifically, never-responders ($(R_i(0), R_i(1)) = (0,0)$), always-responders ($(R_i(0), R_i(1)) = (1,1)$), treatment-only responders ($(R_i(0), R_i(1)) = (0,1)$), and control-only responders ($(R_i(0), R_i(1)) = (1,0)$). As shown in Figure 3, the treatment and control respondents and attritors are composed of different subpopulations $(R_i(0), R_i(1))$.

Figure 3: Respondent and Attritor Subgroups

	Control ($T_i = 0$)	Treatment ($T_i = 1$)
Attritors ($R_i = 0$)	$(R_i(0), R_i(1)) = (0,1)$ $(R_i(0), R_i(1)) = (0,0)$	$(R_i(0), R_i(1)) = (1,0)$ $(R_i(0), R_i(1)) = (0,0)$
Respondents ($R_i = 1$)	$(R_i(0), R_i(1)) = (1,0)$ $(R_i(0), R_i(1)) = (1,1)$	$(R_i(0), R_i(1)) = (0,1)$ $(R_i(0), R_i(1)) = (1,1)$

We can now examine the difference in attrition rates and what it measures in terms of the proportions of the aforementioned subpopulations, which we define as:

$$\begin{aligned} p_{00} &\equiv P((R_i(0), R_i(1)) = (0,0)), \quad p_{01} \equiv P((R_i(0), R_i(1)) = (0,1)), \\ p_{10} &\equiv P((R_i(0), R_i(1)) = (1,0)), \quad p_{11} \equiv P((R_i(0), R_i(1)) = (1,1)). \end{aligned} \tag{4}$$

Now note that by random assignment, $(R_i(0), R_i(1)) \perp T_i$, hence the attrition rates in the treatment and control groups are given by

$$P(R_i = 0|T_i = 0) = p_{00} + p_{01}, \quad P(R_i = 0|T_i = 1) = p_{00} + p_{10}. \tag{5}$$

The difference in attrition rates across groups measures the difference between the proportion of treatment-only and control-only responders, i.e. $P(R_i = 0|T_i = 0) - P(R_i = 0|T_i = 1) = p_{01} - p_{10}$. Thus, in general, equal attrition rates occur if $p_{01} = p_{10}$.

Since internal validity for respondents holds if the distribution of unobservables that affect outcome is the same for treatment and control respondents, as we point out in Proposition 1, we now express the distribution of these unobservables, (U_{i0}, U_{i1}), for treatment and control respondents as a mixture of the distributions conditional on $(R_i(0), R_i(1))$ under random assignment,

$$\begin{aligned} F_{U_{i0},U_{i1}|T_i=1,R_i=1} &= \frac{p_{01} F_{U_{i0},U_{i1}|(R_i(0),R_i(1))=(0,1)} + p_{11} F_{U_{i0},U_{i1}|(R_i(0),R_i(1))=(1,1)}}{P(R_i = 1|T_i = 1)}, \\ F_{U_{i0},U_{i1}|T_i=0,R_i=1} &= \frac{p_{10} F_{U_{i0},U_{i1}|(R_i(0),R_i(1))=(1,0)} + p_{11} F_{U_{i0},U_{i1}|(R_i(0),R_i(1))=(1,1)}}{P(R_i = 1|T_i = 0)}. \end{aligned}$$

The two distributions on the left hand side of the above equations may agree in three different cases. First, if the distributions of treatment-only, control-only and always-responders all agree. That is, if $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$, or equivalently $(U_{i0}, U_{i1}) \perp V_i$. Otherwise, the different subgroups $(R_i(0), R_i(1))$ may not only differ in terms of the distribution of V_i but also in terms of the distribution of (U_{i0}, U_{i1}). Second, if there were no treatment-only or control-only responders, i.e. $p_{10} = p_{01} = 0$, which is a special case of monotonicity as discussed in Lee (2009). Third, if $p_{10} = p_{01}$ and the distribution of unobservables that affect the outcome for the treatment-only and the control-only responders are identical. The equality of distribution for treatment-only and control-only responders is implied by an exchangeability restriction (Altonji and Matzkin, 2005) given below. These three sets of assumptions imply internal validity for respondents as formally stated in the following proposition.

Proposition 2. *Suppose, in addition to* $(U_{i0}, U_{i1}, V_i) \perp T_i$*, one of the following is true,*

(i) $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$ *(Unobservables in* $Y \perp$ *Potential Response)*

(ii) $R_i(0) \leqslant R_i(1)$ *(wlog),* *(Monotonicity)*
& $P(R_i = 0|T_i) = P(R_i = 0)$ *(Equal Attrition Rates)*

(iii) $(U_{i0}, U_{i1})|R_i(0), R_i(1) \overset{d}{=} (U_{i0}, U_{i1})|R_i(0) + R_i(1)$ *(Exchangeability)*
& $P(R_i = 0|T_i) = P(R_i = 0)$ *(Equal Attrition Rates)*

then $(U_{i0}, U_{i1}) \perp T_i|R_i$.

The proof of the proposition is given in Appendix B. Note that in (i) there are no restrictions on the attrition rates. Furthermore, the assumption implies MAR, specifically $Y_{it}|T_i, R_i \overset{d}{=} Y_{it}|T_i$, which together with random assignment not only implies internal validity for respondents, but also for the study population. On the other hand, in (ii) and (iii), equal attrition rates together with another assumption are sufficient for internal validity for respondents. In (ii), where equal attrtion rates and monotonicity coincide, the respondent subpopulation is solely composed of always-responders ($(R_i(0), R_i(1)) = (1, 1)$).[32] The exchangeability restriction in (iii) merits some discussion. First, it is weaker than monotonicity, since it allows for both treatment-only and control-only responders, but it assumes that these "inconsistent" types have the same distribution of (U_{i0}, U_{i1}). While strong in general, this assumption may be more realistic in experiments with two treatments. If coupled with equal attrition rates, exchangeability implies internal validity for respondents.

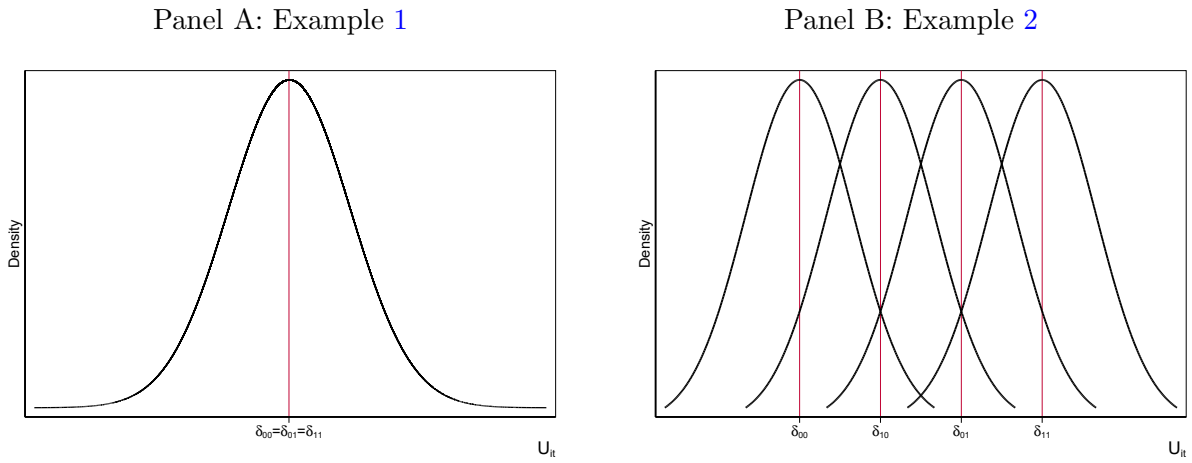
The above discussion and proposition illustrate that equal attrition rates without further assumptions do not imply internal validity for respondents. To illustrate this point further, we present two examples. The first shows that internal validity can coincide with differential

[32]The bounds proposed in Lee (2009) are for the average treatment effect of always-responders.

attrition rates, whereas the second shows that equal attrition rates can coincide with a violation of internal validity.

Example 1. *(Internal Validity & Differential Attrition Rates)*
Assume that the potential response satisfies monotonicity, i.e. $p_{10} = 0$, *and* $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$.[33] *Furthermore, there is a group of individuals for whom it is too costly to respond if they are in the control group. Hence, this group will only respond if assigned the treatment* ($p_{01} > 0$). *By the above proposition, under random assignment,* $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1)) \Rightarrow (U_{i0}, U_{i1})|T_i, R_i \overset{d}{=} (U_{i0}, U_{i1})$. *Panel A of Figure 4 illustrates the resulting distribution of* U_{it}. *However,* $P(R_i = 0|T_i = 1) = p_{00}$, *and* $P(R_i = 0|T_i = 0) = p_{00} + p_{01}$. *Hence, even though we have internal validity not only for the respondents, but also for the study population, we have differential attrition rates across treatment and control groups.*

Figure 4: Distribution of U_{it} for Subpopulations in terms of Potential Responses



Notes: The above figure illustrates the distribution of U_{it} for the different subpopulations for Examples 1 and 2, where we assume $U_{it}|(R_i(0), R_i(1)) = (r_0, r_1) \overset{i.i.d.}{\sim} N(\delta_{r_0 r_1}, 1)$ for all $r_0, r_1 \in \{0, 1\}^2$ for $t = 0, 1$. Panel A represents Example 1 where we assume $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$, hence $\delta_{00} = \delta_{01} = \delta_{11}$. Panel B represents Example 2 where $\delta_{r_0 r_1}$ is unrestricted for $(r_0, r_1) \in \{0, 1\}^2$.

Example 2. *(Equal Attrition Rates & Violation of Internal Validity)*
Assume that potential response violates monotonicity, such that there are treatment-only and control-only responders,[34] *but their proportions are equal* ($p_{10} = p_{01} > 0$), *which yields equal*

[33] For instance, suppose reciprocity is the unobservable that determines response, V_i. If the unobservables that affect outcome are independent of reciprocity, then MAR would hold.

[34] Violations of monotonicity are especially plausible in settings where we have two treatments. For the classical treatment-control case, a nice example of a violation of monotonicity of response is given in Glennerster and Takavarasha (2013). Suppose the treatment is a remedial program for public schools targeted toward students that have identified deficiencies in mathematics. Response in this setting is determined by whether students remain in the public school, which depends on their treatment status and initial mathematical ability, V_i. On one

attrition rates across treatment and control groups.[35] *If* $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$*, then the different subpopulations will have different distributions of unobservables, as illustrated in Panel B of Figure 4. As a result, the distribution of* (U_{i0}, U_{i1}) *for treatment and control respondents defined in* (19)-(20) *will be different and hence internal validity for respondents is violated.*

We recognize that under conditions that may be realistic in certain settings, differential attrition rates indicate a violation of internal validity for respondents. The above examples point out however that the emphasis on the differential attrition rate test as a test of internal validity for respondents can be misleading. While *Example 1* shows that differential attrition rates can coincide with internal validity, *Example 2* illustrates that internal validity can be violated even though we have equal attrition rates. In Section 5, we design simulation experiments that mimic the above examples to illustrate these points numerically.

A further limitation of the focus on differential attrition rates is that we cannot learn from them whether we have internal validity for the study population, even in cases where the differential attrition rate test is a valid test of internal validity for respondents. For instance, consider the case in which monotonicity holds and the attrition rates are equal across groups. We then have internal validity for respondents, which are composed solely of always-responders as pointed out above. However, if the researcher is interested in the treatment effect for the study population, she/he would have to test whether the counterfactual distribution provided by the always-responders is "representative" of the study population. To do so, one would have to test the restriction in Proposition 1(b.ii).

3.3 Stratified Randomization and Heterogeneous Treatment Effects

In many field experiments, randomization is performed within strata or blocks to identify heterogeneous treatment effects, more formally defined as conditional average treatment effects (CATE). In this section, we extend Proposition 1 to this case. We examine the question of identifying the counterfactual distribution for the respondent subpopulation as well as for the study population. The results in this section also apply to completely randomized experiments when heterogeneous treatment effects are of interest.

In the following, let S_i denote the stratum of individual i which has support $\mathcal{S}$, where $|\mathcal{S}| < \infty$. To exclude trivial strata, we assume that $P(S_i = s) > 0$ for all $s \in \mathcal{S}$ throughout the

side, low-achieving students would drop out of school if they are assigned to the control group, but would remain in school if assigned the treatment. On the other side, parents of high-achieving students in the treatment group may be induced to switch their children to private schools because they are unhappy with the larger class sizes, while in the control group those students would remain in the public school. Furthermore, in the context of the LATE framework, de Chaisemartin (2017) provides several applications where monotonicity is implausible and establishes identification of a local average treatment effect under an alternative assumption.

[35] In the multiple treatment case, equal attrition rates are possible without imposing that any subpopulations have equal proportions. See Section B.1 in the supplementary appendix for a derivation.

paper.

Proposition 3. *Assume* $(U_{i0}, U_{i1}, V_i) \perp T_i | S_i$.

(a) If $(U_{i0}, U_{i1}) \perp T_i | S_i, R_i$, *then*

(i) (Identification) $Y_{i1}|T_i = 0, S_i = s, R_i = 1 \stackrel{d}{=} Y_{i1}(0)|T_i = 1, S_i = s, R_i = 1$, *for* $s \in \mathcal{S}$.

(ii) (Sharp Testable Restriction) $Y_{i0}|T_i = 0, S_i = s, R_i = r \stackrel{d}{=} Y_{i0}|T_i = 1, S_i = s, R_i = r$ *for* $r = 0, 1$, $s \in \mathcal{S}$.

(b) If $(U_{i0}, U_{i1}) \perp R_i | T_i, S_i$, *then*

(i) (Identification) $Y_{i1}|T_i = \tau, S_i = s, R_i = 1 \stackrel{d}{=} Y_{i1}(\tau)|S_i = s$, *for* $\tau = 0, 1$, $s \in \mathcal{S}$.

(ii) (Sharp Testable Restriction) $Y_{i0}|T_i = \tau, S_i = s, R_i = r \stackrel{d}{=} Y_{i0}(0)|S_i = s$ *for* $\tau = 0, 1$, $r = 0, 1$, $s \in \mathcal{S}$.

The equality in (a.i) implies that we can identify the average treatment effect conditional on S for respondents as the difference in mean outcomes between treatment and control respondents in each stratum,

$$\begin{aligned} &E\big[Y_{i1}(1) - Y_{i1}(0) | T_i = 1, S_i = s, R_i = 1\big] \\ =&E\big[Y_{i1} | T_i = 1, S_i = s, R_i = 1\big] - E\big[Y_{i1} | T_i = 0, S_i = s, R_i = 1\big], \text{(CATE-R)} \end{aligned} \tag{6}$$

Alternatively, the ATE-R can then be identified by averaging over S_i, i.e. $\sum_{s \in \mathcal{S}} P(S_i = s | R_i = 1) \left(E\big[Y_{i1} | T_i = 1, S_i = s, R_i = 1\big] - E\big[Y_{i1} | T_i = 0, S_i = s, R_i = 1\big]\right)$. The testable restriction in (a.ii) is the identity of the distribution of baseline outcome for treatment and control groups conditional on response status *and* stratum. In other words, the equality of the outcome distribution for treatment and control respondents (as well as for treatment and control attritors) conditional on stratrum is the sharp testable restriction of internal validity for respondents in the case of block randomization. The results in part (b) of the proposition refer to the internal validity for the study population in the context of block randomization. Thus, they are also conditional versions of the results in Proposition 1(b).

The testable restrictions given in the above proposition are also appropriate in completely randomized experiments where heterogeneous treatment effects within strata S_i are of interest. Suppose that in a completely randomized experiment, we are interested in heterogeneous treatment effects within strata S_i. Since complete randomization in this setting is given by $(U_{i0}, U_{i1}, V_i, S_i) \perp T_i$, it implies that treatment is randomly assigned conditional on stratum, i.e. $(U_{i0}, U_{i1}, V_i) \perp T_i | S_i$. In this case, the appropriate testable restriction of internal validity for the respondent subpopulation is that treatment is randomly assigned conditional on stratum

and response status, whereas the testable restriction of internal validity for the study population is that the treatment-response status is randomly assigned conditional on stratum.

4 Randomization Tests of Internal Validity

In this section, we present the distributional test statistics for the restrictions in Propositions 1 and 3 and the randomization procedures to obtain their p-values. We first briefly review randomization procedures in general. Then, we present the statistics for completely and stratified randomized trials. In each case, we explain how to implement the distributional tests of the implications for both internal validity for the respondent subpopulation and internal validity for the study population.

The testable restrictions for internal validity consist of multiple distributional equalities that must hold jointly. Thus, in order to obtain the p-values for joint distributional statistics, we propose a "subgroup"-randomization procedure.[36] This procedure approximates the exact p-values under the cross-sectional i.i.d. assumption when the outcome distribution is absolutely continuous. It can also be adapted to accommodate possibly discrete distributions by applying the procedure in Dufour (2006).

We now present a brief overview of randomization procedures.[37] Given a dataset $\mathbf{Z}$ and a statistic $T_n = T(\mathbf{Z})$ that tests a null hypothesis H_0, we use the following procedure to provide a stochastic approximation of the exact p-value for the test statistic T_n exploiting invariant transformations $g \in \mathcal{G}_0$ (Lehmann and Romano, 2005, Chapter 15.2). Specifically, the transformations $g \in \mathcal{G}_0$ satisfy $\mathbf{Z} \overset{d}{=} g(\mathbf{Z})$ under H_0 only.

Procedure 1. *(Randomization)*

1. *For g_b, which is i.i.d. Uniform($\mathcal{G}_0$), compute $\hat{T}_n(g_b) = T(g_b(\boldsymbol{Z}))$,*
2. *Repeat Step 1 for $b = 1, \ldots, B$ times,*
3. *Compute the p-value, $\hat{p}_{n,B} = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} 1\{\hat{T}_n(g_b) \geqslant T_n\}\right)$.*

A test that rejects when $\hat{p}_{n,B} \leqslant \alpha$ is level α for any B (Lehmann and Romano, 2005, Chapter 15.2). In our application, the invariant transformations in $\mathcal{G}_0$ consist of permutations of individuals across certain subgroups in our data set. The subgroups are defined by the combination of response and treatment in the case of completely randomized trials, and all the combinations of response, treatment, and stratum in the case of trials that are randomized within strata.

[36] The randomization procedure we propose can be used to obtain p-values for t-tests and other statistics that test the equality of two distributions. However, in this section we only focus on distributional statistics.

[37] See Lehmann and Romano (2005); Canay et al. (2017) for a more detailed review.

4.1 Completely Randomized Trials

The testable restriction of internal validity for the respondent subpopulation, stated in Proposition 1(a.ii), implies that the distribution of baseline outcome is identical for treatment and control respondents as well as treatment and control attritors. Thus, the joint hypothesis is given by,

$$H_0^1 : F_{Y_{i0}|T_i=0,R_i=r} = F_{Y_{i0}|T_i=1,R_i=r} \text{ for } r = 0, 1. \tag{7}$$

The general form of the distributional statistic for *each* of the equalities in the null hypothesis above is,

$$T_{n,r}^1 = \left\| \sqrt{n} \left(F_{n,Y_{i0}|T_i=0,R_i=r} - F_{n,Y_{i0}|T_i=1,R_i=r} \right) \right\| \quad \text{for } r = 0, 1,$$

where for a random variable X_i, F_{n,X_i} denotes the empirical cdf, i.e. the sample analogue of F_{X_i}, and $\|.\|$ denotes some non-random or random norm. Different choices of the norm give rise to different statistics. We use the KS and CM statistics in the simulations since they are the most widely known and used. The former is obtained by using the L^∞ norm over the sample points, i.e. $\|f\|_{n,\infty} = \max_i |f(y_i)|$, whereas the latter is obtained by using an L^2-type norm, i.e. $\|f\|_{n,2} = \sum_{i=1}^n f(y_i)^2/n$. Then, in order to test the *joint* hypothesis in (7), we use the two following joint statistics that aggregate over $T_{n,r}^1$ for $r = 0, 1$ as follows[38]

$$T_{n,m}^1 = \max\{T_{n,0}, T_{n,1}\}$$
$$T_{n,p}^1 = p_{n,0} T_{n,0} + p_{n,1} T_{n,1}, \quad \text{where } p_{n,r} = \sum_{i=1}^n 1\{R_i = r\}/n \text{ for } r = 0, 1.$$

To apply the randomization procedure, let $\mathcal{G}$ denote the set of all possible permutations of n elements and $\mathcal{G}_0^1$ be the subset of $\mathcal{G}$ that only includes permutations within respondent and attritor subgroups. Specifically, for $g \in \mathcal{G}_0^1$, $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, R_{g(i)}) : R_{g(i)} = R_i, 1 \leqslant i \leqslant n\}$. Under H_0^1 and the i.i.d. assumption across individuals, $\mathbf{Z} \overset{d}{=} g(\mathbf{Z})$ for $g \in \mathcal{G}_0^1$. Hence, we can obtain p-values for $T_{n,m}^1$ and $T_{n,p}^1$ under H_0^1 by applying Procedure 1 using the set of permutations $\mathcal{G}_0^1$.

We now consider testing the restriction of internal validity for the study population stated in Proposition 1(b.ii). This restriction implies that the distribution of the baseline outcome variable is identically distributed across all four subgroups defined by treatment and response status. Let $(T_i, R_i) = (\tau, r)$, where $(\tau, r) \in \mathcal{T} \times \mathcal{R} = \{(0,0), (0,1), (1,0), (1,1)\}$ and (τ_j, r_j)

[38] There are other possible choices for statistics of the joint hypothesis. The aforementioned statistics are standard choices (Imbens and Rubin, 2015).

denote the j^{th} element of $\mathcal{T} \times \mathcal{R}$. Then, the joint hypothesis is given wlog by

$$H_0^2 : \; F_{Y_{i0}|T_i=\tau_j,R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1},R_i=r_{j+1}} \text{ for } j = 1, \ldots, |\mathcal{T} \times \mathcal{R}| - 1. \tag{8}$$

In this case, the two statistics that we propose to test the *joint* hypothesis are:

$$\begin{aligned} T_{n,m}^2 &= \max_{j=1,\ldots,|\mathcal{T}\times\mathcal{R}|-1} \left\| \sqrt{n} \left(F_{n,Y_{i0}|T_i=\tau_j,R_i=r_j} - F_{n,Y_{i0}|T_i=\tau_{j+1},R_i=r_{j+1}} \right) \right\|, \\ T_{n,p}^2 &= \sum_{j=1}^{|\mathcal{T}\times\mathcal{R}|-1} w_j \left\| \sqrt{n} \left(F_{n,Y_{i0}|T_i=\tau_j,R_i=r_j} - F_{n,Y_{i0}|T_i=\tau_{j+1},R_i=r_{j+1}} \right) \right\| \end{aligned}$$

for some fixed or data-dependent non-negative weights w_j for $j = 1, \ldots, |\mathcal{T} \times \mathcal{R}| - 1$.

Under H_0^2 and the cross-sectional i.i.d. assumption, any random permutation of individuals across the four treatment-response subgroups will yield the same joint distribution of the data. Specifically, for $g \in \mathcal{G}_0^2$, $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, R_{g(i)}) : 1 \leqslant i \leqslant n\}$. We can hence apply Procedure 1 using $\mathcal{G}_0^2$ to obtain p-values for the statistic $T_{n,m}^2$ or $T_{n,p}^2$ under H_0^2.

4.2 Stratified Randomized Trials

As pointed out in Section 3.3, the testable restrictions in the case of stratified or block randomized trials (Proposition 3) are conditional versions of those in the case of completely randomized trials (Proposition 1). Thus, in what follows we lay out the conditional versions of the null hypotheses, the distributional statistics, and the invariant transformations presented in Section 4.1.

We first consider the restriction in Proposition 3(a.ii). The resulting null hypothesis is given by

$$H_0^{1,\mathcal{S}} : F_{Y_{i0}|T_i=0,S_i=s,R_i=r} = F_{Y_{i0}|T_i=1,S_i=s,R_i=r} \text{ for } r = 0, 1, \; s \in \mathcal{S}. \tag{9}$$

To obtain the test statistics for the joint hypothesis $H_0^{1,\mathcal{S}}$, we first construct test statistics for a given $s \in \mathcal{S}$,

$$\begin{aligned} T_{n,m,s}^{1,\mathcal{S}} &= \max_{r=0,1} \left\| \sqrt{n} \left(F_{n,Y_{i0}|T_i=0,S_i=s,R_i=r} - F_{n,Y_{i0}|T_i=1,S_i=s,R_i=r} \right) \right\|, \\ T_{n,p,s}^{1,\mathcal{S}} &= \sum_{r=0,1} p_n^{r|s} \left\| \sqrt{n} \left(F_{n,Y_{i0}|T_i=0,S_i=s,R_i=r} - F_{n,Y_{i0}|T_i=1,S_i=s,R_i=r} \right) \right\|, \end{aligned}$$

where $p_n^{r|s} = \sum_{i=1}^n 1\{R_i = r, S_i = s\} / \sum_{i=1}^n 1\{S_i = s\}$. We then aggregate over each of those

statistics to get

$$T^{1,\mathcal{S}}_{n,m} = \max_{s\in\mathcal{S}} T^{1,\mathcal{S}}_{n,m,s},$$
$$T^{1,\mathcal{S}}_{n,p} = \sum_{s\in\mathcal{S}} p^s_n T^{1,\mathcal{S}}_{n,p,s}, \text{ where } p^s_n = \sum_{i=1}^{n} 1\{S_i = s\}/n \text{ for } s \in \mathcal{S}.$$

In this case, the invariant transformations under $H_0^{1,\mathcal{S}}$ are the ones where n elements are permuted within response-strata subgroups. Formally, for $g \in \mathcal{G}_0^{1,\mathcal{S}}$, $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, S_{g(i)}, R_{g(i)}) : S_{g(i)} = S_i, R_{g(i)} = R_i, 1 \leqslant i \leqslant n\}$, where $\mathbf{Z} = \{(Y_{i0}, T_i, S_i, R_i) : 1 \leqslant i \leqslant n\}$. Under $H_0^{1,\mathcal{S}}$ and the cross-sectional i.i.d. assumption within strata, $\mathbf{Z} \overset{d}{=} g(\mathbf{Z})$ for $g \in \mathcal{G}_0^{1,\mathcal{S}}$. Hence, using $\mathcal{G}_0^{1,S}$, we can obtain p-values for $T^{1,\mathcal{S}}_{n,m}$ and $T^{1,\mathcal{S}}_{n,p}$ under $H_0^{1,\mathcal{S}}$.

We now consider testing the restriction in Proposition 3(b.ii). The resulting null hypothesis is given wlog by the following

$$H_0^{2,\mathcal{S}} : F_{Y_{i0}|T_i=\tau_j,S_i=s,R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1},S_i=s,R_i=r_{j+1}} \text{ for } j = 1,\ldots,|\mathcal{T}\times\mathcal{R}| - 1,\ s \in \mathcal{S}. \quad (10)$$

To obtain the test statistics for the joint hypothesis $H_0^{2,\mathcal{S}}$, we first construct test statistics for a given $s \in \mathcal{S}$,

$$T^{2,\mathcal{S}}_{n,m,s} = \max_{j=1,\ldots,|\mathcal{T}\times\mathcal{R}|-1} \left\| \sqrt{n}\left(F_{n,Y_{i0}|T_i=\tau_j,S_i=s,R_i=r_j} - F_{n,Y_{i0}|T_i=\tau_{j+1},S_i=s,R_i=r_{j+1}}\right)\right\|,$$
$$T^{2,\mathcal{S}}_{n,p,s} = \sum_{j=1}^{|\mathcal{T}\times\mathcal{R}|-1} w_{j,s} \left\| \sqrt{n}\left(F_{n,Y_{i0}|T_i=\tau_j,S_i=s,R_i=r_j} - F_{n,Y_{i0}|T_i=\tau_{j+1},S_i=s,R_i=r_{j+1}}\right)\right\|,$$

given fixed or random non-negative weights $w_{j,s}$ for $j = 1,\ldots,|\mathcal{T}\times\mathcal{R}| - 1$ and $s \in \mathcal{S}$. We then aggregate over each of those statistics to get

$$T^{2,\mathcal{S}}_{n,m} = \max_{s\in\mathcal{S}} T^{2,\mathcal{S}}_{n,m,s},$$
$$T^{2,\mathcal{S}}_{n,p} = \sum_{s\in\mathcal{S}} w^s_n T^{2,\mathcal{S}}_{n,p,s},$$

given fixed or random non-negative weights w^s_n for $s \in \mathcal{S}$.

Under the above hypothesis and the cross-sectional i.i.d. assumption within strata, the distribution of the data is invariant to permutations within strata, i.e. for $g \in \mathcal{G}_0^{2,\mathcal{S}}$, $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, S_{g(i)}, R_{g(i)}) : S_{g(i)} = S_i, 1 \leqslant i \leqslant n\}$. Thus, applying Procedure 1 to $T^{2,\mathcal{S}}_{n,m}$ or $T^{2,\mathcal{S}}_{n,p}$ using $\mathcal{G}_0^{2,\mathcal{S}}$ yields p-values for these statistics under $H_0^{2,\mathcal{S}}$.

Before we proceed to the simulation study, it is important to note that, in practice, it may be possible that response problems could lead to violations of internal validity in some strata but not in others. If that is the case, it may be more appropriate to test interval validity for

each stratum separately. Recall that when the goal is to test internal validity for the respondent subpopulation, the stratum-specific hypothesis is $H_0^{1,s} : F_{Y_{i0}|T_i=0,S_i=s,R_i=r} = F_{Y_{i0}|T_i=1,S_i=s,R_i=r}$ for $r = 0, 1$. Hence, for each $s \in \mathcal{S}$, one can use $\mathcal{G}_0^{1,\mathcal{S}}$ in the above procedure to obtain p-values for $T_{n,m,s}^{1,\mathcal{S}}$ and $T_{n,p,s}^{1,\mathcal{S}}$, and then perform a multiple testing correction that controls either family-wise error rate or false discovery rate. We can follow a similar approach when the goal is to test internal validity for the study population conditional on stratum.

5 Simulation Study

In this section, we illustrate the theoretical results in the paper using simulations. We construct a simulation design that allows for treatment effect heterogeneity and for the unobservables in the outcome equation to be correlated with response. We then conduct simulations to demonstrate the performance of the differential attrition rate test as well as both the mean and distributional tests of internal validity.[39] Using both the mean and the distributional restrictions, we conduct tests of internal validity for respondents only as well as the study population. We report rejection probabilities for a range of overall and differential attrition rates.

We begin by outlining our simulation design. We describe our data-generating processes (DGPs) in Panel A of Table 3. First, we construct our outcome equation. We allow treatment to enter into two terms of the outcome equation: $\beta_1 D_{it}$ and $\beta_2 D_{it}\alpha_i$. Letting β_2 be non-zero introduces treatment effect heterogeneity, which allows for the ATE-R to differ from the ATE. We do, however, require that $E[\alpha_i] = 0$ in all variants of our simulation design, so the ATE equals β_1. Next, we randomly assign individual observations into the treatment ($T_i = 1$) and control ($T_i = 0$) groups, and generate the response equation by further assigning individuals to one of the four types of compliance behavior according to proportions given by $p_{r_0 r_1}$ for $(r_0, r_1) \in \{0, 1\}^2$. Now it remains to generate the unobservable components. We let α_i and η_{it} denote the time-invariant and time-varying unobservable in the outcome equation, respectively. We assume that the time-varying unobservable, η_{i1}, follows an AR(1) process and is independent of potential response. We do allow dependence, however, between the time-invariant unobservable, α_i, and potential response. Specifically, the mean of the conditional distribution differs for each subpopulation (i.e. $\delta_{r_0 r_1}$ for all $(r_0, r_1) \in \{0, 1\}^2$). When the conditional mean is different for at least two subpopulations, α_i and potential response are not independent. Conversely, when the conditional mean is the same for all subpopulations, α_i and potential response are independent.

We conduct simulations using four variants of this simulation design, which are summarized

[39] Although our theoretical results indicate the use of distributional tests, the mean restrictions are closer to current applied practice and hence serve as an initial check of our general approach.

Table 3: Simulation Design

Panel A. Data-Generating Process

Outcome:	$Y_{it} = \beta_1 D_{it} + \beta_2 D_{it}\alpha_i + \alpha_i + \eta_{it}$ for $t = 0, 1$ where $\beta_1 = \beta_2 = 0.25$.
Treatment:	$T_i \overset{i.i.d.}{\sim} Bernoulli(0.5)$, $D_{i0} = 0$, $D_{i1} = T_i$.
Response:	$R_i = (1 - T_i)R_i(0) + T_i R_i(1)$ where $p_{r_0 r_1} = P((R_i(0), R_i(1)) = (r_0, r_1))$ for $r_0, r_1 \in \{0, 1\}^2$
Unobservables:	$\begin{cases} U_{it} = (\alpha_i, \eta_{it})', \ t = 0, 1, \\ \alpha_i \vert R_i(0), R_i(1) \overset{i.i.d.}{\sim} \begin{cases} N(\delta_{00}, 1) \ if \ (R_i(0), R_i(1)) = (0, 0), \\ N(\delta_{01}, 1) \ if \ (R_i(0), R_i(1)) = (0, 1), \\ N(\delta_{10}, 1) \ if \ (R_i(0), R_i(1)) = (1, 0), \\ N(\delta_{11}, 1) \ if \ (R_i(0), R_i(1)) = (1, 1). \end{cases} \\ \eta_{i1} = 0.5\eta_{i0} + \epsilon_{i0}, \ (\eta_{i0}, \epsilon_{i0})' \overset{i.i.d.}{\sim} N(0, 0.5 I_2) \end{cases}$

Panel B. Variants of the Design

Design	I	II	III	IV
Monotonicity in the Response Equation	Yes ($p_{10} = 0$)	Yes ($p_{10} = 0$)	Yes ($p_{10} = 0$)	No
Equal Attrition Rates	No	Yes ($p_{01} = 0$)	No	Yes ($p_{10} = p_{01}$)
$U_{it} \perp (R_i(0), R_i(1))$	No	No	Yes	No

Notes: For an integer k, I_k denotes a $k \times k$ identity matrix. In Designs I, II, and IV, we let $\delta_{00} = -0.5$, $\delta_{01} = 0.25$, $\delta_{10} = -0.25$, and $\delta_{11} = -(\delta_{00}p_{00} + \delta_{01}p_{01} + \delta_{10}p_{10})/p_{11}$, such that $E[\alpha_i] = 0$. In Design III, we let $\delta_{r_0 r_1} = 0$ for all $(r_0, r_1) \in \{0, 1\}^2$, which implies $U_{it} \perp (R_i(0), R_i(1))$ for $t = 0, 1$. As for the proportions of the different subpopulations, in Designs I-III, we let $p_{00} = P(R_i = 0 | T_i = 1)$, $p_{01} = P(R_i = 0 | T_i = 0) - P(R_i = 0 | T_i = 1)$, and $p_{11} = 1 - p_{00} - p_{01}$, whereas in Design IV, we fix $p_{00} = p_{10}/4$, and $P(R_i = 0 | T_i = 0) = p_{00} + p_{10}$.

in Panel B of Table 3.[40] Design I demonstrates the case in which the differential attrition rate test would in fact detect a violation of internal validity. This case requires both monotonicity in the response equation as well as dependence between the unobservables that affect the outcome and the potential response ($\alpha_i \not\perp (R_i(0), R_i(1))$). We also allow attrition rates to differ across the treatment and control groups. Design II demonstrates a setting in which there is internal validity for the respondent subpopulation, but not for the study population as a whole. For that set-up, we impose monotonicity in the response equation as well as equal attrition rates, while allowing for dependence between α_i and $(R_i(0), R_i(1))$.

Designs III and IV illustrate *Example 1* and *Example 2* in Section 3.2, respectively. Thus, Design III demonstrates a setting in which we have differential attrition rates but no violation of internal validity. Specifically, Design III relies on the assumptions of monotonicity and differential attrition rates as in Design I, but assumes independence between α_i and $(R_i(0), R_i(1))$. Finally, Design IV follows *Example 2* in demonstrating a case in which there are equal attrition rates and a violation of internal validity. Thus, we allow for dependence between α_i and $(R_i(0), R_i(1))$, and a violation of monotonicity by letting p_{10} and p_{01} be non-zero. We generate equal attrition rates by imposing equality across the proportion of treatment-only and control-only responders, i.e $p_{01} = p_{10}$.

We use a sample size of $n = 2,000$ as well as 2,000 replications. We chose a range of attrition rates from the results of our review of the empirical literature (see Figure 1). We targeted the following pairs of attrition rates in the control and treatment groups in the differential attrition rates designs: $\{(5\%, 2.5\%); (10\%, 5\%); (15\%, 10\%); (20\%, 15\%); (30\%, 20\%)\}$. And, the following pairs in the equal attrition rates design: $\{(5\%, 5\%); (10\%, 10\%); (15\%, 15\%);$
$(20\%, 20\%); (30\%, 30\%)\}$.

5.1 Differential Attrition Rates and Mean Tests of Internal Validity

In this section, we discuss the finite-sample performance of the differential attrition rate test, the mean tests of internal validity for respondents as well as the study population. Table 4 reports simulation rejection probablilities for the relevant tests across Designs I-IV. We also report the estimated difference in mean outcomes for the treatment and control respondents,

$$\bar{Y}^{TR} - \bar{Y}^{CR} = \frac{\sum_{i=1}^{n} Y_{i1} D_{i1} R_i}{\sum_{i=1}^{n} D_{i1} R_i} - \frac{\sum_{i=1}^{n} Y_{i1}(1 - D_{i1}) R_i}{\sum_{i=1}^{n} (1 - D_{i1}) R_i}. \tag{11}$$

its standard deviation, and the rejection probability of a 5% t-test of its significance ($\hat{p}_{0.05}$).

[40] We choose to focus on these four designs to keep the presentation clear. However, it is possible to combine different assumptions. For instance, if we assume $p_{01} = p_{10}$ and $U_{it} \perp (R_i(0), R_i(1))$, then we would have equal attrition rates and internal validity for the study population. We can also obtain a design that satisfies exchangeability by assuming $\delta_{01} = \delta_{10}$. If combined with $p_{01} = p_{10}$, then we would have equal attrition rates and internal validity for respondents only (Proposition 2.iii).

The first three columns of Table 4 report the simulation mean of the attrition rates for the control (C) and treatment (T) groups, as well as the probability of rejecting a 5% differential attrition rate test, i.e. two-sample t-test of the equality of attrition rates between groups.

The differential attrition rate test rejects above the nominal level (5%) in the different variants of Designs I and III, whereas it rejects at nominal levels in the different variants of Designs II and IV. This is not surprising, since the former designs allow for differential attrition rates, whereas the latter impose that the attrition rates are equal. Designs I and II, which obey monotonicity, illustrate the typical cases in which the differential attrition rate test can be viewed as a test of internal validity for respondents. Since in Design I internal validity is violated, the simulation rejection probabilities of the differential attrition rate test is higher than the nominal level. In Design II, however, internal validity for respondents holds, and thus the test controls size.

Design III and IV, on the other hand, illustrate the concerns we raise regarding the use of the differential attrition rate test as a test of internal validity for respondents. In Design III, $U_{it} \perp (R_i(0), R_i(1))$ holds for $t = 0, 1$, i.e. the unobservables that enter the outcome equation are independent of potential response, which implies MAR. Hence, regardless of the response equation and the attrition rates, we not only have internal validity for respondents but also for the study population. The differential attrition rate test however rejects at a frequency higher than the nominal level because the attrition rates are different. Design IV, on the other hand, is a case where we have equal attrition rates but a violation of internal validity. In this case, the differential attrition rate test does not reject above nominal levels. Hence, an empirical researcher using this test as a test of internal validity is likely to falsely reject internal validity in Design III, and less likely to reject internal validity in Design IV.

The next three columns of Table 4 report simulation results of three mean tests of internal validity for respondents which are based on the following mean testable restrictions from Proposition 1(a.ii),

$$\begin{aligned}
H^{1,1}_{0,\mathcal{M}}: &\quad E[Y_{i0}|T_i = 0, R_i = 0] = E[Y_{i0}|T_i = 1, R_i = 0], && (CA - TA) \\
H^{1,2}_{0,\mathcal{M}}: &\quad E[Y_{i0}|T_i = 0, R_i = 1] = E[Y_{i0}|T_i = 1, R_i = 1], && (CR - TR) \\
H^{1}_{0,\mathcal{M}} &: \quad H^{1,1}_{0,\mathcal{M}} \ \& \ H^{1,2}_{0,\mathcal{M}}. && (Joint)
\end{aligned} \tag{12}$$

where the subscript $\mathcal{M}$ denotes the *mean* implication of the relevant distributional restriction. $H^{1,1}_{0,\mathcal{M}}$ ($H^{1,2}_{0,\mathcal{M}}$) tests the implication of internal validity for attritors (respondents) only, whereas $H^{1}_{0,\mathcal{M}}$ is a joint hypothesis of the two. Note that $H^{1}_{0,\mathcal{M}}$ tests the mean implication of the sharp testable restriction in Proposition 1(a.ii). The simulation rejection probabilities reported in the table are based on the asymptotic critical values for two-sample t-tests $t^{1,r}$ of the two simple

hypotheses $H_{0,\mathcal{M}}^{1,r+1}$ for $r = 0, 1$,

$$t^{1,r} = \frac{\bar{Y}_0^{1r} - \bar{Y}_0^{0r}}{s.e.(\bar{Y}_0^{1r} - \bar{Y}_0^{0r})},$$

where $\bar{Y}_0^{\tau r} = \sum_{i=1}^n Y_{i0} 1\{T_i = \tau, R_i = r\} / \sum_{i=1}^n 1\{T_i = \tau, R_i = r\}$ and $s.e.(\bar{Y}^{1r} - \bar{Y}_0^{0r})$ is computed assuming equal variance. The joint null hypothesis is tested using a Wald statistic constructed from the two t-statistics.

The three mean tests of internal validity for respondents behave according to our theoretical predictions. In Designs II and III, where internal validity for respondents holds, the three tests control size. In Designs I and IV, where internal validity for respondents is violated, they reject with simulation probability above the nominal level. The simulation results illustrate the potential power gains in finite samples from using the attritors in performing this test. In our simulation design, the rejection probability of the t-test that compares the means of the treatment and control attritors (CA-TA) rejects at a significantly higher proportion than the test based on the mean difference between the treatment and control respondents (CR-TR).[41] In general, the relative power of the test statistics based on the attritor or respondent subsamples may differ depending on the data-generating process. We recommend testing the joint null hypothesis of internal validity for respondents, $H_{0,\mathcal{M}}^1$, since it is based on the sharp testable restriction of internal validity for respondents.

We next consider the mean test of internal validity for the study population given in Proposition 1(b.ii), specifically

$$\begin{aligned}
H_{0,\mathcal{M}}^{2,1}: &\quad E[Y_{i0}|T_i = 0, R_i = 1] = E[Y_{i0}|T_i = 0, R_i = 0], && (CR - CA) \\
H_{0,\mathcal{M}}^{2,2}: &\quad E[Y_{i0}|T_i = 1, R_i = 1] = E[Y_{i0}|T_i = 1, R_i = 0], && (TR - TA) \\
H_{0,\mathcal{M}}^{2}: &\quad H_{0,\mathcal{M}}^{2,1} \ \& \ H_{0,\mathcal{M}}^{2,2}. && (Joint)
\end{aligned} \tag{13}$$

The rejection probabilities reported in the table are based on the asymptotic critical values of the two-sample t-tests $t^{2,\tau}$ of the simple hypotheses $H_{0,\mathcal{M}}^{2,\tau+1}$ for $\tau = 0, 1$,

$$t^{2,\tau} = \frac{\bar{Y}_0^{\tau 1} - \bar{Y}_0^{\tau 0}}{s.e.(\bar{Y}_0^{\tau 1} - \bar{Y}_0^{\tau 0})}$$

where $s.e.(\bar{Y}_0^{\tau 1} - \bar{Y}_0^{\tau 0})$ is computed assuming equal variance. A Wald test is constructed from the two t-statistics to test the joint hypothesis. These test statistics also behave according to our theoretical predictions. In Designs I, II and IV, they reject internal validity for the study population at a simulation frequency higher than the nominal level. Design II is interesting

[41] This is because the treatment-only responders are proportionately larger in the control attritor subgroup than in the treatment respondent subgroup.

because we only have internal validity for respondents. Hence, while the mean tests of its testable implications are not rejected above the nominal level, the tests of internal validity for the study population are rejected above the nominal level. This is illustrated empirically by examining the difference in mean outcomes between treatment and control respondents, which is internally valid for the respondents in this case. However, its simulation mean is different from the ATE (0.25). In Design III, which is the only design where internal validity for the study population holds, the tests control size. Note in this design, the simulation results illustrate that the difference in mean outcomes between treatment and control respondents is unbiased for the ATE across all combinations of attrition rates.

5.2 Distributional Tests of Internal Validity

Now we examine the finite-sample behavior of the randomization procedure proposed in Section 4 to obtain p-values for the KS and CM statistics of the testable restrictions of internal validity for the respondent subpopulation and the study population.

Table 5 presents the simulation rejection probabilities of the distributional tests of internal validity for respondents. As in the previous section, we present test statistics of the two simple distributional null hypotheses as well as their joint hypothesis,

$$\begin{aligned} H_0^{1,1}: &\quad Y_{i0}|T_i=1,R_i=0 \overset{d}{=} Y_{i0}|T_i=0,R_i=0, && (CA-TA) \\ H_0^{1,2}: &\quad Y_{i0}|T_i=1,R_i=1 \overset{d}{=} Y_{i0}|T_i=0,R_i=1, && (CR-TR) \\ H_0^{1}: &\quad H_0^{1,1} \ \& \ H_0^{1,2}. && (Joint) \end{aligned} \tag{14}$$

For $r=0,1$, the KS and CM statistics to test $H_0^{1,r+1}$ is given by

$$\begin{aligned} KS_{n,r}^1 &= \max_{i:R_i=r} \left| \sqrt{n} \left(F_{n,Y_{i0}}(y_{i0}|T_i=1,R_i=r) - F_{n,Y_{i0}}(y_{i0}|T_i=0,R_i=r) \right) \right|. \\ CM_{n,r}^1 &= \frac{\sum_{i:R_i=r} \left(\sqrt{n} (F_{n,Y_{i0}}(y_{i0}|T_i=1,R_i=r) - F_{n,Y_{i0}}(y_{i0}|T_i=0,R_i=r) \right)^2}{\sum_{i=1}^n 1\{R_i=r\}} \end{aligned} \tag{15}$$

For the joint hypothesis H_0^1, which is the sharp testable restriction in Proposition 1(b.ii), we consider either $KS_{n,m}^1 = \max\{KS_{n,0}^1, KS_{n,1}^1\}$ or $KS_{n,p}^1 = p_{n,0}KS_{n,0}^1 + p_{n,1}KS_{n,1}^1$, where $p_{n,r} = \sum_{i=1}^n 1\{R_i=r\}/n$ for $r=0,1$. $CM_{n,m}^1$ and $CM_{n,p}^1$ are similarly defined.

In Table 5, we present the rejection probabilities for the KS statistics of the simple hypotheses, $KS_{n,0}^1$ and $KS_{n,1}^1$, using asymptotic critical values (KS $(Asym.)$) as a benchmark to compare the performance of our randomization procedure. We also report the rejection probabilities for different variants of the KS (KS (R)) and the CM (CM (R)) statistics using the p-values obtained from our randomization procedure ($B = 199$). The simulation results illustrate that the randomization procedure yields rejection probabilities for the two-sample KS

statistics, $KS^1_{n,0}$ and $KS^1_{n,1}$, that are very similar to those obtained from the asymptotic critical values.

Next, we consider the relative finite sample performance of KS and CM statistics in testing internal validity for respondents. The different variants of the KS and CM test statistics control size under Designs II and III, where internal validity for respondents holds. They also have non-trivial power in finite samples in Designs I and IV, when internal validity for respondents is violated. The simulation results for the distributional statistics also illustrate the potential power gains in finite samples from using the attritor subgroup in testing internal validity for respondents. In testing the joint null hypothesis, we find that $KS^1_{n,m}$ and $CM^1_{n,m}$ $(Joint\ (m))$ exhibit better finite-sample power properties than $KS^1_{n,p}$ and $CM^1_{n,p}$ $(Joint\ (p))$. Finally, the CM statistics generally have better finite-sample properties than their respective KS statistics, while maintaining comparable size control, in our simulation design.

We then examine the finite-sample performance of the distributional statistics of internal validity for the study population presented in Table 6. Proposition 1(b.ii) implies the three simple null hypotheses as well as their joint hypothesis below,

$$\begin{aligned}
H_0^{2,1}: &\quad Y_{i0}|T_i=0,R_i=0 \overset{d}{=} Y_{i0}|T_i=0,R_i=1, && (CA-CR)\\
H_0^{2,2}: &\quad Y_{i0}|T_i=0,R_i=1 \overset{d}{=} Y_{i0}|T_i=1,R_i=0, && (CR-TA)\\
H_0^{2,3}: &\quad Y_{i0}|T_i=1,R_i=0 \overset{d}{=} Y_{i0}|T_i=1,R_i=1, && (TA-TR)\\
H_0^{2}: &\quad H_0^{2,1}\ \&\ H_0^{2,2}\ \&\ H_0^{2,3}. && (Joint)
\end{aligned} \tag{16}$$

Let (τ_j, r_j) denote the j^{th} element of $\mathcal{T}\times\mathcal{R}=\{(0,0),(0,1),(1,0),(1,1)\}$. We can define the KS and CM statistics for $H_0^{2,j}$ for each $j=1,2,3$ by the following,

$$\begin{aligned}
KS^2_{n,j} &= \max_{i:(T_i,R_i)\in\{(\tau_j,r_j),(\tau_{j+1},r_{j+1})\}} \left|\sqrt{n}\left(F_{n,Y_{i0}|T_i=\tau_{j-1},R_i=r_{j-1}}-F_{n,Y_{i0}|T_i=\tau_j,R_i=r_j}\right)\right|,\\
CM^2_{n,j} &= \frac{\sum_{i:(T_i,R_i)\in\{(\tau_j,r_j),(\tau_{j+1},r_{j+1})} \left(\sqrt{n}\left(F_{n,Y_{i0}|T_i=\tau_{j-1},R_i=r_{j-1}}-F_{n,Y_{i0}|T_i=\tau_j,R_i=r_j}\right)\right)^2}{\sum_{i=1}^n 1\left\{(T_i,R_i)\in\{(\tau_j,r_j),(\tau_{j+1},r_{j+1})\}\right\}},
\end{aligned} \tag{17}$$

The joint hypothesis H_0^2 is tested using the joint statistics $KS^2_{n,m}=\max_{j=1,2,3} KS^2_{n,j}$ and $CM^2_{n,m}=\max_{j=1,2,3} CM^2_{n,j}$.

Similar to Table 5, we report the simulation rejection probabilities for the KS test of the single hypotheses using asymptotic critical values in addition to the rejection probabilities computed using our randomization procedure for all variants of the KS and CM statistics. Under Designs I, II and IV, internal validity for the study population is violated, the rejection probabilities for all the test statistics we consider tend to be higher than the nominal level, as we would expect. The joint KS and CM test statistics behave similarly in this design and

have comparable finite-sample power properties to the test statistic of the single hypothesis (TA-TR), which has the best finite-sample power properties in our simulation design. Finally, in Design III, where internal validity holds for the study population, our simulation results illustrate that the test statistics we consider control size.

Overall, the simulation results illustrate that the randomization procedure we propose performs well in finite samples. We present a more thorough examination of the relative power properties of the test statistics using the simple and joint null hypotheses in Appendix C.

6 Conclusion

This paper has important implications for the empirical practice in testing for attrition bias in field experiments with baseline outcome data. It highlights the importance of selective attrition tests and points out that differential attrition rates should be interpreted with caution. These findings have important implications for the procedures of the WWC (WWC, 2017). There are however several directions for future research. Despite the availability of several approaches to correct for attrition bias (Lee, 2009; Behagel et al., 2015; Millán and Macours, 2017), alternative approaches building on the framework presented here is left for future work. Furthermore, if the test statistics proposed in this paper are used as pre-tests in the empirical analysis, then standard inference methods fail. Inference procedures that correct for the resulting pre-test bias are another important direction for future research.

Table 4: Simulation Results on Differential Attrition Rates and Mean Tests of Internal Validity ($ATE = 0.25$)

Design	Attrition Rates		Differential Attrition Rate Test	Mean Tests of Internal Validity for Respondents			Mean Tests of Internal Validity for Study Population			Difference in Mean Outcomes between Treatment & Control Respondents ($\bar{Y}_1^{TR} - \bar{Y}_1^{CR}$)		
	C	T	$\hat{p}_{0.05}$	CR-TR	CA-TA	Joint	CR-CA	TR-TA	Joint	Mean	SD	$\hat{p}_{0.05}$
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$												
I	0.05	0.025	0.866	0.049	0.394	0.323	0.064	0.605	0.519	0.265	0.057	0.997
	0.10	0.05	0.995	0.077	0.682	0.596	0.063	0.897	0.839	0.282	0.058	0.998
	0.15	0.10	0.935	0.073	0.600	0.514	0.510	0.993	0.997	0.288	0.061	0.997
	0.20	0.15	0.867	0.074	0.499	0.419	0.933	1.000	1.000	0.296	0.063	0.996
	0.30	0.20	1.000	0.151	0.882	0.834	0.852	1.000	1.000	0.334	0.066	0.999
Equal Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))^{\dagger}$												
II	0.05	0.05	0.049	0.046	0.048	0.053	0.912	0.899	0.989	0.255	0.058	0.993
	0.10	0.10	0.053	0.043	0.046	0.045	0.996	0.994	1.000	0.262	0.060	0.991
	0.15	0.15	0.052	0.044	0.049	0.052	1.000	1.000	1.000	0.271	0.062	0.992
	0.20	0.20	0.049	0.044	0.044	0.051	1.000	1.000	1.000	0.280	0.064	0.990
	0.30	0.30	0.048	0.053	0.045	0.048	1.000	1.000	1.000	0.303	0.068	0.991
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$ (*Example 1*)*												
III	0.05	0.025	0.866	0.055	0.052	0.059	0.064	0.054	0.061	0.248	0.058	0.990
	0.10	0.05	0.995	0.054	0.051	0.055	0.054	0.046	0.050	0.248	0.059	0.985
	0.15	0.10	0.935	0.056	0.051	0.057	0.050	0.043	0.051	0.247	0.061	0.983
	0.20	0.15	0.867	0.059	0.049	0.052	0.044	0.052	0.047	0.247	0.063	0.974
	0.30	0.20	1.000	0.055	0.051	0.052	0.048	0.049	0.047	0.248	0.066	0.964
Equal Attrition Rates + Violation of Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$ (*Example 2*)												
IV	0.05	0.05	0.012	0.067	0.404	0.329	0.096	0.487	0.430	0.273	0.058	0.997
	0.10	0.10	0.013	0.131	0.697	0.646	0.157	0.805	0.768	0.302	0.059	0.999
	0.15	0.15	0.007	0.250	0.865	0.846	0.216	0.935	0.920	0.333	0.061	0.999
	0.20	0.20	0.004	0.423	0.933	0.950	0.282	0.987	0.980	0.367	0.063	0.999
	0.30	0.30	0.001	0.800	0.990	0.997	0.437	1.000	1.000	0.452	0.067	1.000

Notes: The above table reports simulation summary statistics for $n = 2,000$ across $2,000$ simulation replications. C denotes the control group, T denotes the treatment group, and $\hat{p}_{0.05}$ denotes the simulation rejection probability of a 5% test. The *mean tests of internal validity for respondents (study population)* refer to the two-sample selective attrition tests in (12) ((13)). And, the *difference in mean outcomes between treatment and control respondents* refers to the two-sample test in (11). All tests are conducted using a nominal level $\alpha = 0.05$. Additional details of the design are provided in Table 3.
† (*) indicates internal validity for respondents only (study population).

Table 5: Simulation Results on the KS & CM Randomization Test of Internal Validity for Respondents

Design	Att. Rate		KS (*Asym.*)		KS (R)				CM(R)			
	C	T	CR-TR	CA-TA	CR-TR	CA-TA	Joint (m)	Joint (p)	CR-TR	CA-TA	Joint (m)	Joint (p)
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \measuredangle (R_i(0), R_i(1))$												
I	0.050	0.025	0.058	0.316	0.058	0.324	0.324	0.081	0.058	0.353	0.353	0.285
	0.100	0.050	0.066	0.589	0.071	0.582	0.582	0.157	0.072	0.636	0.636	0.568
	0.150	0.100	0.067	0.460	0.067	0.483	0.483	0.167	0.069	0.544	0.544	0.460
	0.200	0.150	0.070	0.392	0.073	0.412	0.412	0.180	0.069	0.462	0.462	0.385
	0.300	0.200	0.111	0.790	0.123	0.801	0.801	0.502	0.135	0.855	0.855	0.803
Equal Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \measuredangle (R_i(0), R_i(1))^{\dagger}$												
II	0.050	0.050	0.052	0.059	0.053	0.062	0.062	0.052	0.054	0.056	0.056	0.061
	0.100	0.100	0.049	0.054	0.053	0.056	0.056	0.050	0.054	0.054	0.054	0.053
	0.150	0.150	0.044	0.049	0.049	0.055	0.055	0.051	0.049	0.054	0.054	0.055
	0.200	0.200	0.052	0.044	0.052	0.050	0.050	0.058	0.052	0.049	0.049	0.052
	0.300	0.300	0.051	0.043	0.051	0.042	0.043	0.053	0.049	0.047	0.048	0.057
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$ (*Example 1*)*												
III	0.050	0.025	0.049	0.051	0.054	0.052	0.052	0.056	0.048	0.051	0.051	0.049
	0.100	0.050	0.047	0.042	0.050	0.046	0.046	0.047	0.053	0.047	0.047	0.043
	0.150	0.100	0.047	0.038	0.052	0.045	0.045	0.047	0.049	0.049	0.049	0.048
	0.200	0.150	0.054	0.031	0.053	0.036	0.036	0.047	0.055	0.036	0.036	0.044
	0.300	0.200	0.050	0.043	0.050	0.043	0.043	0.050	0.051	0.042	0.042	0.050
Equal Attrition Rates + Violation of Monotonicity + $(U_{i0}, U_{i1}) \measuredangle (R_i(0), R_i(1))$ (*Example 2*)												
IV	0.050	0.050	0.059	0.332	0.065	0.329	0.329	0.093	0.067	0.375	0.375	0.302
	0.100	0.100	0.102	0.569	0.102	0.577	0.577	0.230	0.116	0.663	0.663	0.593
	0.150	0.150	0.178	0.740	0.190	0.758	0.758	0.465	0.211	0.816	0.816	0.805
	0.200	0.200	0.313	0.854	0.319	0.859	0.859	0.709	0.368	0.917	0.916	0.910
	0.300	0.300	0.683	0.970	0.680	0.972	0.974	0.974	0.760	0.985	0.991	0.996

Notes: The above table presents the rejection probabilities of the KS and CM tests for the simple and joint null hypotheses in (14). We use the nominal level $\alpha = 0.05$, 2,000 simulation replications and $n = 2{,}000$. C denotes the control group, T denotes the treatment group. $KS(Asym.)$ refers to the two-sample KS test using the asymptotic critical values. $KS(R)$ and $CM(R)$ refer to the randomization KS and CM tests, respectively, for the two-sample problem as well as the joint hypothesis. *Joint* (m) and *Joint* (p) denote the randomization procedure applied to $KS^1_{n,m}$ ($CM^1_{n,m}$) and $KS^1_{n,p}$ ($CM^1_{n,p}$), respectively. Additional details of the design are provided in Table 3.

† (*) indicates internal validity for respondents only (study population).

Table 6: Simulation Results on the KS & CM Randomization Test of Internal Validity for Study Population

Design	Att. Rate		KS (*Asym.*)			KS (R)				CM(R)			
	C	T	CA-CR	CR-TA	TA-TR	CA-CR	CR-TA	TA-TR	*Joint* (m)	CA-CR	CR-TA	TA-TR	*Joint* (m)
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \measuredangle (R_i(0), R_i(1))$													
I	0.050	0.025	0.051	0.451	0.456	0.064	0.482	0.485	0.476	0.053	0.492	0.497	0.483
	0.100	0.050	0.053	0.746	0.787	0.055	0.763	0.801	0.787	0.058	0.806	0.837	0.824
	0.150	0.100	0.414	0.970	0.980	0.420	0.969	0.978	0.980	0.463	0.983	0.986	0.989
	0.200	0.150	0.865	0.999	0.998	0.870	0.998	0.998	1.000	0.902	1.000	0.999	1.000
	0.300	0.200	0.774	1.000	1.000	0.771	1.000	1.000	1.000	0.825	1.000	1.000	1.000
Equal Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \measuredangle (R_i(0), R_i(1))^{\dagger}$													
II	0.050	0.050	0.772	0.788	0.788	0.780	0.797	0.804	0.902	0.831	0.840	0.841	0.939
	0.100	0.100	0.984	0.983	0.980	0.985	0.981	0.981	0.999	0.994	0.989	0.986	1.000
	0.150	0.150	1.000	1.000	0.998	1.000	1.000	0.998	1.000	1.000	1.000	0.999	1.000
	0.200	0.200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.300	0.300	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$ (*Example 1*)*													
III	0.050	0.025	0.040	0.042	0.043	0.044	0.050	0.051	0.050	0.047	0.053	0.053	0.054
	0.100	0.050	0.051	0.041	0.048	0.058	0.052	0.052	0.055	0.056	0.050	0.057	0.056
	0.150	0.100	0.040	0.051	0.052	0.046	0.056	0.057	0.059	0.047	0.054	0.055	0.059
	0.200	0.150	0.037	0.040	0.045	0.041	0.046	0.050	0.048	0.046	0.045	0.054	0.050
	0.300	0.200	0.048	0.044	0.044	0.050	0.049	0.046	0.048	0.049	0.044	0.051	0.054
Equal Attrition Rates + Violation of Monotonicity + $(U_{i0}, U_{i1}) \measuredangle (R_i(0), R_i(1))$ (*Example 2*)													
IV	0.050	0.050	0.075	0.325	0.361	0.082	0.350	0.384	0.311	0.097	0.363	0.407	0.342
	0.100	0.100	0.113	0.548	0.668	0.125	0.558	0.681	0.582	0.152	0.605	0.742	0.661
	0.150	0.150	0.169	0.683	0.854	0.180	0.694	0.858	0.792	0.220	0.756	0.908	0.861
	0.200	0.200	0.234	0.759	0.947	0.239	0.762	0.950	0.913	0.288	0.822	0.974	0.952
	0.300	0.300	0.371	0.805	0.999	0.376	0.813	0.999	0.998	0.440	0.875	1.000	1.000

Notes: The above table presents the rejection probabilities of the KS and CM tests for the simple and joint null hypotheses in (16). We use the nominal level $\alpha = 0.05$, 2,000 simulation replications and $n = 2,000$. C denotes the control group, T denotes the treatment group. $KS(Asym.)$ refers to the two-sample KS test using the asymptotic critical values. $KS(R)$ and $CM(R)$ refer to the randomization KS and CM tests, respectively, for the two-sample problem as well as the joint hypothesis. *Joint* (m) denotes the randomization procedure applied to $KS^2_{n,m}$ ($CM^2_{n,m}$). Additional details of the design are provided in Table 3.
† (*) indicates internal validity for respondents only (study population).

A Attrition Tests in the Field Experiments Literature

In this section, we describe and classify the different econometric strategies that are carried out to test for attrition bias in the papers that we review. We use the following notation to facilitate the exposition of each strategy and the comparison across them:
-Let R_i take the value of 1 if individual i belongs to the follow-up sample.
-Let T_i take the value of 1 if individual i belongs to the treatment group.
-Let X_{i0} be a $k \times 1$ vector of baseline variables.
-Let Y_{i0} be a $l \times 1$ vector of outcomes collected at baseline.
-Let $Z_{i0} = (X'_{i0}, Y'_{i0})'$.

In the following, for the regression tests, we intentionally do not specify the null hypothesis. Since we seek to categorize papers as generously as possible, we include any paper that performs a regression under the following categories as performing the relevant test, whether or not the paper specifies the null hypothesis.

A.1 Differential Attrition Rate Test

The *differential attrition rate test* determines whether the rates of attrition are statistically significantly different across treatment and control groups.

1. t-test of the equality of attrition rate by treatment group, i.e. $H_0 : P(R_i = 0|T_i = 1) = P(R_i = 0|T_i = 0)$.

2. $R_i = \gamma + T_i\beta + U_i$; may include strata fixed effects.

3. $R_i = \gamma + T_i\beta + X'_{i0}\theta + Y'_{i0}\alpha + U_i$; may include strata fixed effects.

A.2 Selective Attrition Test

The *selective attrition test* determines whether, conditional on response status, the distribution of observable characteristics is the same across treatment and control groups. We also identify two sub-types of selective attrition tests: i) a simple test conducted only on respondents or attritors, and ii) a joint test conducted on both respondents and attritors. Let Z^j_{i0} be the j^{th} element of Z_{i0}.

A.2.1 Single tests:

1. t-test of baseline characteristics by treatment group among respondents, i.e. $H^j_0 : E[Z^j_{i0}|T_i = 1, R_i = 1] = E[Z^j_{i0}|T_i = 0, R_i = 1])$ for $j = 1, 2, \ldots, (l + k)$.

2. $Z^j_{i0} = \gamma + T_i\beta^j + U^j_i$ if $R_i = 1$, for $j = 1, 2, \ldots, (l + k)$; may include strata fixed effects.

3. $T_i = \gamma + X'_{i0}\theta + Y'_{i0}\alpha + U_i$ if $R_i = 1$; may include strata fixed effects.

4. Kolmogorov-Smirnov (KS) test of baseline characteristics by treatment group among respondents, $H_0^j : F_{Z^j_{i0}|T_i,R_i=1} = F_{Z^j_{i0}|R_i=1}$ for $j = 1, 2, \ldots, (l+k)$.

5. $Z^j_{i0} = \gamma + T_i\beta^j + U^j_i$ if $R_i = 1$, for $j = 1, 2, \ldots, (l+k)$; may include strata fixed effects.

6. $Z^j_{i0} = \gamma + T_i\beta^j + U^j_i$ if $R_i = 0$, for $j = 1, 2, \ldots, (l+k)$; may include strata fixed effects.

A.2.2 Joint tests:

1. $Z^j_{i0} = \gamma^j + T_i\beta^j + (1 - R_i)\lambda^j + T_i(1 - R_i)\phi^j + U^j_i$; may include strata fixed effects.

2. $R_i = \gamma + T_i\beta + X'_{i0}\theta + Y'_{i0}\alpha + T_iX'_{i0}\lambda_1 + T_iY'_{i0}\lambda_2 + U_i$; may include strata fixed effects.

3. t-test of the null hypothesis: $E[Y_{i0}|T_i = 1, R_i = 1] - E[Y_{i0}|T_i = 1, R_i = 0] = E[Y_{i0}|T_i = 0, R_i = 1] - E[Y_{i0}|T_i = 0, R_i = 0]$.

A.3 Determinants of Attrition Test

The *determinants of attrition test* determines whether attritors are significantly different from respondents regardless of treatment assignment. Such test aims to better understand the impact of attrition on internal validity for the study population.

1. $R_i = \gamma + T_i\beta + X'_{i0}\theta + Y'_{i0}\alpha + U_i$; may include strata fixed effects.

2. $Z^j_{i0} = \gamma^j + (1 - R_i)\lambda^j + U^j_i$, $j = 1, 2, \ldots, (l+k)$; may include strata fixed effects.

3. $R_i = \gamma + X'_{i0}\theta + Y'_{i0}\alpha + U_i$; may include strata fixed effects.

4. Let $Reason_i$ take the value of 1 if the individual identifies it as one of the reasons for which she dropped out of the program. The test consists of a Probit estimation of: $Reason_i = \gamma + T_i\beta + U_i$ if $R_i = 1$; may include strata fixed effects.

B Proofs

Proof. (Proposition 1)
(a) Under the assumptions imposed it follows that $F_{U_{i0},U_{i1}|T_i,R_i} = F_{U_{i0},U_{i1}|R_i}$, which implies that for $d = 0, 1$ $F_{Y_{it(d)}|T_i,R_i} = \int 1\{\mu_t(d,u) \leqslant .\} dF_{U_{it}|T_i,R_i}(u) = \int 1\{\mu_t(d,u) \leqslant .\} dF_{U_{it}|R_i}(u) = F_{Y_{it(d)}|R_i}$. (i) follows by letting $t = 1$ and $d = 0$, while conditioning the left-hand side of the last equality on $T_i = 0$ and $R_i = 1$, and the testable implication in (ii) follows by letting $t = d = 0$.

Following Hsu et al. (2016), we show that the testable restriction is sharp by showing that if $(Y_{i0}, Y_{i1}, T_i, R_i)$ satisfy $Y_{i0}|T_i = 0, R_i = r \overset{d}{=} Y_{i0}|T_i = 1, R_i = r$ for $r = 0, 1$, then there

exists (U_{i0}, U_{i1}) such that $Y_{it}(d) = \mu_t(d, U_{it})$ for some $\mu_t(d, .)$ for $d = 0, 1$ and $t = 0, 1$ and $(U_{i0}, U_{i1}) \perp T_i | R_i$ that generate the observed distributions. By the arbitrariness of U_{it} and μ_t, we can let $U_{it} = (Y_{it}(0), Y_{it}(1))'$ and $\mu_t(d, U_{it}) = dY_{it}(1) + (1-d)Y_{it}(0)$ for $d = 0, 1$, $t = 0, 1$. Note that $Y_{i0} = Y_{i0}(0)$ since $D_{i0} = 0$ w.p.1. Now we need to construct a distribution of $U_i = (U_{i0}', U_{i1}')$ that satisfies

$$F_{U_i|T_i,R_i} \equiv F_{Y_{i0}(0),Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|T_i,R_i} = F_{Y_{i0}(0),Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|R_i}$$

as well as the relevant equalities between potential and observed outcomes. We proceed by first constructing the unobservable distribution for the respondents. By setting the appropriate potential outcomes to their observed counterparts, we obtain the following equalities for the distribution of U_i for the treatment and control respondents

$$\begin{aligned}
F_{U_i|T_i=0,R_i=1} &= F_{Y_{i0}(0),Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|T_i=0,R_i=1} = F_{Y_{i0}(1),Y_{i1},Y_{i1}(1)|Y_{i0},T_i=0,R_i=1} F_{Y_{i0}|T_i=0,R_i=1} \\
F_{U_i|T_i=1,R_i=1} &= F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}|Y_{i0},T_i=1,R_i=1} F_{Y_{i0}|T_i=1,R_i=1}
\end{aligned}$$

By construction, $F_{Y_{i0}|T_i,R_i=1} = F_{Y_{i0}|R_i=1}$. Now generating the above two distributions using $F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|Y_{i0},T_i,R_i=1}$ which satisfies $F_{Y_{i0}(1),Y_{i1},Y_{i1}(1)|Y_{i0},T_i=0,R_i=1} = F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}|Y_{i0},T_i=1,R_i=1}$ yields $U_i \perp T_i | R_i = 1$ and we can construct the observed outcome distribution $(Y_{i0}, Y_{i1})|R_i = 1$ from $U_i|R_i = 1$.

The result for the attritor subpopulation follows trivially from the above arguments,

$$\begin{aligned}
F_{U_i|T_i=0,R_i=0} &= F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|Y_{i0},T_i=0,R_i=0} F_{Y_{i0}|T_i=0,R_i=0}, \\
F_{U_i|T_i=1,R_i=0} &= F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|Y_{i0},T_i=1,R_i=0} F_{Y_{i0}|T_i=1,R_i=0},
\end{aligned}$$

Since $F_{Y_{i0}|T_i,R_i=0} = F_{Y_{i0}|R_i=0}$ by construction, it remains to generate the above two distributions using the same $F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|Y_{i0},R_i=0}$. This leads to a distribution of $U_i|R_i = 0$ that is independent of T_i and that generates the observed outcome distribution $Y_{i0}|R_i = 0$.

(b) Under the given assumptions, it follows that $F_{U_{i0},U_{i1}|T_i,R_i} = F_{U_{i0},U_{i1}|T_i} = F_{U_{i0},U_{i1}}$ where the last equality follows by random assignment. Similar to (a), the above implies that $F_{Y_{it}(d)|T_i,R_i}(.) = \int 1\{\mu_t(d,u) \leqslant .\} dF_{U_{it}|T_i,R_i}(u) = \int 1\{\mu_t(d,u) \leqslant .\} dF_{U_{it}}(u) = F_{Y_{it}(d)}$. (i) follows by letting $d = 0$ and $t = 1$, while conditioning the left-hand side of the last equality on $T_i = 0$ and $R_i = 1$, whereas (ii) follows by letting $d = t = 0$ while conditioning on $T_i = \tau$ and $R_i = r$ for $\tau = 0, 1$, $r = 0, 1$.

To show that the testable restriction is sharp, it remains to show that if $(Y_{i0}, Y_{i1}, T_i, R_i)$ satisfies $Y_{i0}|T_i, R_i \overset{d}{=} Y_{i0}(0)$, then there exists (U_{i0}, U_{i1}) such that $Y_{it}(d) = \mu_t(d, U_{it})$ for some $\mu_t(d, .)$ for $d = 0, 1$ and $t = 0, 1$ and $(U_{i0}, U_{i1}) \perp (T_i, R_i)$. Similar to (a.ii), we let $U_{it} =$

$(Y_{it}(0), Y_{it}(1))'$ and $\mu_t(d, U_{it}) = dY_{it}(1) + (1-d)Y_{it}(0)$. Then $Y_{i0} = Y_{i0}(0)$ by similar arguments as in the above. Furthermore, $F_{Y_{i0}|T_i,R_i} = F_{Y_{i0}}$ by construction and it follows immediately that

$$\begin{aligned}
F_{U_i|T_i=0,R_i=1} &= F_{Y_{i0}(1),Y_{i1},Y_{i1}(1)|Y_{i0}T_i=0,R_i=1}F_{Y_{i0}},\\
F_{U_i|T_i=1,R_i=1} &= F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}|Y_{i0},T_i=1,R_i=1}F_{Y_{i0}},\\
F_{U_i|T_i=0,R_i=0} &= F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|Y_{i0},T_i=0,R_i=0}F_{Y_{i0}},\\
F_{U_i|T_i=1,R_i=0} &= F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|Y_{i0},T_i=1,R_i=0}F_{Y_{i0}}.
\end{aligned}$$

Now constructing all of the above distributions using the same $F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}(1)|T_i,R_i}$ that satisfies $F_{Y_{i0}(1),Y_{i1},Y_{i1}(1)|Y_{i0},T_i=0,R_i=1} = F_{Y_{i0}(1),Y_{i1}(0),Y_{i1}|Y_{i0},T_i=1,R_i=1}$ implies the result. □

Proof. (Proposition 2) For notational brevity, let $U_i = (U_{i0}', U_{i1}')$. We first note that by random assignment, it follows that

$$F_{U_i|T_i,R_i(0),R_i(1)} = F_{U_i|T_i,\xi(0,V_i),\xi(1,V_i)} = F_{U_i|\xi(0,V_i),\xi(1,V_i)} = F_{U_i|R_i(0),R_i(1)}. \tag{18}$$

As a result,

$$F_{U_i|T_i=1,R_i=1} = \frac{p_{01}F_{U_i|(R_i(0),R_i(1))=(0,1)} + p_{11}F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i=1|T_i=1)}, \tag{19}$$

$$F_{U_i|T_i=0,R_i=1} = \frac{p_{10}F_{U_i|(R_i(0),R_i(1))=(1,0)} + p_{11}F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i=1|T_i=0)}. \tag{20}$$

If (i) holds, then $F_{U_i|R_i(0),R_i(1)} = F_{U_i|\xi(0,V_i),\xi(1,V_i)} = F_{U_i}$, hence

$$F_{U_i|R_i=1,T_i=1} = \frac{p_{01}F_{U_i} + p_{11}F_{U_i}}{P(R_i=1|T_i=1)} = F_{U_i}.$$

It follows trivially that $U_i|T_i, R_i \overset{d}{=} U_i|R_i$.

Alternatively, if we assume (ii), $R_i(0) \leqslant R_i(1)$ implies $p_{10} = 0$. As a result, $P(R_i=0|T_i=1) = P(R_i=0|T_i=0)$ iff $p_{01} = 0$. It follows that the terms in (19) and (20) both equal $F_{U_i|(R_i(0),R_i(1))=(1,1)}$. Similar analysis implies that $F_{U_i|T_i=1,R_i=0} = F_{U_i|T_i=0,R_i=0}$, which implies the result.

Finally, suppose (iii) holds, then equal attrition rates imply that $p_{01} = p_{10}$. The exchangeability restriction implies that $F_{U_i|(R_i(0),R_i(1))=(0,1)} = F_{U_i|(R_i(0),R_i(1))=(1,0)}$. Hence,

$$\begin{aligned}
F_{U_i|T_i=1,R_i=1} &= \frac{p_{01}F_{U_i|(R_i(0),R_i(1))=(0,1)} + p_{11}F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i=1|T_i=1)}\\
&= \frac{p_{10}F_{U_i|(R_i(0),R_i(1))=(1,0)} + p_{11}F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i=1|T_i=0)} = F_{U_i|T_i=0,R_i=1}.
\end{aligned} \tag{21}$$

Similarly, it follows that $F_{U_i|T_i=1,R_i=0} = F_{U_i|T_i=0,R_i=0}$. Hence, the result follows. □

Proof. (Proposition 3) The proof is immediate from the proof of Proposition 1 by conditioning all statements on S_i. □

C Additional Simulation Results

In this section, we present additional simulation results to illustrate the relative power properties of using the simple vs joint tests of internal validity. Panel A in Figure 5 displays the simulation rejection probabilities of the tests of internal validity for respondents while Panel B displays the simulation rejection probabilities of the tests of internal validity for the study population. We show the results of the distributional tests for the case where $P(R_i = 0|T_i = 0) = 0.15$, and report the simulation results for the CM statistics since they seemed to perform better than the KS statistics in our simulation study in Section 5.[42]

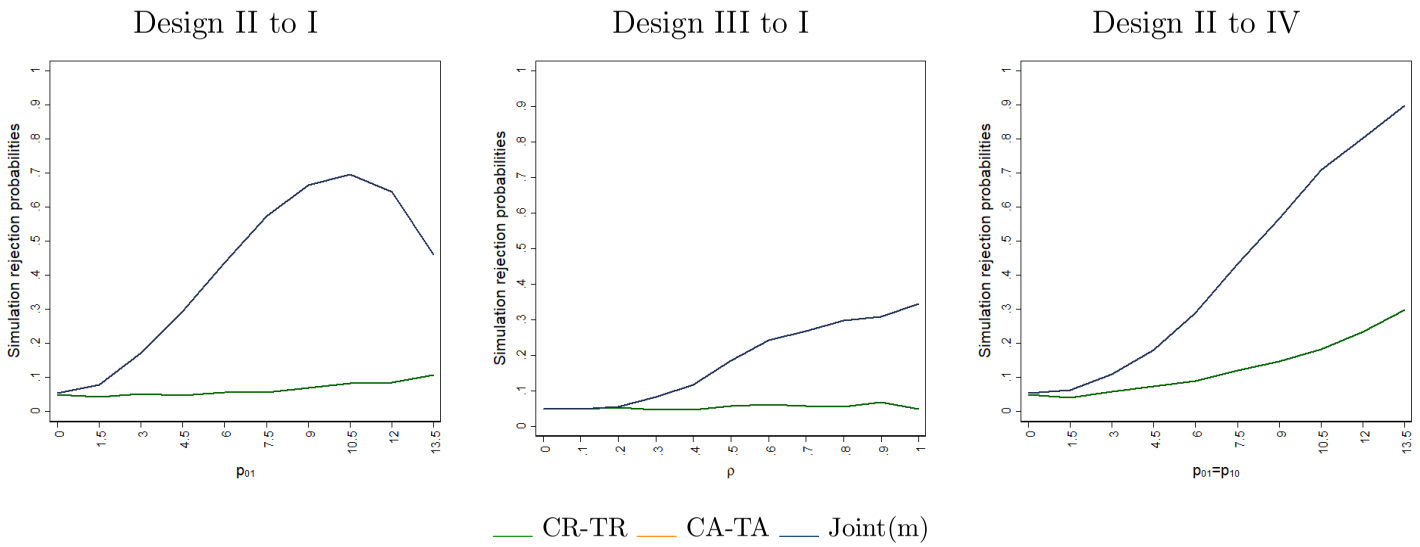
To illustrate the relative power properties of the tests that we propose, we present the simulation rejection probabilities for alternative parameter values of the designs we consider in Section 5. *Design II to I* depicts the case in which we vary the proportion of treatment-only responders, p_{01}, from zero to $0.9 \times P(R_i = 0|T_i = 0)$, where $p_{01} = 0$ corresponds to Design II and $p_{01} > 0$ to variants of Design I. *Design III to I* depicts the case in which we vary the correlation parameter between the unobservables in the outcome equation and the unobservables in the response equation, ρ, from zero to one. Hence, $\rho = 0$ corresponds to Design III while $\rho > 0$ corresponds to different versions of Design I. Finally, the results under *Design II to IV* are obtained by fixing $p_{01} = p_{10}$ and varying them from zero to $0.9 \times P(R_i = 0|Ti = 0)$. Design II corresponds to the case in which $p_{01} = p_{10} = 0$ and $p_{01} = p_{10} > 0$ corresponds to different versions of Design IV.

Overall, the simulation results illustrate that the *joint* tests that we propose in Section 4 have better finite-sample power properties relative to the statistics of the simple null hypotheses. Most notably, the results under *Design II to I* in Panel A of Figure 5 show that when internal validity for respondents does not hold (i.e. $p_{01} > 0$), the simulation rejection probabilities of the joint test are generally above the simulation rejection probabilities of the simple test that only uses the subpopulation of respondents.

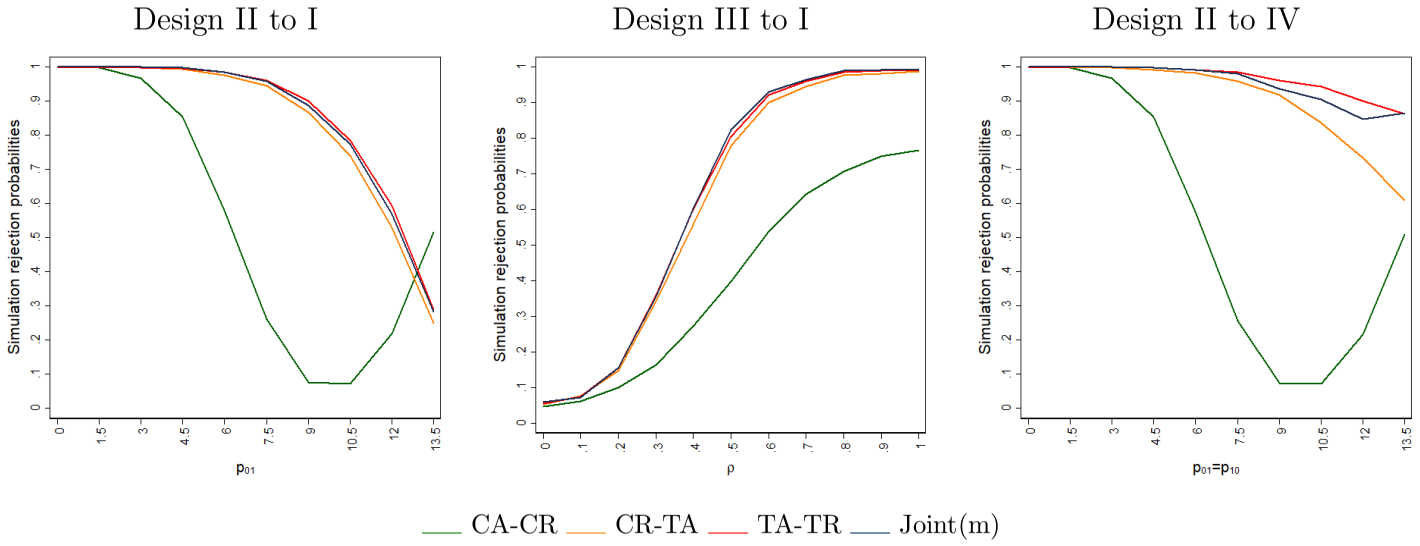
[42] We use an attrition rate of 15% in the control group as reference since that is the average attrition rate in our review of field experiments. See Section 2 for details.

Figure 5: Simulation Rejection Probabilities of the *CM* Statistic of the Distributional Tests of Internal Validity

Panel A. Internal Validity for Respondents



Panel B. Internal Validity for the Study Population



D Regression Tests of Internal Validity

In this section, we show how to implement regression-based tests of internal validity for respondents ($H^1_{0,\mathcal{M}}$) and internal validity for the study population ($H^2_{0,\mathcal{M}}$). We follow the same notational conventions as in the paper.

D.1 Completely and Clustered Randomized Experiments

$$
\begin{aligned}
Y_{i0} &= \gamma_{11} T_i R_i + \gamma_{01}(1-T_i)R_i + \gamma_{10} T_i (1-R_i) + \gamma_{00}(1-T_i)(1-R_i) + \epsilon_i \\
H^1_{0,\mathcal{M}} &: \gamma_{11} = \gamma_{01} \;\&\; \gamma_{10} = \gamma_{00}, \\
H^2_{0,\mathcal{M}} &: \gamma_{11} = \gamma_{01} = \gamma_{10} = \gamma_{00}.
\end{aligned}
$$

Both hypotheses are joint hypotheses of linear restrictions on linear regression coefficients. Hence, they are straightforward to test using the appropriate standard errors.

It is worth noting that since treatment is randomly assigned, the baseline outcome of the treatment and control groups are identically distributed by definition. Hence, an intuitive testable restriction of internal validity for the study population is that the baseline outcome has the same mean across attritors and respondents within treatment and control groups, i.e. $H^{2'}_{0,\mathcal{M}} : \gamma_{11} = \gamma_{10} \;\&\; \gamma_{01} = \gamma_{00}$. This restriction is weaker than $H^2_{0,\mathcal{M}}$, hence there may be consequences for finite-sample power.

D.2 Stratified Randomized Experiments

$$
Y_{i0} = \sum_{s \in \mathcal{S}} \left[\gamma^s_{11} T_i R_i + \gamma^s_{10} T_i (1-R_i) + \gamma^s_{01}(1-T_i)R_i + \gamma^s_{00}(1-T_i)(1-R_i) \right] 1\{S_i = s\} + \epsilon_i
$$

Hence, for $s \in \mathcal{S}$,

$$
\begin{aligned}
H^{1,s}_{0,\mathcal{M}} &: \gamma^s_{11} = \gamma^s_{01} \;\&\; \gamma^s_{10} = \gamma^s_{00}, \\
H^{2,s}_{0,\mathcal{M}} &: \gamma^s_{11} = \gamma^s_{01} = \gamma^s_{10} = \gamma^s_{00}.
\end{aligned}
$$

One could either test the above null hypotheses jointly for all $s \in \mathcal{S}$ or approach it as a multiple testing problem for each $s \in \mathcal{S}$ and perform an appropriate correction.

References

Abadie, Alberto, Matthew M. Chingos, and Martin R. West, "Endogenous Stratification in Randomized Experiments," *The Review of Economics and Statistics*, 2018, *100* (4),

567–580.

Ahn, Hyungtaik and James L. Powell, "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 1993, *58* (1), 3–29.

Altonji, Joseph and Rosa Matzkin, "Cross-section and Panel Data Estimators for Nonseparable Models with Endogenous Regrssors," *Econometrica*, 2005, *73* (3), 1053–1102.

Andrews, Isaiah and Emily Oster, "Weighting for External Validity," 2017. Unpublished Manuscript.

Angrist, Joshua D., "Conditional independence in sample selection models," *Economics Letters*, 1997, *54* (2), 103 – 112.

_ , Guido W. Imbens, and Donald B. Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 1996, *91* (434), 444–455.

Athey, S. and G.W. Imbens, "Chapter 3 - The Econometrics of Randomized Experimentsa," in Abhijit Vinayak Banerjee and Esther Duflo, eds., *Handbook of Field Experiments*, Vol. 1 of *Handbook of Economic Field Experiments*, North-Holland, 2017, pp. 73 – 140.

Athey, Susan, Dean Eckles, and Guido W. Imbens, "Exact p-Values for Network Interference," *Journal of the American Statistical Association*, 2017, *0* (0), 1–11.

Azzam, Tarek, Michael Bates, and David Fairris, "Do Learning Communities Increase First Year College Retention? Testing the External Validity of Randomized Control Trials," 2018. Unpublished Manuscript.

Baird, Sarah, J. Aislinn Bohren, Craig McIntosh, and Berk Özler, "Optimal Design of Experiments in the Presence of Interference," *The Review of Economics and Statistics*, 2018, *Forthcoming.*

Behagel, Luc, Bruno Crépon, Marc Gurgand, and Thomas Le Barbanchon, "Please Call Again: Correcting Nonresponse Bias in Treatment Effect Models," *The Review of Economics and Statistics*, 2015, *97*, 1070–1080.

Bester, C. Alan and Christian Hansen, "Identification of Marginal Effects in a Nonparametric Correlated Random Effects Model," *Journal of Business and Economic Statistics*, 2009, *27* (2), 235–250.

Brownstone, David, "Multiple Imputation Methodology For Missing Data, Non-Random Response, And Panel Attrition.," in T. Gärling, T. Laitila, and K. Westin, eds., *Theoretical Foundations of Travel Choice Modeling*, Elsevier, 1998, pp. 421–450.

Bruhn, Miriam and David McKenzie, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, October 2009, *1* (4), 200–232.

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh, "Inference under Covariate-Adaptive Randomization," *Journal of the American Statistical Association*, 2017, *0* (ja), 0–0.

Canay, Ivan A., Joseph P. Romano, and Azeem M. Shaikh, "Randomization Tests Under an Approximate Symmetry Assumption," *Econometrica*, 2017, *85* (3), 1013–1030.

Chernozhukov, Victor, Ivan Fernandez-Val, Jinyong Hahn, and Whitney Newey, "Average and Quantile Effects in Nonseparable Panel Data Models," *Econometrica*, 2013, *81* (2), pp.535–580.

Das, Mitali, Whitney K. Newey, and Francis Vella, "Nonparametric Estimation of Sample Selection Models," *The Review of Economic Studies*, 2003, *70* (1), 33–58.

de Chaisemartin, Clément and Luc Behagel, "Next Please! Estimating the Effect of Treatments Allocated by Randomized Waiting Lists," 2017. Working Paper.

de Chaisemartin, Clément, "Tolerating defiance? Local average treatment effects without monotonicity," *Quantitative Economics*, 2017, *8* (2), 367–396.

Dufour, Jean-Marie, "Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics," *Journal of Econometrics*, 2006, *133* (2), 443 – 477.

_ , Abdeljelil Farhat, Lucien Gardiol, and Lynda Khalaf, "Simulation-Based Finite Sample Normality Tests of Linear Regressions," *The Econometrics Journal*, 2008, *1* (1).

Ghanem, Dalia, "Testing Identifying Assumptions in Nonseparable Panel Data Models," *Journal of Econometrics*, 2017, *197*, 202–217.

Glennerster, Rachel and Kudzai Takavarasha, *Running Randomized Evaluations: A Practical Guide*, student edition ed., Princeton University Press, 2013.

Heckman, James J., "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models,"

in "Annals of Economic and Social Measurement, Volume 5, number 4" NBER Chapters, National Bureau of Economic Research, Inc, April 1976, pp. 475–492.

_ , "Sample Selection Bias as a Specification Error," *Econometrica*, 1979, *47* (1), 153–161.

Hoderlein, Stefan and Halbert White, "Nonparametric Identification of Nonseparable Panel Data Models with Generalized Fixed Effects," *Journal of Econometrics*, 2012, *168* (2), 300–314.

Horowitz, Joel L. and Charles F. Manski, "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 2000, *95* (449), 77–84.

Hsu, Yu-Chin, Chu-An Liu, and Xiaoxia Shi, "Testing Generalized Regression Monotonicity," 2016. Unpublished Manuscript.

Imbens, Guido W. and Donald B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press, 2015.

_ **and Joshua D. Angrist**, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

Kitagawa, Toru, "A Test for Instrument Validity," *Econometrica*, 2015, *83* (5), 2043–2063.

Kline, Patrick and Andres Santos, "Sensitivity to missing data assumptions: Theory and an evaluation of the U.S. wage structure," *Quantitative Economics*, 2013, *4* (2), 231–267.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 2009, *76* (3), 1071–1102.

Lehmann, E. L. and Joseph P. Romano, *Testing statistical hypotheses* Springer Texts in Statistics, third ed., New York: Springer, 2005.

Manski, Charles F., "Partial identification with missing data: concepts and findings," *International Journal of Approximate Reasoning*, 2005, *39* (2), 151 – 165. Imprecise Probabilities and Their Applications.

McKenzie, David, "Beyond baseline and follow-up: The case for more T in experiments," *Journal of Development Economics*, 2012, *99* (2), 210–221.

Millán, Teresa Molina and Karen Macours, "Attrition in Randomized Control Trials: Using Tracking Information to Correct Bias," 2017. Working Paper.

Mourifié, Ismael and Yuanyuan Wan, "Testing Local Average Treatment Effect Assumptions," *The Review of Economics and Statistics*, 2017, *99* (2), 305–313.

Vella, Francis, "Estimating Models with Sample Selection Bias: A Survey," *The Journal of Human Resources*, 1998, *33* (1), 127–169.

Vytlacil, Edward, "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 2002, *70* (1), 331–341.

Wooldridge, Jeffrey M., "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, 2007, *141* (2), 1281 – 1301.

WWC, *What Works Clearinghouse: Standards Handbook, Version 4.0, 7/25/2017 Draft* 2017.

Young, Alwyn, "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*," 11 2018.