Testing Attrition Bias in Field Experiments*

Dalia Ghanem
UC Davis

Sarojini Hirshleifer
UC Riverside

Karen Ortiz-Becerra
UC Davis

March 29, 2020

Abstract

We approach attrition in field experiments with baseline outcome data as an identification problem in a panel model. A systematic review of the literature indicates that there is no consensus on how to test for attrition bias. We establish identifying assumptions for treatment effects for both the respondent subpopulation and the study population. We then derive their sharp testable implications on the baseline outcome distribution and propose randomization procedures to test them. We demonstrate that the most commonly used test does not control size in general when internal validity holds. Simulations and applications illustrate the empirical relevance of our analysis.

JEL Codes: C12, C21, C23, C93
Keywords: attrition, non-response, treatment effects, field experiment, randomized experiment, randomized control trial, internal validity, identifying assumptions, randomization test, panel data

*E-mail: dghanem@ucdavis.edu, sarojini.hirshleifer@ucr.edu, kaortizb@ucdavis.edu.
We thank Alberto Abadie, Josh Angrist, Stephen Boucher, Federico Bugni, Pamela Jakiela, Tae-hwy Lee, Jia Li, Aprajit Mahajan, Matthew Masten, Craig McIntosh, David McKenzie, Adam Rosen, Monica Singhal and Aman Ullah for helpful discussions.

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. In high-income countries, survey response rates are often low and may be declining.[2] 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 (ATE-R)? Second, is this estimand equal to the average treatment effect for the study population (ATE)?[3] To answer these questions, we examine the testable implications of the relevant identifying assumptions and propose procedures to test them. Our results provide insights that are relevant to current empirical practice.

We first conduct a systematic review of 91 recent field experiments with baseline data in order to document attrition rates and understand how authors test for attrition bias. Attrition and attrition tests are both common in published field experiments. Although the implementation of attrition tests varies widely, 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 observable characteristics differs across the treatment and control groups conditional on response status. While authors report a differential attrition rate test for 81% of field experiments, they report a selective attrition test only 60% of the time. In addition, for a substantial minority of field experiments (34%), authors conduct a *determinants of attrition test* for differences in the distributions of respondents and attritors.

Next, we present a formal treatment of attrition in field experiments with baseline out-

[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] See, for example, Meyer et al. (2015) and Barrett et al. (2014).

[3] We refer to the population selected for the evaluation as the study population.

come 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 (IV-R), the identifying assumption is random assignment conditional on response status (IV-R assumption). This implies that the difference in the mean outcome across the treatment and control respondents identifies the ATE-R, a local average treatment effect for the respondents. In the second case, where internal validity for the study population (IV-P) is of interest, the identifying assumption is that the unobservables that affect response and outcome are independent in addition to the initial random assignment of the treatment (IV-P assumption). If this identifying assumption holds, the ATE for the study population is identified. 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 IV-R assumption 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 IV-P assumption implies a joint hypothesis of equality on the baseline outcome distribution across all four treatment/response subgroups. Like all tests of identifying assumptions, a test of attrition bias can only be tested by implication in general. Hence, we show that the aforementioned testable restrictions are sharp, meaning that they are the strongest implications that we can test given our data.[4] We apply our two proposed tests to data from a large-scale RCT of the *Progresa* program in Mexico, in which the study population is representative of a broader population of interest. Across two main outcomes collected in the same survey, we reject the IV-P identifying assumption for one outcome while not rejecting it for another.

Since the IV-R and IV-P assumptions are random-assignment-type restrictions, randomization tests are a natural choice in this context.[5] We therefore propose "subgroup"-randomization procedures (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 for attrition bias given stratified randomization and to identify heterogeneous treatment effects.

Given their relevance to current empirical practice, we also provide a formal treatment of

[4]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 and applied in Kitagawa (2015), Hsu et al. (2019), and Mourifié and Wan (2017).

[5]The mean versions of our sharp testable restrictions for both the IV-R and IV-P identifying assumptions can be implemented using simple regression tests which we outline in Section B.

the differential attrition rate test and the use of covariates. In order to understand the role of differential attrition rates for internal validity, we apply the framework of partial compliance from the local average treatment effect (LATE) literature to potential response.[6] We demonstrate that even though equal attrition rates are sufficient for IV-R under additional assumptions, they are not a necessary condition for internal validity in general. We illustrate using an analytical example and simulations that it is possible to have differences in attrition rates across treatment and control groups while IV-P holds. Next, we examine the use of covariates in testing the IV-R or IV-P assumption, which is useful for settings where baseline outcome data is not available. We note two types of covariates may be included: (i) determinants of the outcome, and (ii) "proxy" variables which are determined by the same variables as the outcome in question. Using covariates that do not fulfill either of these criteria can lead to a false rejection of the IV-R or IV-P assumption.

To illustrate the empirical relevance of our results, we apply our tests of the IV-R and IV-P assumptions to outcomes from five published field experiments in our review with available data and the highest overall attrition rates. We also consider the authors' attrition tests and note that their approach differs from ours in several ways. Using our tests, we do not reject the IV-R assumption for any of the outcomes we examine, even though two of the experiments did not conduct a selective attrition test. More surprisingly, for about two thirds of the outcomes we examine, we cannot reject the IV-P assumption. We also find several empirical examples consistent with the theoretical conditions under which the differential attrition rate test does not control size, thereby providing evidence of their empirical relevance. Overall, our empirical results are promising for field experiments where IV-P is of interest.

This paper has several implications for empirical practice. First, our theoretical and empirical results imply that the most widely used test in the literature, the differential attrition rate test, *may* lead to a false rejection of internal validity in practice. The second most widely used test, the selective attrition test, is implemented using a variety of approaches, the majority of which focus on IV-R and only use respondents. Our theoretical results indicate, however, that the implication of the relevant identifying assumption is a joint test that uses all of the available information in the baseline data, and thus includes both respondents and attritors. Finally, while the majority of testing procedures pertain to IV-R and not IV-P, the use of determinants of attrition tests suggests that some researchers may be interested in implications of the estimated treatment effects for the study population. More generally, this paper highlights the importance of understanding the implications of attrition for a broader population when interpreting field experiment results for policy.[7]

[6] See the foundational work in the LATE literature (Imbens and Angrist, 1994; Angrist et al., 1996).

[7] External validity can be assessed in a number of ways (see, for example, Andrews and Oster (2019) and

This paper contributes to a growing literature that considers methodological questions relevant to field experiments.[8] Given the wide use of attrition tests, we formally examine the testing problem here. Our focus complements a thread in this literature that outlines various approaches to correcting attrition bias in field experiments (Lee, 2009; Huber, 2012; Behagel et al., 2015; Millán and Macours, 2019).[9] These corrections build on the vast sample selection literature in econometrics going back to Heckman (1976, 1979).[10] While the larger sample selection literature is broadly concerned with population objects, work relevant for program evaluation propose corrections for objects pertaining to subpopulations (e.g. Lee, 2009; Huber, 2012; Chen and Flores, 2015). Our paper provides tests of identifying assumptions emphasizing the distinction between the (study) population and the respondent subpopulation. Finally, the randomization tests we propose contribute to recent work that examines the potential use of randomization tests in analyzing field experiment data (Young, 2018; Athey and Imbens, 2017; Athey et al., 2018; Bugni et al., 2018).

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 in this paper fall under the nonparametric correlated random effects category (Altonji and Matzkin, 2005). Furthermore, we build on the literature on randomization tests for distributional statistics (Dufour, 2006; Dufour et al., 1998).

The paper proceeds as follows. Section 2 presents the review of the field experiment literature. Section 3 formally presents the identifying assumptions and their sharp testable restrictions. It also includes a formal treatment of differential attrition rates and of the role

Azzam et al. (2018)). In our setting, we note that if IV-R holds but not IV-P, we may be able to draw inference from the local average treatment effect for respondents to a broader population.

[8] 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 Behaghel (2018) 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.

[9] Other work considers corrections for settings with sample selection and noncompliance. Chen and Flores (2015) rely on monotonicity restrictions to construct bounds for average treatment effects in the presence of partial compliance and sample selection. Fricke et al. (2015) consider instrumental variables approaches to address these two identification problems.

[10] Nonparametric Heckman-style corrections have been proposed for linear and nonparametric outcome models (e.g. Ahn and Powell, 1993; Das et al., 2003). Inverse probability weighting (Horvitz and Thompson, 1952; Hirano et al., 2003; Robins et al., 1994) is another important category of corrections for sample selection bias, frequently used in the field experiment literature. Attrition corrections for panel data have also been proposed (e.g. Hausman and Wise, 1979; Wooldridge, 1995; Hirano et al., 2001). Finally, nonparametric bounds is an alternative approach relying on weaker conditions (Horowitz and Manski, 2000; Manski, 2005; Lee, 2009; Kline and Santos, 2013).

of covariates in testing internal validity. In Section 4, we propose a subgroup-randomization procedure to obtain p-values for the distributional null hypotheses. Section 5 presents simulation experiments to illustrate the theoretical results. Section 6 presents the results of the empirical application exercise. Section 7 concludes.

2 Attrition in the Field Experiment Literature

We systematically reviewed 88 recent articles published in economics journals that report the results of 91 field experiments. The objective of this review is to understand both the extent to which attrition is observed and the implementation of tests for attrition bias in the literature.[11] Our categorization imposes some structure on the variety of different estimation strategies used to test for attrition bias in the literature.[12] In keeping with our panel approach, we focus on field experiments in which the authors had baseline data on at least one main outcome variable.[13]

We review reported overall and differential attrition rates in field experiment papers and find that attrition is common. As depicted in Panel A in Figure 1, 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 the average of 15%.[14] 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

[11] We included articles from 2009 to 2015 that were published in 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*. Section A.1 in the online appendix includes additional details on the selection of papers and relevant attrition rates. Section F in the online appendix contains a list of all the papers included in the review.

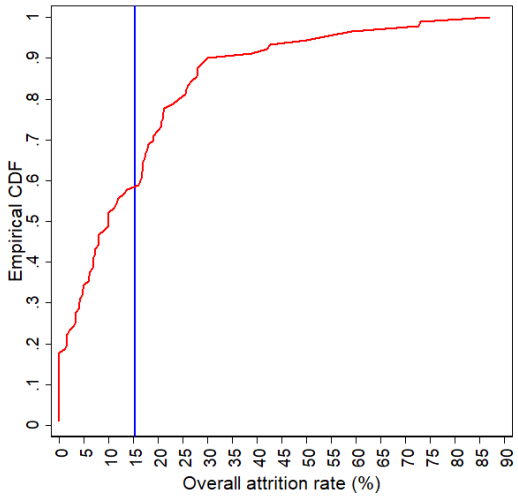
[12] We identify fifteen estimation strategies used to conduct attrition tests (see Section B in the online appendix).

[13] We exclude 58 field experiments that were published during that time period, since they lack baseline data for any outcome mentioned in the abstract. Of those, slightly less than half (45%) are experiments for which the baseline outcome is the same for everyone by design and hence is not informative (see Section A.1 in the online appendix).

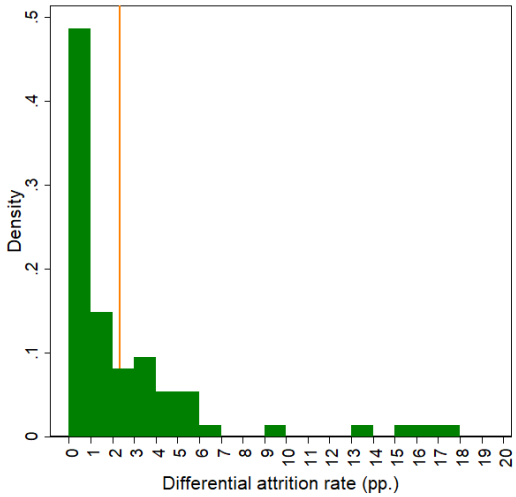
[14] To understand the extent of attrition that is relevant to the main outcomes in the paper, we focus on attrition rates that are relevant to outcomes reported in the abstract (i.e. "abstract results"). Most papers report attrition rates at the level of the data source or subsample, rather than at the level of the outcome. Since the number of data sources and/or subsamples that are relevant to the abstract results vary by experiment, we include one attrition rate per field experiment for consistency. Specifically, we report the highest attrition rate relevant to an abstract result. Authors do not in general report attrition rates conditional on baseline response. A noteworthy finding from Table B.1 in the online appendix is that attrition rates are higher on average for experiments in high-income countries.

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 a result reported in the abstract of the article. The *Overall* rate is the attrition rate for the full sample, which is composed of 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 (orange) line depicts the average overall (differential) attrition rate in our sample of field experiments. Panel A includes 90 field experiments and Panel B includes 74 experiments since the relevant attrition rates are not reported in some articles.

differential attrition rate that is greater than 5 percentage points.[15]

We then study how authors test for attrition bias. Notably, attrition tests are widely used in the literature: 90% of field experiments with an attrition rate of at least 1% for an outcome with baseline data conduct at least one attrition test. We first 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* determines whether the rates of attrition are statistically significantly different across treatment and control groups. In contrast, a *selective attrition test* aims to determine whether, conditional on being a respondent and/or attritor, the mean of observable characteristics is the same across treatment and control groups. We find that there is no consensus on whether to conduct a differential attrition rate test or a selective attrition test, however (Panel A in Table 1). In the field experiments that we reviewed, the differential attrition rate test is substantially more common (81%) than the selective attrition test (60%). In fact, 30% of the articles that conducted a differential attrition rate do not conduct a selective attrition

[15]It is possible, however, that these numbers reflect authors' exclusion of results with higher differential attrition rates than those that were reported or published.

test.[16]

Table 1: Distribution of Field Experiments by Attrition Test

Panel A: Differential and Selective Attrition Tests

Proportion of field experiments that conduct:		Selective attrition test		
		No	*Yes*	*Total*
Differential attrition rate test	*No*	10%	10%	19%
	Yes	30%	51%	81%
	Total	40%	60%	100%

Panel B: Types of Selective Attrition Test

Conditional on conducting a selective attrition test:	
Test using respondents and attritors	27%
Test using respondents only	68%
Test using attritors only	5%
Total†	100%

Panel C: Determinants of Attrition Tests

Proportion of field experiments that conduct:	Determinants of attrition test		
	Yes	*No*	*Total*
Differential attrition rate test only	11%	19%	30%
Selective attrition test only	1%	8%	10%
Differential & selective attrition tests	22%	29%	51%
No differential & no selective attrition test	0%	10%	10%
Total	34%	66%	100%

Notes: Panel A and C include 73 field experiments that have an attrition rate of at least 1% for an outcome with baseline data. Panel B includes 44 of those experiments that conducted a selective attrition test (†). For details on the classification of the empirical strategies, see Section B in the online appendix.

We further consider if selective attrition tests include both the respondents and the attritors or if they include either only the respondents or only the attritors (Panel B in Table 1). Conditional on having conducted any type of selective attrition test, authors include both respondents and attritors in only 27% of those field experiments. Instead, authors conduct a selective attrition test on the sample of respondents in most cases (68%). Although our review is limited to experiments in which baseline outcome data is available, covariates are typically included in attrition tests along with the baseline outcome. In particular, 98% of field experiments that report a selective attrition test include more than one baseline variable

[16]We also consider some potential determinants of the use of selective attrition tests: overall attrition rates, differential rates, year of publication, journal of publication. We do not find any strong correlations given the available data.

in that test.[17] A key issue that arises with the inclusion of covariates is how to approach the issue of multiple testing. We find that 77% of the experiments that implement a selective attrition test conduct it on an average of 16 variables, and none of those implement a multiple testing correction (Table B.2 in online appendix). Only a minority of authors conduct a joint test across all of the baseline variables included in the test (21%).

Another important aspect of testing for attrition bias is testing for differences in the distributions of respondents and attritors. Such tests can illustrate the implications of the main results of the experiment for the study population. We define a *determinants of attrition test* as a test of whether baseline outcomes and covariates correlate with response status and find that authors conduct such a test in approximately one-third of field experiments (Panel C of Table 1). Table 1 illustrates that conducting the determinants of attrition test does not have a one-to-one relationship with either conducting a differential attrition rate test or conducting a selective attrition test.[18]

3 Identifying Treatment Effects in the Presence of Attrition

This section presents 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 data is available for both completely and stratified randomized experiments. We further examine the role of differential attrition rates in this context and discuss the implications of our theoretical analysis for empirical practice.

3.1 Internal Validity in the Presence of Attrition

An empirical example motivates our treatment of internal validity in the presence of attrition. After deriving the implications of our identifying assumptions, we demonstrate how to test those implications in that example. We also consider the limits of testing identifying assumptions and present the extension of the results to stratified randomization and heterogeneous treatment effects.

[17] Although identifying which variables are outcomes or covariates is beyond the scope of this paper, we note that in 91% of the experiments the selective attrition test includes at least one variable that we can easily identify as a covariate (such as age or gender).

[18] Approximately half of the determinants of attrition tests are conducted using the same regression used to test for differential attrition rates. We categorize this strategy as both types of tests since authors typically interpret both the coefficients on treatment and the baseline covariates.

3.1.1 Motivating Example

To illustrate the problem of attrition in field experiments, we use data collected for the randomized evaluation of *Progresa*, a social program in Mexico that provides cash to eligible poor households on the condition that children attend school and family members visit health centers regularly (Skoufias, 2005). The evaluation of *Progresa* relied on the random assignment of 320 localities into the treatment group and 186 localities into the control group. These localities, which constitute the study population, were selected to be representative of a larger population of 6396 eligible localities across seven states in Mexico.[19] The surveys conducted for the experiment include a baseline and three follow-up rounds collected 5, 13, and 18 months after the program began.[20] We examine two outcomes of the evaluation that have been previously studied: (i) current school enrollment for children 6 to 16 years old, and (ii) paid employment for adults in the last week.

Table 2: Summary Statistics for the Outcomes of Interest for *Progresa*

	Full Sample				Respondent Subsample at Follow-up			
Round	N	Control Mean	$T-C$	p-value	Attrition Rate	Control Mean	$T-C$	p-value
Panel A. School Enrollment (6-16 years old)								
Baseline	24353	0.824	0.007	0.455				
Pooled					0.183	0.793	0.046	0.000
1st					0.142	0.814	0.043	0.000
2nd					0.234	0.829	0.046	0.000
3rd					0.174	0.740	0.047	0.000
Panel B. Employment Last Week (18+ years old)								
Baseline	31237	0.471	-0.006	0.546				
Pooled					0.161	0.464	0.014	0.002
1st					0.096	0.460	0.016	0.016
2nd					0.196	0.459	0.009	0.138
3rd					0.192	0.472	0.018	0.001

Notes: T and C refer to treatment and control group, respectively. $T-C$ is the difference in means between the treatment and control groups and the p-value is estimated with a regression of outcome on treatment that clusters standard errors at the locality level. The attrition rates reported are conditional on responding to the baseline survey. *Pooled* refers to data from all three follow-ups combined.

In Table 2, we report the initial sample size for each outcome of interest as well as summary statistics of the outcome by treatment group at baseline and follow-up. The failure to reject the null hypothesis of the equality of means across the treatment and control groups at baseline is suggestive evidence that the randomization procedure was implemented as

[19] Localities were eligible if they ranked high on an index of deprivation, had access to schools and a clinic, and had a population of 50 to 2500 people. See INSP (2005) for details about the experiment. For this analysis, we use the evaluation panel dataset, which can be found on the official website of the evaluation at `https://evaluacion.prospera.gob.mx/es/eval_cuant/p_bases_cuanti.php`.

[20] The baseline was collected in October 1997 and the three follow-ups were collected in October 1998, June 1999, and November 1999.

intended. In the context of treatment randomization and absence of attrition, differences in a mean outcome across treatment and control groups at follow-up would identify the average treatment effect of *Progresa* for the study population. Pooling data from the three follow-up rounds, we would conclude that the impact of *Progesa* on the probability that children attend school (adults work) is an increase of 4.6 (1.4) percentage points. The attrition rate, however, varies from 10% to 24% depending on the outcome and the follow-up round. These attrition rates raise the question of whether the differences in mean outcomes in respondents identify at least one of two objects of interest: (i) the average treatment for the respondent subpopulation (ATE-R) or (ii) the average treatment effect for the entire study population (ATE).

3.1.2 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, where $D_{it} \in \{0, 1\}$.[21] 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 individual i belongs to the treatment group and $T_i = 0$ if 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 $d_U \times 1$ vector of time-invariant and time-varying variables, 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 $d = 0, 1$.[22] To simplify illustration, we postpone the discussion of covariates to Section 3.3.2.

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 a valid counterfactual outcome distribution for the treatment group, i.e. $Y_{i1}(0)|T_i = 1 \stackrel{d}{=} Y_{i1}|T_i = 0$, where $\stackrel{d}{=}$ denotes the equality in distribution. In this case, any difference in the outcome distribution

[21] The extension to the multiple treatment case is in Section D of the online appendix.

[22] 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 understanding internal validity questions in our problem.

between treatment and control groups in the follow-up period can be attributed to the treatment. The ATE can be identified as the difference in mean outcomes between the treatment and control group,

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

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 ($t = 1$) as in Hirano et al. (2001). Response status in the follow-up period is determined by the following equation,[23]

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

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 outcome 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 at follow-up, 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, $Y_{i1}|T_i = 0, R_i = 1 \stackrel{d}{=} Y_{i1}(0)|T_i = 1, R_i = 1$? This would imply that we can obtain internally valid estimands for the respondent subpopulation, such as the ATE-R, $E[Y_{i1}(1) - Y_{i1}(0)|R_i = 1]$. Second, do the outcome distributions of treatment and control respondents in the follow-up period identify the potential outcome distribution of the study population with and without the treatment, $Y_{i1}|T_i = \tau, R_i = 1 \stackrel{d}{=} Y_{i1}(\tau)$ for $\tau = 0, 1$? This would imply that we can obtain internally valid estimands for the study population, such as the ATE.

The next proposition 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*

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

(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 A. The assumption in (a) is random assignment conditional on response status. 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* (IV-R) and the assumption in (a) as the IV-R assumption. The restriction in (a.ii) implies that the appropriate test of the implication of the IV-R assumption is a *joint test* of the equality of the baseline outcome distribution between treatment and control respondents as well as treatment and control attritors.[24]

The assumption in (b) implies *missing-at-random* as defined in Manski (2005).[25] Together with random assignment, it implies that 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* (IV-P) and the assumption in (b) as the IV-P assumption. 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. The restriction in (b.ii) is the testable implication of the IV-P assumption under random assignment. The resulting null hypothesis

[24] If IV-R is of interest, a natural question is whether one should simply test the implication of $(U_{i0}, U_{i1}) \perp T_i|R_i = 1$ in lieu of the IV-R assumption ($(U_{i0}, U_{i1}) \perp T_i|R_i$). This would be empirically relevant 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, we note that a primitive condition for this 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 condition is not empirically plausible since it implies that the unobservable distribution is the same for always-responders, treatment-only and control-only responders, but different for the never-responders.

[25] In the cross-sectional setup, the missing-at-random assumption is given by $Y_i|T_i, R_i \stackrel{d}{=} Y_i|T_i$. Manski (2005) establishes that this assumption is not testable in that context. We obtain the testable implications by exploiting the panel structure. It is important to emphasize that this definition of missing-at-random is different from the assumption in Hirano et al. (2001) building on Rubin (1976), which would translate to $Y_{i1} \perp R_i|Y_{i0}, T_i$ in our notation. Finally, while we do not distinguish between observables and unobservables here, it is worth noting that Assumption 3 in Huber (2012) provides a set of conditions that imply the assumption in Proposition 1(b).

in this case is the equality of the baseline outcome distribution regardless of both treatment and response status.

3.1.3 Application of Tests to Motivating Example

Returning to our motivating example from the Progresa evaluation, we aim to understand whether the differences in mean outcomes across treatment and control respondents at follow-up reported in Table 2 are estimating an internally valid object, such as the ATE-R or the ATE. We do so by testing the implications of the relevant identifying assumptions. Since both outcomes in our example are binary, the restrictions in Proposition 1 simplify to restrictions on the baseline mean for each outcome across the four treatment-response subgroups.

We first inspect the mean baseline outcome across the four subgroups presented in Table 3 and notice distinct patterns across the two outcomes of interest. The share of children who attend school at baseline is similar across treatment and control respondents as well as treatment and control attritors. This is consistent with the testable restriction in Proposition 1(a.ii) implied by the IV-R assumption, which is random assignment conditional on response status. When we compare respondents and attritors, however, we find meaningful differences. At baseline, school enrollment for the respondents in the pooled follow-up sample was around 87%, while enrollment for the attritors in the same sample was 61%. Thus, children that are observed in the follow-up data are substantially different from those that are not. This suggests a violation of the testable restriction of the IV-P assumption in Proposition 1(b.ii), which requires all four treatment-response subgroups to have the same mean outcome at baseline. In contrast, the share of employed adults at baseline is similar in all four subgroups, which is consistent with the testable implication of the IV-P assumption.

Table 3 also presents the p-values of the tests of the IV-R and IV-P assumptions based on the restrictions in Proposition 1(a.ii) and (b.ii), respectively. For school enrollment, we specifically cannot reject the IV-R assumption, but we do reject the IV-P assumption at the 5% significance level.[26] Thus, we do not reject the assumption that the difference in school attendance rates across treatment and control respondents at follow-up identifies the ATE-R. We do, however, reject the assumption that this difference could identify the ATE. In contrast, for the outcome of employment, we do not reject either the IV-R or the IV-P assumption.[27] In other words, we do not reject the assumption that the difference

[26] It is worth noting that a multiple testing correction would not change the decisions of any of the tests in our example. For instance, applying the Bonferroni correction for each outcome would yield a significance level for each hypothesis of 0.63% to control a family-wise error rate of 5% across the eight tests we conduct.

[27] A natural question that arises from this example is why we find different patterns of response across two outcomes that were collected from the same surveys. We conduct a determinants of attrition test, and find that the probability that a household responds to the employment question for all adults and does not

Table 3: Internal Validity in the Presence of Attrition for *Progresa*

Follow-up	Attrition Rate		Mean Baseline Outcome by Group				Test of IV-R	Test of IV-P
	C	Differential	TR	CR	TA	CA	*p*-value	*p*-value
Panel A. School Enrollment (6-16 years old)								
Pooled	0.187	-0.007	0.878	0.874	0.615	0.605	0.836	0.000
1st	0.150	-0.013	0.875	0.871	0.550	0.554	0.810	0.000
2nd	0.244	-0.017	0.901	0.897	0.590	0.595	0.824	0.000
3rd	0.168	0.009	0.859	0.856	0.697	0.663	0.217	0.000
Panel B. Employment Last Week (18+ years old)								
Pooled	0.157	0.007	0.463	0.468	0.472	0.486	0.698	0.132
1st	0.100	-0.007	0.464	0.471	0.472	0.473	0.825	0.860
2nd	0.195	0.001	0.463	0.465	0.474	0.496	0.566	0.058
3rd	0.175	0.027	0.463	0.469	0.471	0.481	0.769	0.503

Notes: The mean baseline outcomes correspond to the groups of treatment respondents (TR), control respondents (CR), treatment attritors (TA), and control attritors (CA). *Pooled* refers to all the three follow-ups. The tests of internal validity were conducted using the regression tests proposed in Section B. All regression tests use clustered standard errors at the locality level. For further details on the implementation of the tests, see Sections 4 and 6.

in employment rates between treatment and control respondents at follow-up identifies the ATE.

Understanding treatment effects for the study population is especially relevant to understanding the impact of large-scale programs such as *Progresa*, where the study population is representative of a larger population. In this type of study, if we do reject the IV-P assumption but not the IV-R assumption for an outcome such as school enrollment, we can still draw inferences about an average treatment effect on a larger population. That average treatment effect, however, is a local average treatment effect for the type of participants for which there would be follow-up data available for a given outcome.

3.1.4 Attrition Tests as Identification Tests

Like other tests of identifying assumptions, tests of internal validity in the presence of attrition can only be tested by implication in general. In our problem, if we impose time homogeneity on the structural function and the unobservable distribution (Chernozhukov et al., 2013), specifically $\mu_0 = \mu_1$ and $U_{i0}|T_i, R_i \stackrel{d}{=} U_{i1}|T_i, R_i$, then the testable restriction in Proposition 1(a.ii) holds if and only if identification (a.i) holds. This equivalence relationship does not hold in general, however. Hence, while rejection of a test of the implication in (a.ii) allows us to refute the identifying assumption in question, it is possible not to reject the test

respond to the school enrollment question for all children is positively correlated with household size, and is even more closely correlated with the number of children 6-16 years old in the household. This suggests that non-response on the school enrollment question may be driven by survey fatigue.

even when identification fails.[28] This point is illustrated in the following example.

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.*[29]

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

[28] While the converse (i.e. that identification holds while the testable implication on the baseline outcome distribution is violated) is *theoretically* possible, it is not an interesting case empirically. If a field experimentalist finds violations of the testable implication of the IV-R assumption, it is highly unlikely that he/she will discount this evidence and argue that identification of the ATE-R remains possible from a simple difference of mean outcomes between treatment and control respondents.

[29] We could however have a case in which the control respondents provide a valid counterfactual 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 an average treatment effect, $E[Y_{i1}(1) - Y_{i1}(0)|T_i = 1, R_i = 1] = E[U_{i1}^1 + \beta_1 U_{i1}^1 + \beta_2 U_{i1}^2 | T_i = 1, R_i = 1] - E[U_{i1}^1 | T_i = 1, R_i = 1] = E[Y_{i1}|T_i = 1, R_i = 1] - E[Y_{i1}|T_i = 0, R_i = 1]$, 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 longer internally valid for the entire respondent subpopulation, but for the smaller subpopulation of treatment respondents.

3.1.5 Heterogeneous Treatment Effects and Stratified Randomization

In this section, we extend our analysis to discuss heterogeneous treatment effects and stratified randomization. Heterogeneous treatment effects, more formally referred to as conditional average treatment effects (CATE), are of interest in many experiments. Stratified randomization is also common in empirical practice. Sometimes it is a necessity of the design, such as when the study is randomized within roll-out waves or locations. At other times, it is included in the experimental design with the aim of increasing precision and reducing bias of both average and heterogeneous treatment effects. The results in this section are relevant both for stratified randomized experiments and for completely randomized experiments that estimate heterogeneous treatment effects.[30]

In the following, let S_i denote the stratum of individual i which has support $\mathcal{S}$, where $|\mathcal{S}| < \infty$.[31] To exclude trivial strata, we assume that $P(S_i = s) > 0$ for all $s \in \mathcal{S}$ throughout the paper. In a stratified randomized experiment, random assignment is defined by $(U_{i0}, U_{i1}, V_i) \perp T_i | S_i$, whereas in a completely randomized experiment this conditional independence assumption holds as an implication of simple randomization ($(S_i, U_{i0}, U_{i1}, V_i) \perp T_i$). As a result, the following proposition applies to both completely and stratified randomized experiments.

Proposition 2. *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) $Y_{i1}|T_i = 0, S_i = s, R_i = 1 \overset{d}{=} Y_{i1}(0)|T_i = 1, S_i = s, R_i = 1$, *for* $s \in \mathcal{S}$.

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

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

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

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

[30] This framework can also be extended to test unconfoundedness assumptions, which motivate IPW-type attrition corrections (Huber, 2012), using baseline data. While interesting, this issue is outside the scope of the present paper.

[31] The finiteness of the number of strata motivates the finite-support assumption on $\mathcal{S}$. It is worth noting however that the results in the proposition hold for continuous conditioning variables as well.

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,

$$\begin{aligned} &E[Y_{i1}(1) - Y_{i1}(0)|T_i = 1, S_i = s, R_i = 1] \\ =&E[Y_{i1}|T_i = 1, S_i = s, R_i = 1] - E[Y_{i1}|T_i = 0, S_i = s, R_i = 1]. \text{ (CATE-R)} \end{aligned} \tag{4}$$

Alternatively, the ATE-R can then be identified by averaging over S_i, i.e. $\sum_{s\in\mathcal{S}} P(S_i = s|R_i = 1)\left(E[Y_{i1}|T_i = 1, S_i = s, R_i = 1] - E[Y_{i1}|T_i = 0, S_i = s, R_i = 1]\right)$. 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 the IV-R assumption in the case of block randomization. The results in part (b) of the proposition refer to IV-P in the context of block randomization. Thus, they are also conditional versions of the results in Proposition 1(b).

3.2 Differential Attrition Rates and Internal Validity

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. Since the differential attrition rate test is widely used, we examine the relationship between equal attrition rates and IV-R as well as IV-P.

In order to understand the role of differential attrition rates in testing IV-R, we use potential response to characterize different response types 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: 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 2, the treatment and control respondents and attritors are composed of different response types $(R_i(0), R_i(1))$.

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

$$\begin{aligned} &p_{00} \equiv P((R_i(0), R_i(1)) = (0, 0)), \quad p_{01} \equiv P((R_i(0), R_i(1)) = (0, 1)), \\ &p_{10} \equiv P((R_i(0), R_i(1)) = (1, 0)), \quad p_{11} \equiv P((R_i(0), R_i(1)) = (1, 1)). \end{aligned} \tag{5}$$

Figure 2: 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 ($R_i = 1$)	$(R_i(0), R_i(1)) = (1, 0)$ $(R_i(0), R_i(1)) = (1, 1)$	$(R_i(0), R_i(1)) = (0, 1)$ $(R_i(0), R_i(1)) = (1, 1)$

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

$$P(R_i = 0|T_i = 0) = p_{00} + p_{01}, \quad P(R_i = 0|T_i = 1) = p_{00} + p_{10}. \tag{6}$$

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, equal attrition rates occur if $p_{01} = p_{10}$.

Next, we illustrate the relationship between the differential attrition rates and the IV-R assumption (Proposition 1(a)), $(U_{i0}, U_{i1}) \perp T_i|R_i$. To do so, we express the distribution of unobservables, (U_{i0}, U_{i1}), for treatment and control respondents as a mixture of the distributions of response types $(R_i(0), R_i(1))$. We omit the analysis for attritors for brevity, since it is analoguous. Under random assignment, the unobservable distribution of treatment and control respondents is given by the following

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

When the IV-R assumption holds, the two distributions on the left hand side of the above equations agree. This equality holds in three different cases outlined in the following proposition.

Proposition 3. *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) $R_i(0) \leq R_i(1)$ *(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 Section A. Note that in (i) there are no restrictions on the attrition rates. This assumption requires that all four treatment-response subgroups have the same unobservable distribution, which not only implies IV-R, but also IV-P, under random assignment. In (ii), where both equal attrition rates and monotonicity are required for IV-R to hold, the respondent subpopulation is solely composed of always-responders ($(R_i(0), R_i(1)) = (1, 1)$). Lee (2009) uses the monotonicity assumption to bound the average treatment effect for the always-responders when attrition rates are not equal. The exchangeability restriction in (iii) merits some discussion. Specifically, 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 the IV-R assumption.

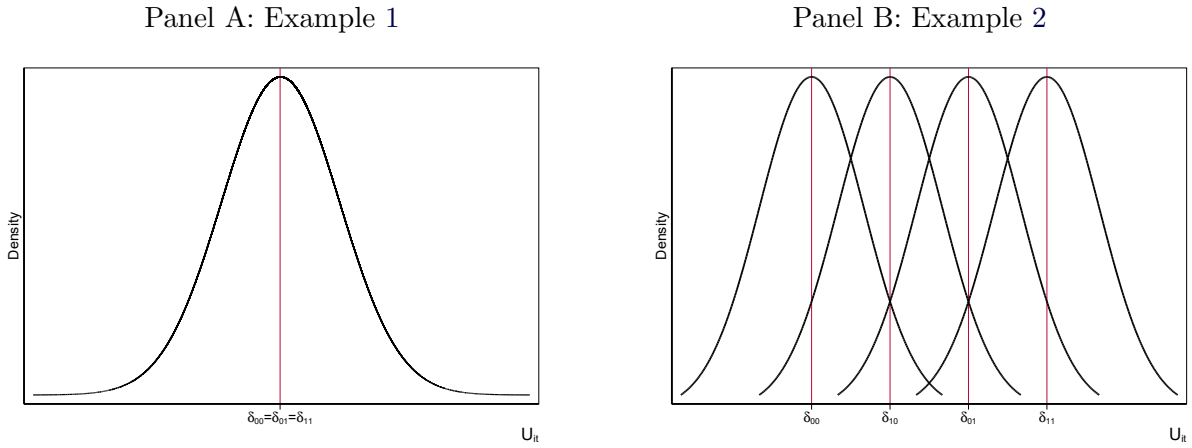
The above discussion and proposition illustrate that equal attrition rates without further assumptions do not imply IV-R. To illustrate this point further, we present two examples.

Example 1. *(Internal Validity & Differential Attrition Rates)*
Assume that potential response satisfies monotonicity, i.e. $p_{10} = 0$*, and* $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$*. 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 3 illustrates the resulting distribution of* U_{it}*. However, due to the presence of treatment-only responders,* $P(R_i = 1 | T_i = 1) = p_{11} + p_{01}$*, and* $P(R_i = 1 | T_i = 0) = p_{11}$*. Hence, even though we not only have IV-R but also IV-P, we have differential attrition rates* ($-p_{01}$) *across treatment and control groups.*

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,[32] *but their proportions are equal* ($p_{10} = p_{01} > 0$*), which yields equal*

[32]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 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

Figure 3: Distribution of U_{it} for Different Response Types



Notes: The above figure illustrates the distribution of U_{it} for the different subpopulations in 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 $(r_0, r_1) \in \{0,1\}^2$.

attrition rates across treatment and control groups.[33] *If* $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$*, then the different response types will have different distributions of unobservables, as illustrated in Panel B of Figure 3. As a result, the distribution of* (U_{i0}, U_{i1}) *for treatment and control respondents defined in* (20)-(21) *will be different and hence IV-R is violated.*

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 in empirical practice is that we cannot use it to test IV-P, even in cases where the differential attrition rate test is a valid test of IV-R. For instance, consider the case in which monotonicity holds and the attrition rates are equal across groups. We can then identify the ATE-R, since the respondent subpopulation is composed solely of always-responders as pointed out above. If the researcher is interested in identifying the treatment effect for the study population, however, s/he would have to test whether the always-responders are "representative" of the study population. To do so, one would have to test the restriction of the IV-P assumption in Proposition 1(b.ii).

applications where monotonicity is implausible and establishes identification of a local average treatment effect under an alternative assumption.

[33] In the multiple treatment case, equal attrition rates are possible without requiring any two response types to have equal proportions in the population. See Section C in the online appendix for a derivation.

3.3 Implications for Empirical Practice

Our theoretical analysis underscores the importance of the object of interest in determining the appropriateness of an attrition test. Hence, explicitly stating the object of interest, whether it is the ATE-R, ATE, CATE-R or CATE, is important to justify a particular attrition test. Our results further clarify the interpretation of attrition tests in the field experiment literature. The differential attrition rate test, which is implemented in 81% of papers in our review, is not based on a necessary condition of IV-R. Most of the selective attrition tests, which are performed in 60% of the papers, are based on mean implications of the IV-R assumption. The most common version of this test (40% of all papers) uses respondents only; and hence, it does not exploit all the information in the baseline sample, specifically the attritors. Sixteen percent of papers do implement a selective attrition test that includes both respondents and attritors, suggesting that some authors are aware of the value of this information. Several of the null hypotheses they use, however, do not constitute IV-R or IV-P tests. This is perhaps unsurprising given the wide range of null hypotheses tested (see Section B.2 in the online appendix). Although authors do not in general conduct a direct test of IV-P, the inclusion of respondents and attritors in some selective attrition tests as well as the use of determinants of attrition tests suggest that some authors are likely interested in IV-P.

3.3.1 Mean Tests of Internal Validity

The vast majority of selective attrition tests implemented in the literature are based on restrictions on the mean of the baseline variables in question. The distributional restrictions in Proposition 1 test the IV-R (IV-P) assumption which implies the identification of the entire distribution of the potential outcomes, and as a result the identification of the ATE-R (ATE). In some experiments, however, researchers may be solely interested in average treatment effects. Here, we discuss the weaker sufficient conditions to identify these objects and their sharp testable implications.

If the ATE-R is the object of interest, then the following assumption is sufficient for its identification,

$$E[Y_{it}(0)|T_i, R_i] = E[Y_{it}(0)|R_i],\ t = 0, 1. \tag{7}$$

Note that this assumption is implied by the IV-R assumption in Proposition 1(a). Its sharp testable implication, $E[Y_{i0}|T_i, R_i] = E[Y_{i0}|R_i]$, is the mean version of the testable restriction in Proposition 1(a), so it includes testable restrictions on attritors and respondents. Similarly,

if the object of interest is the ATE, then the relevant identifying assumption is

$$E[Y_{it}(0)|T_i, R_i] = E[Y_{it}(0)], t = 0, 1, \tag{8}$$

which is similarly the mean version of the IV-P assumption in Proposition 1(b). The testable restriction of this assumption, $E[Y_{i0}|T_i, R_i] = E[Y_{i0}]$, involves all treatment-response subgroups as its distributional version in Proposition 1(b.ii).

The mean restrictions of the identifying assumptions of the ATE-R and ATE can be implemented in a straightforward manner using regression-based tests. We present the relevant regressions and null hypotheses in Section B.

3.3.2 The role of covariates

An important question that arises in empirical practice is whether to include covariates in attrition tests. In some cases, using covariates may be the only way to test attrition bias. In particular, some experiments target a population for which the baseline outcome always takes on the same value by design (i.e. if a job training program is targeted to unemployed people and employment is the main outcome). In other field experiments, baseline outcome data may not be available. We therefore provide a formal discussion of the role of covariates in attrition tests in this section.

Suppose that the researcher has the following *a priori* information on W_{it}, a $d_W \times 1$ vector of covariates,

$$W_{it} = \nu_t(U_{it}), \tag{9}$$

i.e. that these covariates are functions of U_{it}, the determinants of the outcome Y_{it}, for $t = 0, 1$. This identifies two types of covariates: (i) covariates that are themselves determinants of the outcome, i.e. $W_{it}^k = U_{it}^j$ for some k, j, $k = 1, \ldots, d_W$, $j = 1, \ldots, d_U$, or (ii) "proxy" variables, which are covariates determined by the same factors as the outcome Y_{it}. If this *a priori* information is true, the sharp testable restrictions of the IV-R and IV-P assumptions in Proposition 1 as well as Proposition 2 would be imposed on the joint distribution of $Z_{i0} = (Y_{i0}, W_{i0}')'$ and not solely on the marginal distribution of Y_{i0}.[34] However, if this *a priori* information is false and W_{it} also depends on unobservables that affect response, V_i, i.e. $W_{it} = \xi_t(U_{it}, V_i)$, then the testable restrictions on W_{i0} may be violated even if the identifying assumption in question holds.

The main takeaway from the above discussion is that the testable restriction of the IV-

[34] See Section B for details on regression tests for the multivariate case.

R or IV-P assumption for the outcome in question, Y_{it}, consists of a joint hypothesis of the relevant restrictions on the vector of baseline outcome and covariates, Z_{i0}, assuming (9). This outcome-specific approach to testing attrition bias is further supported by our Progresa example, which illustrates empirically that attrition may affect internal validity differently depending on the outcome in question.

In our review of field experiments, we find that most authors use covariates in attrition tests regardless of the design of the study. The average number of variables (outcomes and covariates) used in the selective attrition tests in our review is 17 with 75% of those tests using more than 10 variables and a maximum of 46. It is important to point out that studies that implement the selective attrition tests on all baseline variables, $\mathcal{Z}_{i0} = (Y_{i0}, W'_{i0}, X'_{i0})'$, are testing the IV-R assumption for all determinants of those variables, $\mathcal{E}_{it} = (U'_{it}, \eta'_{it})'$, where $\mathcal{Z}_{it} = \xi_t(\mathcal{E}_{it})$. Our results suggest that the inclusion of X_{i0}, which do not satisfy (9) by definition (i.e. not solely determined by U_{it}), may lead to false rejection of the IV-R or IV-P assumption for the outcome in question. Another reason for potential over-rejection of internal validity in the literature is that a substantial proportion of the implementation of selective attrition tests in the literature consists of individual tests for each baseline variable without correcting for multiple testing.

The implications of our analysis for empirical practice resonate with existing recommendations in the literature regarding the random assignment method used to ensure the similarity of treatment and control groups in terms of baseline observables in a given sample (i.e. "balance"). In seminal work on clinical trials, Altman (1985) emphasizes that imbalance should only concern the researcher if the variable in question relates to the outcome. Bruhn and McKenzie (2009) compare different stratified randomization procedures in terms of their ability to achieve balance. They point to the potential cost of using "irrelevant" variables in their simulation study and find that baseline outcome is by far the most informative determinant of future outcomes in various datasets.

4 Randomization Tests of Internal Validity

We present randomization procedures to test the IV-R and IV-P assumptions for completely and stratified randomized experiments. If distributional treatment effects are the object of interest, then the distributional hypotheses are the testable restrictions of the relevant identifying assumptions. Furthermore, the use of randomization tests is an increasingly common approach to estimating treatment effects (Young, 2018). Thus, authors may want to implement randomization procedures when testing for attrition bias, even when their focus is on mean rather than distributional effects.

The proposed procedures approximate the exact p-values of the proposed distributional statistics under the cross-sectional i.i.d. assumption when the outcome distribution is continuous.[35] They can also be adapted to accommodate possibly discrete or mixed outcome distributions, which may result from rounding or censoring in the data collection, by applying the procedure in Dufour (2006). In this section, we focus on distributional statistics for the testable restrictions on the baseline outcome as in Propositions 1 and 2. The randomization procedures we propose, however, can be applied to test joint distributional hypotheses that include covariates as in Section 3.3.2.

We first outline a general randomization procedure that we adapt to the different settings we consider.[36] 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, $\hat{p}_{n,B} = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} 1\{\hat{T}_n(g_b) \geq T_n\}\right)$.*

A test that rejects when $\hat{p}_{n,B} \leq \alpha$ 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.

4.1 Completely Randomized Trials

The testable restriction of the IV-R assumption, 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

$$H_0^1 : F_{Y_{i0}|T_i=0,R_i=r} = F_{Y_{i0}|T_i=1,R_i=r} \text{ for } r = 0, 1. \tag{10}$$

[35] We maintain the cross-sectional i.i.d. assumption to simplify the presentation. The randomization procedures proposed here remain valid under weaker exchangeability-type assumptions.

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

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

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

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 norm, i.e. $\|f\|_{n,2} = \sum_{i=1}^n f(y_i)^2/n$. In order to test the *joint* hypothesis in (10), the two following statistics that aggregate over $T^1_{n,r}$ for $r = 0, 1$ are standard choices in the literature (Imbens and Rubin, 2015),[37]

$$\begin{aligned} T^1_{n,m} &= \max\{T^1_{n,0}, T^1_{n,1}\}, \\ T^1_{n,p} &= p_{n,0} T^1_{n,0} + p_{n,1} T^1_{n,1}, \quad \text{where } p_{n,r} = \sum_{i=1}^n 1\{R_i = r\}/n \text{ for } r = 0, 1. \end{aligned}$$

Let $\mathcal{G}^1_0$ denote the set of all permutations of individual observations within respondent and attritor subgroups, for $g \in \mathcal{G}^1_0$, $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, R_{g(i)}) : R_{g(i)} = R_i, 1 \leq i \leq n\}$. Under H^1_0 and the cross-sectional i.i.d. assumption, $\mathbf{Z} \overset{d}{=} g(\mathbf{Z})$ for $g \in \mathcal{G}^1_0$. Hence, we can obtain p-values for $T^1_{n,m}$ and $T^1_{n,p}$ under H^1_0 by applying Procedure 1 using the set of permutations $\mathcal{G}^1_0$.

We now consider testing the restriction of the IV-P assumption 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) denote the j^{th} element of $\mathcal{T} \times \mathcal{R}$. Then, the joint hypothesis is given wlog by

$$H^2_0 : \; F_{Y_{i0}|T_i=\tau_j,R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1},R_i=r_{j+1}} \text{ for } j = 1, \ldots, |\mathcal{T} \times \mathcal{R}| - 1. \tag{11}$$

[37] There are other possible approaches to construct joint statistics. We compare the finite-sample performance of the two joint statistics we consider numerically in Section E of the online appendix.

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

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

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 $g \in \mathcal{G}_0^2$, $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, R_{g(i)}) : 1 \leq i \leq n\}$. We can hence apply Procedure 1 using $\mathcal{G}_0^2$ to obtain approximately exact p-values for the statistic $T^2_{n,m}$ or $T^2_{n,p}$ under H_0^2.

4.2 Stratified Randomized Trials

As pointed out in Section 3.1.5, the testable restrictions in the case of stratified or block randomized trials (Proposition 2) 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 2(a.ii), which yields the following null hypothesis

$$H_0^{1,\mathcal{S}} : F_{Y_{i0}|T_i=0,S_i=s,R_i=r} = F_{Y_{i0}|T_i=1,S_i=s,R_i=r} \text{ for } r = 0, 1,\ s \in \mathcal{S}. \tag{12}$$

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

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

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

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

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 $g \in \mathcal{G}_0^{1,\mathcal{S}}$, $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, S_{g(i)}, R_{g(i)}) : S_{g(i)} = S_i, R_{g(i)} = R_i, 1 \le i \le n\}$, where $\mathbf{Z} = \{(Y_{i0}, T_i, S_i, R_i) : 1 \le i \le n\}$. 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 $T_{n,m}^{1,\mathcal{S}}$ and $T_{n,p}^{1,\mathcal{S}}$ under $H_0^{1,\mathcal{S}}$.

We now consider testing the restriction in Proposition 2(b.ii). The resulting null hypothesis is given wlog by the following

$$H_0^{2,\mathcal{S}} : F_{Y_{i0}|T_i=\tau_j,S_i=s,R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1},S_i=s,R_i=r_{j+1}} \text{ for } j = 1,\ldots,|\mathcal{T}\times\mathcal{R}|-1,\ s \in \mathcal{S}. \tag{13}$$

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

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

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

$$T_{n,m}^{2,\mathcal{S}} = \max_{s\in\mathcal{S}} T_{n,m,s}^{2,\mathcal{S}},$$
$$T_{n,p}^{2,\mathcal{S}} = \sum_{s\in\mathcal{S}} w_s T_{n,p,s}^{2,\mathcal{S}},$$

given fixed or random non-negative weights w_s 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 $g \in \mathcal{G}_0^{2,\mathcal{S}}$, $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, S_{g(i)}, R_{g(i)}) : S_{g(i)} = S_i, 1 \le i \le n\}$. Thus, applying Procedure 1 to $T_{n,m}^{2,\mathcal{S}}$ or $T_{n,p}^{2,\mathcal{S}}$ using $\mathcal{G}_0^{2,\mathcal{S}}$ yields approximately exact p-values for these statistics under $H_0^{2,\mathcal{S}}$.

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 the IV-R assumption, 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 the IV-P assumption conditional on stratum.

The aforementioned subgroup-randomization procedures split the original sample into respondents and attritors or four treatment-response groups. This approach does not directly extend to cluster randomized experiments.[38] Given the widespread use of regression-based tests in the empirical literature, we illustrate how to test the mean implications of the distributional restrictions of the IV-R and IV-P assumptions using regressions for completely, cluster, and stratified randomized experiments in Section B.

5 Simulation Study

We illustrate the theoretical results in the paper using a numerical study. The simulations demonstrate the performance of the differential attrition rate test as well as both the mean and distributional tests of the IV-R and IV-P assumptions.

5.1 Simulation Design

The data-generating process (DGP) is described in Panel A of Table 4. 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 response types according to proportions given by $p_{r_0r_1}$ for $(r_0, r_1) \in \{0, 1\}^2$. The unobservable, U_{it}, has time-varying and time-invariant components. The time-varying unobservable, η_{i1}, follows an AR(1) process and is independent of potential response in all variants of our design for simplicity. We allow dependence between the time-invariant unobservable, α_i, and potential response by allowing the means of the conditional distributions to differ for each response type (i.e. $\delta_{r_0r_1}$ for all $(r_0, r_1) \in \{0, 1\}^2$), while maintaining $E[\alpha_i] = 0$. Conversely, when the conditional mean is the same for all subpopulations, α_i and potential response are independent. In order to introduce treatment heterogeneity, treatment enters into two terms of the outcome equation: $\beta_1 D_{it}$ and $\beta_2 D_{it}\alpha_i$. Specifically, letting β_2 be non-zero allows for the ATE-R to differ from the ATE. The ATE always equals β_1, however, since $E[\alpha_i] = 0$.

[38] To test the distributional restrictions for cluster randomized experiments, the bootstrap-adjusted critical values for the KS and CM-type statistics in Ghanem (2017) can be implemented.

We conduct simulations using four variants of this simulation design, which are summarized in Panel B of Table 4.[39] 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 potential response ($U_{it} \not\perp (R_i(0), R_i(1))$). We also allow attrition rates to differ across the treatment and control groups. Design II demonstrates a setting in which there is IV-R, but not IV-P. For that set-up, we impose monotonicity in the response equation as well as equal attrition rates, while allowing for dependence between U_{it} and $(R_i(0), R_i(1))$.

Table 4: 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 ($p_{10} = p_{01}$)
$(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$	No	No	Yes	No

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

[39] We only consider 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 $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$, then we would have equal attrition rates and IV-P. 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 IV-R only (Proposition 3(iii)).

Designs III and IV illustrate *Examples 1* and *2* in Section 3.2, respectively. Design III demonstrates a setting in which we have differential attrition rates and IV-P. Specifically, Design III relies on the assumptions of monotonicity and differential attrition rates as in Design I, but assumes independence between U_{it} 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 U_{it} and $(R_i(0), R_i(1))$, and a violation of monotonicity by letting p_{10} and p_{01} be non-zero. We maintain equal attrition rates in this design by imposing $p_{01} = p_{10}$.

We use a sample size of $n = 2,000$ as well as 2,000 simulation replications. We chose a range of attrition rates from the results of our review of the empirical literature (see Figure 1). Specifically, we allow for attrition rates in the control group from 5% to 30%, and differential attrition rates from zero to ten percentage points.

5.2 Differential Attrition Rates and Tests of Internal Validity

Table 5 reports simulation rejection probablilities for the differential attrition rate test as well as the mean and distributional tests of the IV-R and IV-P assumption across Designs I-IV using a 5% level of significance. We also report the estimated difference in mean outcomes for the treatment and control respondents in the follow-up period ($t = 1$),

$$\bar{Y}_1^{TR} - \bar{Y}_1^{CR} = \frac{\sum_{i=1}^n Y_{i1} D_{i1} R_i}{\sum_{i=1}^n D_{i1} R_i} - \frac{\sum_{i=1}^n Y_{i1}(1 - D_{i1}) R_i}{\sum_{i=1}^n (1 - D_{i1}) R_i}, \tag{14}$$

its standard deviation, and the rejection probability of a t-test of its significance ($\hat{p}_{0.05}$) in columns 10 through 12 of Table 5.

First, we consider the performance of the differential attrition rate test. Columns 1 through 3 of Table 5 report the simulation mean of the attrition rates for the control (C) and treatment (T) groups as well as the probability of rejecting a differential attrition rate test, which is a two-sample t-test of the equality of attrition rates between groups. The differential attrition rate test rejects at a simulation frequency above the nominal level (5%) in Designs I and III, whereas it rejects at approximately the nominal level in 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 and allow for dependence between U_{it} and potential response, illustrate the typical cases in which the differential attrition rate test can be viewed as a test of IV-R.

Designs 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 IV-R. In Design III, the unobservables in the outcome equation are independent of potential response. Thus, 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, however, allows for equal attrition rates but a violation of internal validity. Thus, the differential attrition rate test does not reject above nominal levels.

Columns 4 through 7 of Table 5 report simulation results of the tests of the IV-R assumption. The first three tests are based on the following mean testable restrictions from Proposition 1(a.ii),

$$\begin{aligned} H^{1,1}_{0,\mathcal{M}}: &\quad E[Y_{i0}|T_i=0,R_i=1]=E[Y_{i0}|T_i=1,R_i=1], && (CR-TR)\\ H^{1,2}_{0,\mathcal{M}}: &\quad E[Y_{i0}|T_i=0,R_i=0]=E[Y_{i0}|T_i=1,R_i=0], && (CA-TA)\\ H^{1}_{0,\mathcal{M}} &: \quad H^{1,1}_{0,\mathcal{M}}\ \&\ H^{1,2}_{0,\mathcal{M}}, && (Joint) \end{aligned} \tag{15}$$

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 for respondents (attritors) only. We present the tests of these two hypotheses since they rely on an approach that is similar to widely used tests in the literature. The mean implication of the sharp testable restriction in Proposition 1(a.ii), $H^{1}_{0,\mathcal{M}}$, is a joint hypothesis of $H^{1,1}_{0,\mathcal{M}}$ and $H^{1,2}_{0,\mathcal{M}}$. These hypotheses are linear restrictions on the fully saturated regression of baseline outcome on treatment and response given in Section B, which we test using χ^2 statistics. We also examine the finite-sample performance of the KS statistic of the sharp testable restriction of the IV-R assumption in (10). The reported p-values of the KS statistic defined below are obtained using the randomization procedure to test H^1_0 from Section 4,

$$\begin{aligned} KS^1_{n,m} &= \max\{KS^1_{n,0}, KS^1_{n,1}\}, \text{where for } r=0,1\\ KS^1_{n,r} &= \max_{i:R_i=r} \left|\sqrt{n}\left(F_{n,Y_{i0}}(y_{i0}|T_i=1,R_i=r)-F_{n,Y_{i0}}(y_{i0}|T_i=0,R_i=r)\right)\right|. \end{aligned} \tag{16}$$

The tests of the IV-R assumption behave according to our theoretical predictions. In Designs II and III, where IV-R holds, the tests control size. In Designs I and IV, where IV-R is violated, they reject with simulation probability above the nominal level. In general, the relative power of the test statistics may differ depending on the DGP. In our simulation design, however, the rejection probabilities of the attritors-only test (CA-TA) and the joint tests ($Mean$ and KS) are significantly higher than the test based on the difference between the treatment and control respondents (CR-TR).[40]

[40]This may be because the treatment-only responders are proportionately larger in the control attritor subgroup than in the treatment respondent subgroup.

Columns 8 and 9 of Table 5 report the simulation results of the mean and distributional tests of the IV-P assumption given in Proposition 1(b.ii). The distributional hypothesis H_0^2 is given in (11). Its mean version is defined as follows

$$H_{0,\mathcal{M}}^2 : E[Y_{i0}|T_i = \tau_j, R_i = r_j] = E[Y_{i0}|T_i = \tau_{j+1}, R_i = r_{j+1}] \text{ for } j = 1, \ldots, |\mathcal{T} \times \mathcal{R}| - 1, \tag{17}$$

where (τ_j, r_j) denote the j^{th} element of $\mathcal{T} \times \mathcal{R} = \{(0,0), (0,1), (1,0), (1,1)\}$. We test the mean version of the hypothesis using the χ^2 statistic of the linear restrictions on the regression in Section B as in the above. To test the distributional hypothesis, we use the KS statistic given below

$$\begin{aligned} KS_n^2 &= \max_{j=1,2,3} KS_{n,j}^2, \text{ where} \\ KS_{n,j}^2 &= \max_{i:(T_i,R_i)\in\{(\tau_j,r_j),(\tau_{j+1},r_{j+1})\}} \left|\sqrt{n}\left(F_{n,Y_{i0}|T_i=\tau_{j-1},R_i=r_{j-1}} - F_{n,Y_{i0}|T_i=\tau_j,R_i=r_j}\right)\right|. \end{aligned} \tag{18}$$

The p-values of the KS statistic are obtained using the randomization procedure to test H_0^2 in Section 4.

The test statistics of the IV-P assumption also behave according to our theoretical predictions. In Designs I, II and IV, where $U_{it} \not\perp (R_i(0), R_i(1))$, they reject the IV-P assumption at a simulation frequency higher than the nominal level. Design II is notable since IV-R holds, but IV-P does not. Thus, while the mean tests of the IV-R assumption are not rejected at a simulation frequency above the nominal level, the tests of the IV-P assumption are rejected above the nominal level. In addition, the difference in mean outcomes between treatment and control respondents is different from the ATE (0.25), even though it is internally valid for the respondents. In Design III, which is the only design where IV-P holds, both the mean and KS tests control size. Examining the difference in mean outcomes between treatment and control respondents at follow-up in this design, we find that it is unbiased for the ATE across all combinations of attrition rates.

Overall, the simulation results illustrate the limitations of the differential attrition rate test and show that the tests of the IV-R and IV-P assumptions we propose behave according to our theoretical predictions. For a more thorough numerical analysis of the finite-sample behavior of the KS and CM statistics, see Section E in the online appendix.

6 Empirical Applications

To complement the simulations presented above, we apply the proposed tests of attrition bias to five published field experiments. This exercise builds on the simulation results by demonstrating the existence of a few notable regularities on a set of data generated from experiments. The data comes from five articles with both high attrition rates and publicly available data that includes attritors.[41] Thus, the exercise is not intended to draw inference about implications of applying various attrition tests to a representative sample of published field experiments. In addition, field experiments that are published in prestigious journals may not to be representative of all field experiment data–especially if perceptions of attrition bias had an impact on publication.

6.1 Implementation of Attrition Tests

Across the five selected articles included in this exercise, we conduct attrition tests for a total of 33 outcomes. This includes all outcomes that are reported in the abstracts as well as all other unique outcomes.[42] For each outcome included in this exercise, the appropriate attrition test depends on the type of outcome and the approach to randomization used in the experiment. For fully randomized experiments, we apply joint tests of the IV-R and IV-P assumptions in Proposition 1. For stratified experiments, we instead apply the tests of the assumptions in Proposition 2.[43] For continuous outcomes in non-clustered experiments, we report p-values of the KS distributional tests using the appropriate randomization procedure.[44] For binary outcomes and also for all outcomes from clustered experiments, we apply regression-based mean tests (see Section B). For all tests, the results are presented in a way that is designed to preserve the anonymity of the results and papers. Thus, attrition rates are presented as ranges, the results are not linked to specific articles, and we randomize the order of the outcomes such that they are not listed by paper.

In addition to the tests of the restrictions in Propositions 1 and 2, we also consider how our tests might compare to other approaches. We apply a version of the tests commonly used in the literature for completeness, including: the differential attrition rate test, the IV-

[41] We selected the articles with the five highest attrition rates for which the data required to implement the attrition tests is available (see Section A.2 in the online appendix for details).

[42] If the article reports results separately by wave, we report attrition tests for each wave of a given outcome. We did not, however, report results for each heterogeneous treatment effect unless those results were reported in the abstract.

[43] When the number of strata in the experiment is larger than ten, we conduct a test with strata fixed effects only as opposed to the fully interacted regression in Section B in order to avoid high dimensional inference issues. Under the null, this specification is an implication of the sharp testable restrictions proposed in Proposition 2.

[44] We apply the Dufour (2006) randomization procedure to accommodate the possibility of ties.

R test using the respondent subsample only and the IV-R test using the attritor subsample only. We use the same approaches to handling stratification and continuous outcomes in all three IV-R tests to ensure they are directly comparable, but that also means that those tests do not generally replicate versions of those attrition tests that are used in published field experiments. Instead, we indicate whether authors' attrition tests reject for the outcomes for which they are available. In keeping with our findings from Section 2, there is heterogeneity in the application of attrition tests across these articles. Two of the articles reported only a differential attrition rate test, while three also reported some type of selective attrition test.

Our approach differs from the authors of the selected papers in a number of key ways. First, since response rates vary across questions within a survey, we calculate a separate (differential) attrition rate for each outcome. Second, we also conduct separate tests of attrition bias for each outcome reported in the abstract. Third, we focus only on the baseline outcome, while authors typically include covariates.

6.2 Results of the Empirical Applications

First, we consider the results of the attrition tests reported by the authors (Table 6). The authors conduct differential attrition rate tests that are relevant to 30 of the 33 outcomes that we consider. They reject the null hypothesis of equal attrition rates at a significance level of 5% in 23 cases. The authors conduct a selective attrition test for 8 of the 33 outcomes that we include in our exercise.[45] Conditional on implementing a selective attrition test, the authors largely do not find evidence of selective attrition; they reject the null hypothesis at the 10% level for two of the eight outcomes. Since authors do not state their object of interest, it is not clear whether they intend to test for IV-R or IV-P.

We implement outcome-specific differential attrition rate tests that are directly comparable to our IV-R and IV-P tests. We reject the null hypothesis at the 5% level for 9 out of 33 outcomes (3 outcomes after correcting for multiple hypothesis testing).[46] The relatively high differential attrition rates we find in this exercise are perhaps not surprising, given that overall attrition rates and differential attrition rates seem to be correlated, and these outcomes have fairly high attrition rates (McKenzie, 2019).

In contrast, for the proposed joint IV-R test, all of the reported p-values are larger than 10%. Thus, for any of the outcomes reported here, a researcher using this test would

[45] This type of test is much less widely implemented since the authors of two of the articles that we consider do not conduct selective attrition tests, and authors that do conduct selective attrition tests typically implement them on a limited set of outcomes.

[46] We note that our test of differential attrition rate rejects less frequently than the authors' implementation of the test. This difference appears to be driven by authors' use of the survey-level (differential) attrition rate for the test while we focus on the outcome-level (differential) attrition rates.

not reject the identifying assumption that implies that differences between treatment and control respondents are internally valid for the respondent subpopulation. Similarly, the IV-R tests using only respondents or attritors have p-values larger than 5% for all 33 outcomes. Although there is often a substantial difference in the p-values for these two simple tests relative to the joint test for a given outcome, there is no consistent pattern in the direction of those differences.

Next, we consider the results of our proposed IV-P test. We do not reject the IV-P assumption at the 5% level for a majority of outcomes in this exercise, specifically 26 out of 33 (28 of 33 when accounting for multiple hypothesis testing).[47] In addition, we find empirical cases that are consistent with the testable implications of *Example 1*. For 8 out of the 33 outcomes, the differential attrition rate test that we implement rejects the null hypothesis at the 5% level, whereas the IV-P test does not reject at the 5% level. This provides suggestive evidence that the theoretical conditions under which the differential attrition rate test does not control size are empirically relevant. Overall, our results have promising implications for randomized experiments in which the study population is intended to be representative of a larger population.

7 Conclusion

This paper presents the problem of testing attrition bias in field experiments with baseline outcome data as an identification problem in a panel model. The proposed tests are based on the sharp testable restrictions of the identifying assumptions of the specific object of interest: either the average treatment effect for the respondents, the average treatment effect for the study population or a heterogeneous treatment effect. This study also provides theoretical conditions under which the differential attrition rate test, a widely used test, may not control size as a test of internal validity. The theoretical analysis has important implications for current empirical practice in testing attrition bias in field experiments. It also highlights that the majority of testing procedures used in the empirical literature have focused on the internal validity of treatment effects for the respondent subpopulation. The theoretical and empirical results, however, suggest that the treatment effects of the study population are important and possibly attainable in practice.

While this paper is a step forward toward understanding current empirical practice and establishing a standard in testing attrition bias in field experiments, it opens several questions for future research. Despite the availability of several approaches to correct for attrition

[47] Although the number of outcomes from a given field experiment varies widely, the results are not driven by any one experiment or type of outcome.

bias (Lee, 2009; Huber, 2012; Behagel et al., 2015; Millán and Macours, 2019), alternative approaches that exploit the information in baseline outcome data as in the framework here may require weaker assumptions and hence constitute an important direction for future work. The extension of the analysis in this paper to the problem of attrition in the presence of partial compliance is another interesting direction. Furthermore, several practical aspects of the implementation of the proposed test may lead to pre-test bias issues. For instance, the proposed tests may be used in practice to inform whether an attrition correction is warranted or not in the empirical analysis. Empirical researchers may also be interested in first testing the identifying assumption for treatment effects for the respondent subpopulation and then testing their validity for the entire study population. Inference procedures that correct for these and other pre-test bias issues are a priority for future work.

Finally, this paper has several policy implications. 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 rate standards for studies (IES, 2017). Our results indicate an alternative approach to assessing potential attrition bias. Furthermore, questions regarding external validity of treatment effects measured from field experiments are especially important from a policy perspective. This paper points to the possibility that in the presence of response problems, the identified effect in a given field experiment may only be valid for the respondent subpopulation, and hence may not identify the ATE for the study population. This is an important issue to consider when synthesizing results of field experiments to inform policy.

Table 5: Simulation Results on Differential Attrition Rates and Tests of Internal Validity ($ATE = 0.25$)

Design	Attrition Rates		Differential Attrition Rate Test	Tests of the IV-R Assumption				Tests of the IV-P Assumption		Difference in Mean Outcomes between Treatment & Control Respondents ($\bar{Y}_1^{TR} - \bar{Y}_1^{CR}$)		
				Mean Tests			KS Test	Mean Test	KS Test			
	C	T	$\hat{p}_{0.05}$	CR-TR	CA-TA	Joint	Joint	Joint	Joint	Mean	SD	$\hat{p}_{0.05}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
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	0.446	0.353	0.324	0.452	0.476	0.265	0.057	0.997
	0.10	0.05	0.995	0.076	0.719	0.635	0.582	0.792	0.787	0.282	0.058	0.998
	0.15	0.10	0.935	0.072	0.631	0.542	0.483	0.995	0.980	0.288	0.061	0.997
	0.20	0.15	0.867	0.072	0.532	0.442	0.412	1.000	1.000	0.296	0.063	0.996
	0.30	0.20	1.000	0.141	0.894	0.851	0.801	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	0.044	0.053	0.062	0.981	0.902	0.255	0.058	0.993
	0.10	0.10	0.053	0.043	0.045	0.045	0.056	1.000	0.999	0.262	0.060	0.991
	0.15	0.15	0.052	0.043	0.049	0.052	0.055	1.000	1.000	0.271	0.062	0.992
	0.20	0.20	0.049	0.045	0.047	0.050	0.050	1.000	1.000	0.280	0.064	0.990
	0.30	0.30	0.048	0.053	0.044	0.046	0.043	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	0.051	0.056	0.052	0.065	0.050	0.248	0.058	0.990
	0.10	0.05	0.995	0.055	0.050	0.055	0.046	0.053	0.055	0.248	0.059	0.985
	0.15	0.10	0.935	0.057	0.052	0.053	0.045	0.053	0.059	0.247	0.061	0.983
	0.20	0.15	0.867	0.058	0.047	0.053	0.046	0.048	0.048	0.247	0.063	0.974
	0.30	0.20	1.000	0.057	0.053	0.052	0.043	0.049	0.048	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	0.429	0.337	0.329	0.360	0.311	0.273	0.058	0.997
	0.10	0.10	0.013	0.131	0.708	0.653	0.577	0.708	0.582	0.302	0.059	0.999
	0.15	0.15	0.007	0.248	0.873	0.855	0.758	0.888	0.792	0.333	0.061	0.999
	0.20	0.20	0.004	0.422	0.934	0.951	0.859	0.970	0.913	0.367	0.063	0.999
	0.30	0.30	0.001	0.797	0.990	0.997	0.974	0.999	0.998	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 the IV-R (IV-P) assumption refer to the regression tests (Section B) of the null hypothesis in (15) ((17)). The KS statistics of the IV-R (IV-P) assumption are given in (16) ((18)), and their p-values are obtained using the proposed randomization procedures ($B = 199$). The simulation mean, standard deviation (SD), and rejection probability of a two-sample t-test are reported for the difference in mean outcome between treatment and control respondents, $\bar{Y}_1^{TR} - \bar{Y}_1^{CR}$, defined in (14). All tests are conducted using $\alpha = 0.05$. Additional details of the design are provided in Table 4.

† (*) indicates IV-R only (IV-P).

Table 6: Attrition Tests Applied to Outcomes from Five Field Experiments

Outcome	Attrition Rate		Differential Attrition Rate Test	Tests of the IV-R Assumption			Test of the IV-P Assumption	Authors Reject the Null for:	
	Control (%)	Differential (percentage points)		CR-TR	CA-TA	Joint	Joint	Differential Attrition Rates Test	Selective Attrition Test
1	[10 - 30]	(10 - 20]	0.025†	0.567	0.948	0.832	0.563	Yes: 5%	No
2	[10 - 30]	(0 - 5]	0.887	0.514	0.546	0.571	0.60	No	Yes: 10%
3	[10 - 30]	(0 - 5]	0.109	0.834	0.751	0.879	0.956	Yes: 5%	–
4	[10 - 30]	(0 - 5]	0.486	0.351	0.701	0.576	0.000*	Yes: 5%	–
5	[10 - 30]	(0 - 5]	0.100	0.421	0.526	0.668	0.755	Yes: 5%	–
6	[10 - 30]	(0 - 5]	0.086	0.392	0.098	0.187	0.313	Yes: 5%	–
7	[10 - 30]	(0 - 5]	0.056	0.315	0.575	0.490	0.652	Yes: 5%	–
8	[10 - 30]	(0 - 5]	0.027	0.359	0.381	0.537	0.679	Yes: 5%	–
9	[10 - 30]	(0 - 5]	0.129	0.190	0.532	0.312	0.008*	Yes: 5%	–
10	[30 - 50]	(0 - 5]	0.301	0.202	0.191	0.198	0.002*	Yes: 5%	–
11	[10 - 30]	(0 - 5]	0.030	0.688	0.966	0.917	0.979	Yes: 5%	–
12	[10 - 30]	(0 - 5]	0.955	0.120	0.114	0.250	0.000*	No	–
13	[10 - 30]	(10 - 20]	0.039†	0.827	0.120	0.277	0.441	Yes: 5%	–
14	[10 - 30]	(0 - 5]	0.788	0.861	0.194	0.423	0.525	No	–
15	[10 - 30]	(10 - 20]	0.048†	0.682	0.558	0.800	0.609	Yes: 5%	No
16	[10 - 30]	(0 - 5]	0.798	0.802	0.180	0.404	0.590	No	No
17	[10 - 30]	(10 - 20]	0.037†	0.685	0.428	0.711	0.843	Yes: 5%	–
18	[10 - 30]	(0 - 5]	0.784	0.833	0.169	0.384	0.546	No	–
19	[30 - 50]	(0 - 5]	0.127	0.700	0.494	0.690	0.010*	Yes: 5%	–
20	[30 - 50]	(0 - 5]	0.241	0.605	0.476	0.720	0.697	Yes: 5%	–
21	[10 - 30]	(0 - 5]	0.084	0.796	0.261	0.518	0.671	Yes: 5%	–
22	[30 - 50]	(0 - 5]	0.218	0.748	0.183	0.385	0.022†	Yes: 5%	–
23	[30 - 50]	(0 - 5]	0.128	0.328	0.632	0.615	0.053	Yes: 5%	–
24	[30 - 50]	(0 - 5]	0.134	0.133	0.976	0.337	0.528	Yes: 5%	–
25	[30 - 50]	(0 - 5]	0.118	0.718	0.510	0.707	0.029†	Yes: 5%	–
26	[30 - 50]	(0 - 5]	0.348	0.663	0.370	0.691	0.807	Yes: 5%	–
27	[30 - 50]	(0 - 5]	0.217	0.883	0.768	0.858	0.423	Yes: 5%	–
28	[10 - 30]	(0 - 5]	0.061	0.218	0.986	0.518	0.609	Yes: 5%	–
29	[10 - 30]	(5 - 10]	0.036*	0.276	0.698	0.832	0.106	–	–
30	[10 - 30]	(10 - 20]	0.000*	0.354	0.984	0.864	0.064	–	No
31	[30 - 50]	(10 - 20]	0.047*	0.144	0.440	0.526	0.692	–	Yes: 10%
32	[10 - 30]	(0 - 5]	0.867	0.580	0.509	0.798	0.720	No	No
33	[10 - 30]	(5 - 10]	0.437	0.421	0.887	0.683	0.447	No	No

Notes: The table reports p-values for the differential attrition rate test as well as tests of the IV-R and IV-P assumptions. The symbol * (†) next to the p-value indicates that the relevant test statistic remains statistically significant after applying the Benjamini-Hochberg correction at 5% (10%) (see Benjamini and Hochberg (1995) for details on this procedure). $CR-TR$ ($CA-TA$) indicates difference across treatment and control respondents (attritors). Joint tests include all four treatment-response sub-groups. Regression tests are implemented for (i) the differential attrition rate test, (ii) for the IV-R and IV-P tests with binary outcomes, and (iii) for cluster-randomized trials. Standard errors are clustered (if treatment is randomized at the cluster level) and strata fixed effects are included (if treatment is randomized within strata). For continuous outcomes in non-clustered trials, p-values of the KS tests are implemented using the appropriate randomization procedures ($B = 499$). For stratified experiments with less than ten strata, the test proposed in Proposition 2 is implemented. The last two columns of the table report whether (and the significance level at which) the authors reject their tests of differential attrition rates and selective attrition, respectively. The dash indicates that the test was not reported by the authors.

A 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 $d = 0, 1$, $F_{Y_{it}(d)|T_i,R_i} = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}|T_i,R_i}(u) = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}|R_i}(u) = F_{Y_{it}(d)|R_i}$ for $t = 0, 1$. (i) follows by letting $t = 1$ and $d = 0$, while conditioning the left-hand side of the last equation on $T_i = 0$ and $R_i = 1$, and the testable implication in (ii) follows by letting $t = d = 0$.

Following Hsu et al. (2019), 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 two distributions above 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 two distributions above 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 for $d = 0, 1$ and $t = 0, 1$, $F_{Y_{it}(d)|T_i,R_i} = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}|T_i,R_i}(u) = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}}(u) = F_{Y_{it}(d)}$. (i) follows by letting $t = 1$, while conditioning the left-hand side of the last equation on $T_i = \tau$ and $R_i = 1$ for $d = \tau$ and $\tau = 0, 1$, 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 $Y_{i0}|T_i, R_i \stackrel{d}{=} Y_{i0}(0)$, 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) The proof is immediate from the proof of Proposition 1 by conditioning all statements on S_i. □

Proof. (Proposition 3) For notational brevity, let $U_i = (U_{i0}', U_{i1}')$. We first note that by random assignment, it follows that

$$
F_{U_i|T_i,R_i(0),R_i(1)} = F_{U_i|T_i,\xi(0,V_i),\xi(1,V_i)} = F_{U_i|\xi(0,V_i),\xi(1,V_i)} = F_{U_i|R_i(0),R_i(1)}. \tag{19}
$$

As a result,

$$
F_{U_i|T_i=1,R_i=1} = \frac{p_{01} F_{U_i|(R_i(0),R_i(1))=(0,1)} + p_{11} F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i = 1|T_i = 1)}, \tag{20}
$$

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

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

$$
F_{U_i|T_i=1,R_i=1} = \frac{p_{01} F_{U_i} + p_{11} F_{U_i}}{P(R_i = 1|T_i = 1)} = F_{U_i}, \quad F_{U_i|T_i=0,R_i=1} = \frac{p_{10} F_{U_i} + p_{11} F_{U_i}}{P(R_i = 1|T_i = 0)} = F_{U_i}.
$$

We can similarly show that $F_{U_i|T_i,R_i=0} = F_{U_i}$, it follows trivially that $U_i|T_i, R_i \stackrel{d}{=} U_i|R_i$.

Alternatively, if we assume (ii), $R_i(0) \leq R_i(1)$ 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 (20) and (21) both equal $F_{U_i|(R_i(0),R_i(1))=(1,1)}$. Similarly, it follows that $F_{U_i|T_i=1,R_i=0} = F_{U_i|T_i=0,R_i=0} = F_{U_i|(R_i(0),R_i(1))=(0,0)}$, 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,

$$
\begin{aligned}
F_{U_i|T_i=1,R_i=1} &= \frac{p_{01} F_{U_i|(R_i(0),R_i(1))=(0,1)} + p_{11} F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i = 1|T_i = 1)} \\
&= \frac{p_{10} F_{U_i|(R_i(0),R_i(1))=(1,0)} + p_{11} F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i = 1|T_i = 0)} = F_{U_i|T_i=0,R_i=1}.
\end{aligned} \tag{22}
$$

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

B 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.

B.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.

If a column-vector W_{i0} of d_W baseline covariates is also used to test for internal validity, then the regression-based test should consider both the baseline outcome and the baseline covariates, i.e. $Z_{i0} = (Y_{i0}, W'_{i0})'$, $\forall j = 1, \ldots, (d_W + 1)$

$$
\begin{aligned}
Z^j_{i0} &= \gamma^j_{11} T_i R_i + \gamma^j_{01}(1-T_i)R_i + \gamma^j_{10} T_i(1-R_i) + \gamma^j_{00}(1-T_i)(1-R_i) + \epsilon_i \\
H^1_{0,\mathcal{M}} &: \gamma^j_{11} = \gamma^j_{01} \;\&\; \gamma^j_{10} = \gamma^j_{00} \quad \forall \quad j = 1, \ldots, (d_W + 1) \\
H^2_{0,\mathcal{M}} &: \gamma^j_{11} = \gamma^j_{01} = \gamma^j_{10} = \gamma^j_{00} \quad \forall \quad j = 1, \ldots, (d_W + 1)
\end{aligned}
$$

B.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.

If a column-vector W_{i0} of d_W baseline covariates is also used to test for internal validity, then the regression-based test should consider both the baseline outcome and the baseline covariates, i.e. $Z_{i0} = (Y_{i0}, W^{'}_{i0})'$

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

$$H_{0,\mathcal{M}}^{1,s} : \gamma_{11}^{s,j} = \gamma_{01}^{s,j} \;\&\; \gamma_{10}^{s,j} = \gamma_{00}^{s,j} \quad \forall \quad j = 1, \ldots, (d_W + 1)$$
$$H_{0,\mathcal{M}}^{2,s} : \gamma_{11}^{s,j} = \gamma_{01}^{s,j} = \gamma_{10}^{s,j} = \gamma_{00}^{s,j} \quad \forall \quad j = 1, \ldots, (d_W + 1)$$

References

Abadie, Alberto, Matthew M. Chingos, and Martin R. West, "Endogenous Stratification in Randomized Experiments," *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.

Altman, Douglas G., "Comparability of Randomised Groups," *Journal of the Royal Statistical Society: Series D (The Statistician)*, 1985, *34* (1), 125–136.

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

Andrews, Isaiah and Emily Oster, "A simple approximation for evaluating external validity bias," *Economics Letters*, 2019, *178*, 58 – 62.

Angrist, Joshua D., 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 Experiments," 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*, 2018, *113* (521), 230–240.

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.

Baird, Sarah, J. Aislinn Bohren, Craig McIntosh, and Berk Özler, "Optimal Design of Experiments in the Presence of Interference," *Review of Economics and Statistics*, 2018, *100* (5), 844–860.

Barrett, Garry, Peter Levell, and Kevin Milligan, "A Comparison of Micro and Macro Expenditure Measures across Countries Using Differing Survey Methods," in "Improving the Measurement of Consumer Expenditures" NBER Chapters, National Bureau of Economic Research, Inc, 2014, pp. 263–286.

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

Benjamini, Yoav and Yosef Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1995, *57* (1), 289–300.

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.

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*, 2018, *113* (524), 1784–1796.

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

Chen, Xuan and Carlos A. Flores, "Bounds on Treatment Effects in the Presence of Sample Selection and Noncompliance: The Wage Effects of Job Corps," *Journal of Business & Economic Statistics*, 2015, *33* (4), 523–540.

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," *Review of Economic Studies*, 2003, *70* (1), 33–58.

de Chaisemartin, Clément, "Tolerating Defiance? Local Average Treatment Effects Without Monotonicity," *Quantitative Economics*, 2017, *8* (2), 367–396.

_ **and Luc Behaghel**, "Estimating the Effect of Treatments Allocated by Randomized Waiting Lists," Papers 1511.01453, arXiv.org November 2018.

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 in Linear Regressions," *Econometrics Journal*, 1998, *1* (1), 154–173.

Fricke, Hans, Markus Fröhlich, Martin Huber, and Michael Lechner, "Endogeneity and Non-Response Bias in Treatment Evaluation: Nonparametric Identification of Causal Effects by Instruments," 2015. IZA Discussion Papers, No. 9428, Institute for the Study of Labor (IZA), Bonn.

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.

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

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 Sanford V. Berg, ed., *Annals of Economic and Social Measurement*, Vol. 5, National Bureau of Economic Research, 1976, pp. 475–492.

Heckman, James J., "Sample Selection Bias as A Specification Error," *Econometrica*, 1979, *47* (1), 153–161.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 2003, *71* (4), 1161–1189.

_ , _ , _ , and Donald B. Rubin, "Combining Panel Data Sets with Attrition and Refreshment Samples," *Econometrica*, 2001, *69* (6), 1645–1659.

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.

Horvitz, D. G. and D. J. Thompson, "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 1952, *47* (260), 663–685.

Hsu, Yu-Chin, Chu-An Liu, and Xiaoxia Shi, "Testing Generalized Regression Monotonicity," *Econometric Theory*, 2019, p. 1 – 55.

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

IES, "What Works Clearinghouse. Standards Handbook Version 4.0," Technical Report, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse 2017.

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.

INSP, "General Rural Methodology Note," Technical Report, Instituto Nacional de Salud Publica 2005.

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," *Review of Economic Studies*, 2009, *76* (3), 1071–1102.

Lehmann, E. L. and Joseph P. Romano, *Testing Statistical Hypotheses*, 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.

McKenzie, David, "Beyond Baseline and Follow-Up: The Case for More T in Experiments," *Journal of Development Economics*, 2012, *99* (2), 210–221.

_ , "Attrition Rates Typically Aren't that Different for The Control Group than The Treatment Group – Really? and Why?," *Development Impact Blog*, January 07, 2019. https://blogs.worldbank.org/impactevaluations/attrition-rates-typically-aren-t-different-control-group-treatment-group-really-and-why.

Meyer, Bruce D., Wallace K. C. Mok, and James X. Sullivan, "Household Surveys in Crisis," *Journal of Economic Perspectives*, November 2015, *29* (4), 199–226.

Millán, Teresa Molina and Karen Macours, "Attrition in Randomized Control Trials: Using Tracking Information to Correct Bias," 2019. Unpublished Manuscript.

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

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao, "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 1994, *89* (427), 846–866.

Rubin, Donald B., "Inference and Missing Data," *Biometrika*, 1976, *63* (3), 581–592.

Skoufias, Emmanuel, "PROGRESA and Its Impacts on The Welfare of Rural households in Mexico," Research Report 139, International Food Policy Research Institute (IFPRI) 2005.

Wooldridge, Jeffrey M., "Selection corrections for panel data models under conditional mean independence assumptions," *Journal of Econometrics*, 1995, *68* (1), 115 – 132.

Young, Alwyn, "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*," *Quarterly Journal of Economics*, 11 2018, *134* (2), 557–598.