



*The World's Largest Open Access Agricultural & Applied Economics Digital Library*

**This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.**

**Help ensure our sustainability.**

**Give to AgEcon Search**

AgEcon Search

<http://ageconsearch.umn.edu>

[aesearch@umn.edu](mailto:aesearch@umn.edu)

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

*No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.*

## Testing Attrition Bias in Field

### Experiments

Dalia Ghanem, UC Davis, [dghanem@ucdavis.edu](mailto:dghanem@ucdavis.edu)

Sarojini Hirshleifer, UC Riverside,

[sarojini.hirshleifer@ucr.edu](mailto:sarojini.hirshleifer@ucr.edu)

Karen Ortiz-Becerra, UC Davis, [kaortizb@ucdavis.edu](mailto:kaortizb@ucdavis.edu)

*Selected Paper prepared for presentation at the 2019 Agricultural & Applied Economics Association  
Annual Meeting, Atlanta, GA, July 21 – July 23*

*Copyright 2019 by [authors]. All rights reserved. Readers may make verbatim copies of this document for non-commercial purposes by any means, provided that this copyright notice appears on all such copies.*

# Testing Attrition Bias in Field Experiments\*

Dalia Ghanem  
UC Davis

Sarojini Hirshleifer  
UC Riverside

Karen Ortiz-Becerra  
UC Davis

August 6, 2019

## 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.

*Keywords:* attrition, field experiments, randomized experiments, randomized controlled trials, internal validity, Kolmogorov-Smirnov, Cramer-von-Mises, randomization tests

*JEL Codes:* C12, C21, C33, C93

---

\*E-mail: [dghanem@ucdavis.edu](mailto:dghanem@ucdavis.edu), [sarojini.hirshleifer@ucr.edu](mailto:sarojini.hirshleifer@ucr.edu), [kaortizb@ucdavis.edu](mailto: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.<sup>1</sup> 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 experimental sample. Conflict, intimidation or natural disasters sometimes make it unsafe to collect complete response data. The recent, increased focus on the long-term impacts of interventions has also made non-response especially relevant. Thus, researchers often face the question: How much of a threat is attrition to the internal validity of a given study?

In this paper, we approach attrition in field experiments with baseline outcome data as an identification problem in a nonseparable panel model. We focus on two identification questions generated by attrition in field experiments. First, does the difference in mean outcomes between treatment and control respondents identify the average treatment effect for the respondent subpopulation (ATE-R)? Second, is this estimand equal to the average treatment effect for the study population (ATE)?<sup>2</sup> 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 is common in published field experiments: the majority of such experiments have at least one attrition rate relevant to a main result that is higher than 10%, and a minority have attrition rates that are substantially higher. 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. Our review indicates that attrition tests are widely used, and their implementation varies substantially across papers. 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-

---

<sup>1</sup>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.

<sup>2</sup>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. First, if the researcher’s objective is internal validity for the respondent subpopulation (IV-R), then the identifying assumption is random assignment conditional on response status. 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. Second, if internal validity for the study population (IV-P) is of interest, then the identifying assumption is that the unobservables that affect the outcome are independent of response conditional on treatment assignment. This assumption implies the identification of the ATE for the study population. This second case is especially relevant in settings where the study population is representative of a larger population.

We then derive testable restrictions for each of the above identifying assumptions. The assumption required for IV-R implies a joint hypothesis of two equalities on the baseline outcome distribution; specifically, for treatment and control respondents as well as treatment and control attriters. Meanwhile, the assumption required for IV-P implies a joint hypothesis of equality on the baseline outcome distribution across all four treatment/response subgroups. The approach presented in this paper highlights that a test of attrition bias is a test of an identifying assumption, which (like other identifying assumptions) can only be tested by implication in general. Hence, we show that the aforementioned testable restrictions are sharp, meaning that they are the strongest implications that we can test given our data.<sup>3</sup> 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. This example demonstrates that across two outcomes collected in the same survey it is possible to reject the IV-P identifying assumption for one outcome, while not rejecting it for the other.

Since the assumptions required for identification are random-assignment-type restrictions, randomization tests are a natural choice in this context.<sup>4</sup> 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.

---

<sup>3</sup>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, Liu and Shi (2019), and Mourifié and Wan (2017).

<sup>4</sup>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.

We also provide a formal treatment of the differential attrition rate test, since it is the most frequently used attrition test in the field experiment literature. In order to do so, we apply the framework of partial compliance from the local average treatment effect (LATE) literature to potential response.<sup>5</sup> 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. An analytical example illustrates that it is possible to have differences in attrition rates across treatment and control groups while IV-P holds.

A simulation experiment illustrates our analytical results. In our design, the mean and distributional tests of the IV-R (IV-P) assumption only reject at a higher-than-nominal level when IV-R (IV-P) is violated. In contrast, the differential attrition rate test: (i) does not control size in some cases when internal validity holds, (ii) can have trivial power in some cases when internal validity is violated.

To examine the empirical relevance of our results, we apply the differential attrition rate test as well as our tests of the IV-R and IV-P assumptions to 33 outcomes from five published field experiments with high overall attrition rates. For all outcomes in this exercise, the p-values for the IV-R test are larger than 5%. More surprisingly, the p-values of the IV-P test are also larger than 5% for a large proportion of the outcomes. This is promising for field experiments where IV-P is of interest. Finally, we find multiple cases with large and statistically significant differential attrition rates. In the overwhelming majority of those cases, the p-value of the IV-P test is larger than 5%. These results are consistent with the theoretical conditions under which the differential attrition rate test does not control size, thereby providing evidence of their empirical relevance.

This paper has several implications for empirical practice. Our theoretical and empirical results provide evidence that the differential attrition rate test may lead to a false rejection of internal validity in practice. In addition, most of the approaches in the literature to testing for selective attrition 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 (i.e. respondents and attriters). A substantial minority of researchers also examine differences in the baseline distributions between respondents and attriters. This suggests that some researchers may be interested in implications of the estimated treatment effects for the study population (IV-P). More generally, this paper highlights the importance of understanding the implications of attrition to a broader population when interpreting field experiment results for policy.<sup>6</sup>

---

<sup>5</sup>See the foundational work in the LATE literature (Imbens and Angrist, 1994; Angrist, Imbens and Rubin, 1996).

<sup>6</sup>External validity can be assessed in number of ways (see, for example, Andrews and Oster (2019) and Azzam, Bates and Fairris (2018)). In our setting, we note that even if the IV-P assumption is rejected, if

This paper contributes to a growing literature that considers methodological questions relevant to field experiments.<sup>7</sup> Given the wide use of attrition tests, we formally examine the testing problem here. There is a thread in this literature however that outlines various approaches to correcting attrition bias in field experiments (e.g. [Behagel et al., 2015](#); [Millán and Macours, 2017](#)). Our paper also relates to recent work that examines the potential use of randomization tests in analyzing field experiment data ([Young, 2018](#); [Athey and Imbens, 2017](#); [Athey, Eckles and Imbens, 2018](#); [Bugni, Canay and Shaikh, 2018](#)).

The attrition corrections in the field experiments literature build on the larger sample selection literature in econometrics going back to [Heckman \(1976, 1979\)](#). Nonparametric Heckman-style corrections have been proposed for linear and nonparametric outcome models (e.g. [Ahn and Powell, 1993](#); [Das, Newey and Vella, 2003](#)).<sup>8</sup> Inverse probability weighting is another important category of corrections for sample selection bias (e.g. [Angrist, 1997](#); [Wooldridge, 2007](#)).<sup>9</sup> In addition, a strand in this literature examines attrition corrections for panel data (e.g. [Hausman and Wise, 1979](#); [Wooldridge, 1995](#); [Hirano et al., 2001](#)). Nonparametric bounds is an alternative approach which requires weaker assumptions. [Horowitz and Manski \(2000\)](#), [Manski \(2005\)](#) and [Kline and Santos \(2013\)](#) propose bounds on the conditional outcome distribution and related objects of interest. The sample selection literature is broadly concerned with objects that pertain to the population. [Lee \(2009\)](#) bounds the average treatment effect for a subpopulation assuming monotonicity of selection. Our paper provides tests of identifying assumptions emphasizing the distinction between the (study) population and the respondent subpopulation.

This paper also builds 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

---

we do not reject IV-R we may still be able to draw inferences from the local average treatment effect for respondents to a broader population.

<sup>7</sup>[Bruhn and McKenzie \(2009\)](#) compare the performance of different randomization methods; [McKenzie \(2012\)](#) discusses the power trade-offs of the number of follow-up samples in the experimental design; [Baird et al. \(2018\)](#) propose an optimal method to design field experiments in the presence of interference; [de Chaisemartin and Behagel \(2018\)](#) present how to estimate treatment effects in the context of randomized wait lists; [Abadie, Chingos and West \(2018\)](#) propose alternative estimators that reduce the bias resulting from endogenous stratification in field experiments.

<sup>8</sup>[Vella \(1998\)](#) provides a detailed review of this category of sample selection models. See [Brownstone \(1998\)](#) for some interesting discussions on several sample selection corrections and the trade-offs between them.

<sup>9</sup>[Angrist \(1997\)](#) examines the connection between Heckman-style corrections and the inverse probability weighting approach.

(Dufour, 2006; Dufour et al., 1998).

The paper proceeds as follows. Section 2 presents the review of the field experiments literature. Section 3 formally presents the identifying assumptions and their sharp testable restrictions. In Section 4, we propose a subgroup-randomization procedure to obtain  $p$ -values for the distributional test statistics. Section 5 presents simulation experiments to illustrate the theoretical results. Section 6 presents the results of the empirical applications. 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.<sup>10</sup> 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. We identify two main types of tests that aim to determine the impact of attrition on internal validity: (i) a *differential attrition rate test*, and (ii) a *selective attrition test*. A *differential attrition rate test* determines whether the rates of attrition are statistically significantly different across treatment and control groups. In contrast, a *selective attrition test* determines whether, conditional on being a respondent and/or attritor, the mean of observable characteristics is the same across treatment and control groups. We also consider whether selective attrition tests include both respondents and attritors as well as whether authors test for differences in the baseline distributions of attritors and respondents. Our categorization imposes some structure on the variety of different estimation strategies used to test for attrition bias in the literature.<sup>11</sup>

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

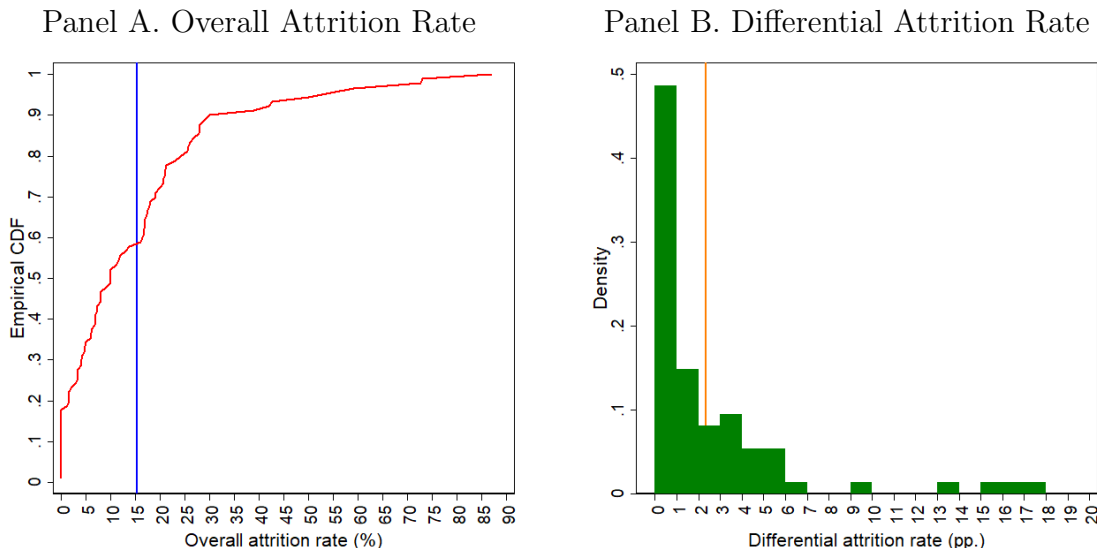
---

<sup>10</sup>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 E in the online appendix contains a list of all the papers included in the review.

<sup>11</sup>We identify fifteen estimation strategies used to conduct attrition tests. See Section A.2 in the online appendix.



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



*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 both the treatment and control groups. The *Differential* rate is the absolute value of the difference in attrition rates across treatment and control groups. The blue (orange) line depicts the average overall (differential) attrition rate in our sample of field experiments. Panel A includes 90 experiments and Panel B includes 74 experiments since the relevant attrition rates are not reported in some articles.

the average of 15%.<sup>12</sup> Of the experiments that report a differential attrition rate, Panel B in Figure 1 illustrates that a majority have little differential attrition for the abstract results: 66% have a differential rate that is less than 2 percentage points, and only 12% have a differential attrition rate that is greater than 5 percentage points.<sup>13</sup> These distributions of overall and differential attrition rates inform our simulations in Section 5.

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 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

<sup>12</sup>We focus on attrition rates that are relevant to outcomes reported in the abstract (i.e. “abstract results”). Although attrition rates may differ at the level of the outcome (given varying response rates across questions in the same survey), most papers report attrition rates at the level of the data source or subsample. For some experiments, all the abstract results are drawn from one data source/subsample, but in other experiments they are not. We include one attrition rate per field experiment for consistency. We report the data source relevant to an abstract result with the highest attrition rate to understand the extent of attrition that is relevant to the main outcomes in the paper. Authors do not in general report attrition rates conditional on baseline response.

<sup>13</sup>It is possible, however, that these numbers reflect authors’ exclusion of results with higher differential attrition rates than those that were reported or published.

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 test do not conduct a selective attrition test.<sup>14</sup>

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:	
Joint test: report at least one test using both samples	20%
Simple test: only using sample of respondents	68%
Simple test: only using sample of attritors	5%
Simple test: one using respondents & one using attritors	7%
Total <sup>†</sup>	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 43 of those experiments that conducted a selective attrition test (†).

We further categorize selective attrition tests as *simple* or *joint* tests. A simple test determines whether treatment and control respondents *or* treatment and control attritors are the same in terms of mean baseline characteristics, while a joint test determines whether treatment and control groups are the same within respondents and attritors jointly. Panel B in Table 1 illustrates the proportion of papers that conduct the simple and joint selective attrition tests. Conditional on having conducted any type of selective attrition test, authors

<sup>14</sup>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 any strong correlations given the available data.

attempt a joint test on only 20% of those field experiments. Instead, authors conduct a simple test of selective attrition on the sample of respondents in most cases (68%).

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 for the full population selected for the study. First, we define a *determinants of attrition test* as a test of whether baseline outcomes and covariates correlate with response status. In approximately one-third of field experiments (34%), the authors conduct a determinants of attrition test (Panel C of Table 1). Table 1 illustrates that conducting such a test does not have a one-to-one relationship with either conducting a differential attrition rate test or conducting a selective attrition test.<sup>15</sup> In addition, most of the articles that implement joint selective attrition tests use an estimation strategy that could test both for differences across treatment and control groups of the baseline means of the observables conditional on response status as well as for differences across the respondents and attritors. We do not find clear evidence, however, that the authors are typically using a null hypothesis that is designed to test for both of these differences jointly.

### 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 as well as present the extension of the results to stratified randomization and heterogeneous treatment effects.

---

<sup>15</sup>Approximately half the determinants of attrition tests are conducted in the same regression as a differential attrition rate test. 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 rely on 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.<sup>16</sup> The surveys conducted for the experiment include a baseline and three follow-up rounds collected 5, 13, and 18 months after the program began.<sup>17</sup> 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*

Round	Full Sample				Respondent Subsample at Follow-up			
	N	Control Mean	$T - C$	$p$ -value	Attrition Rate	Control Mean	$T - C$	$p$ -value
<i>Panel A. School Enrollment (6-16 years old)</i>								
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
<i>Panel B. Employment Last Week (18+ years old)</i>								
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. It 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 random assignment of treatment was implemented as intended. In the absence of attrition, the difference in a mean outcome

<sup>16</sup>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 at the official [website](#) of the evaluation.

<sup>17</sup>The baseline was collected in October 1997 and the three follow-ups were collected in October 1998, June 1999, and November 1999.

across the 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 across treatment and control 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  $i = 1, \dots, n$  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  $D_{it} \in \{0, 1\}$ .<sup>18</sup> 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 vector of time-invariant and time-varying unobservables,  $U_{it}$ ,

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

Given this structural function, we can define the potential outcomes  $Y_{it}(d) = \mu_t(d, U_{it})$  for  $d = 0, 1$ .<sup>19</sup>

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 between the treatment and control groups in the follow-up period can be attributed to the

<sup>18</sup>The extension to the multiple treatment case is in Section C of the online appendix.

<sup>19</sup>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.

treatment. The ATE can be identified from the following difference in mean outcomes,

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

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,<sup>20</sup>

$$R_i = \xi(T_i, V_i), \quad (3)$$

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*

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

---

<sup>20</sup>Since non-response is only allowed in the follow-up period, we omit time subscripts from the response equation for notational convenience.

(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.<sup>21</sup> 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 attriters.

The assumption in (b) implies *missing-at-random* as defined in Manski (2005).<sup>22</sup> Together with random assignment, it implies that both treatment and response status are jointly independent of the unobservables in the outcome equation. We will refer to this case as *internal validity for the study population* (IV-P) and the assumption in (b) as the IV-P assumption. The equality in (b.i) implies the identification of the ATE from 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.<sup>23</sup> The restriction in (b.ii) is the testable implication of the IV-P assumption under random assignment. The resulting null hypothesis is the equality of the baseline outcome distribution regardless of both treatment and response status.

<sup>21</sup>We state our assumptions in terms of the joint distribution of  $(U_{i0}, U_{i1})$  to be consistent with the statement of random assignment. Our results also follow if we replace the assumptions on the joint distribution by their counterparts on the marginal distribution of  $U_{it}$  for  $t = 0, 1$ .

<sup>22</sup>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.

<sup>23</sup>If the linear model holds, the ATE-R equals the ATE, specifically if  $Y_{i1} = \alpha + \beta D_{i1} + U_{i1}$ , then  $\beta = E[Y_{i1}(1) - Y_{i1}(0)|R_i = 1] = E[Y_{i1}(1) - Y_{i1}(0)]$ . This relates to Proposition 1 in Angrist (1997), which establishes that random assignment conditional on response status and missing at random are equivalent in the context of a cross-sectional linear model if treatment is randomly assigned.

### 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 attriters. 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 attriters, 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 attriters in the same sample was 61%. Thus, children that are observed in the follow-up data are substantially different than 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: 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	<i>p</i> -value	<i>p</i> -value
<i>Panel A. School Enrollment (6-16 years old)</i>								
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
<i>Panel B. Employment Last Week (18+ years old)</i>								
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 attriters (TA), and control attriters (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.

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 cannot reject the IV-R assumption, but we do reject the IV-P assumption at the 5% significance level.<sup>24</sup> 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.<sup>25</sup> In other words, we do not reject the assumption that the difference in employment rates between treatment and control respondents at follow-up identifies the ATE.

Understanding treatment effects for the study population is especially relevant for large-scale programs such as *Progres*a, where the study population is representative of a broader population of interest. 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 even when identification fails. 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

---

<sup>24</sup>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 for each outcome.

<sup>25</sup>A natural question that arises in 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 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.

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  $Y_{i0}|T_i = 1, R_i \stackrel{d}{=} Y_{i0}|T_i = 0, R_i$  holds,  $Y_{i1}(0)|T_i = 1, R_i = 1 \neq Y_{i1}|T_i = 0, R_i = 1$ . 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.<sup>26</sup>

The above example illustrates why we cannot test identification “directly”, since it would require us to observe the counterfactual of the treatment respondents. As a result, it is crucial to test identifying assumptions by using their sharp testable restrictions (i.e. their strongest possible implications on the data).

### 3.1.5 Stratified Randomization and Heterogeneous Treatment Effects

In many field experiments, randomization is performed within strata or blocks for a variety of reasons, including implementation design (i.e. the study is randomized within roll-out waves or locations) and concerns about power. One important reason to randomize within strata is to better identify heterogeneous treatment effects, more formally defined as conditional average treatment effects (CATE). In this section, we extend Proposition 1 to the case of stratified randomization. We examine the identification of the counterfactual distribution for

<sup>26</sup>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.

the respondent subpopulation as well as for the study population. The results in this section also apply to completely randomized experiments when heterogeneous treatment effects are of interest.

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

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

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

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

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

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

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

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

The ATE-R can then be identified by averaging over  $S_i$ , i.e.  $\sum_{s \in \mathcal{S}} P(S_i = s | R_i = 1) (E[Y_{i1} | T_i = 1, S_i = s, R_i = 1] - E[Y_{i1} | T_i = 0, S_i = s, R_i = 1])$ . 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 stratum 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 may 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, Imbens and Rubin, 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  $(R_i(0), R_i(1))$ .

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 ( $R_i = 1$ )	$(R_i(0), R_i(1)) = (1, 0)$ $(R_i(0), R_i(1)) = (1, 1)$	$(R_i(0), R_i(1)) = (0, 1)$ $(R_i(0), R_i(1)) = (1, 1)$

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:

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

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 differential attrition rates and the IV-R assumption in 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 unobservable distributions of the different response types  $(R_i(0), R_i(1))$ . We omit the analysis for attriters for brevity, since it is analogous. 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: (i) if the distributions of treatment-only, control-only and always-responders all agree, which is implied by  $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$ ; (ii) if there were no treatment-only or control-only responders, i.e.  $p_{10} = p_{01} = 0$ , which is a special case of monotonicity as discussed in [Lee \(2009\)](#); (iii) if  $p_{10} = p_{01}$  and the distribution of unobservables that affect the outcome for the treatment-only and the control-only responders are identical. The equality of distribution for treatment-only and control-only responders is implied by an exchangeability restriction ([Altonji and Matzkin, 2005](#)) given below. These three sets of assumptions imply the IV-R assumption as formally stated 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)  $(U_{i0}, U_{i1}) | R_i(0), R_i(1) \stackrel{d}{=} (U_{i0}, U_{i1}) | R_i(0) + R_i(1)$  (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. First, it is weaker than monotonicity, since

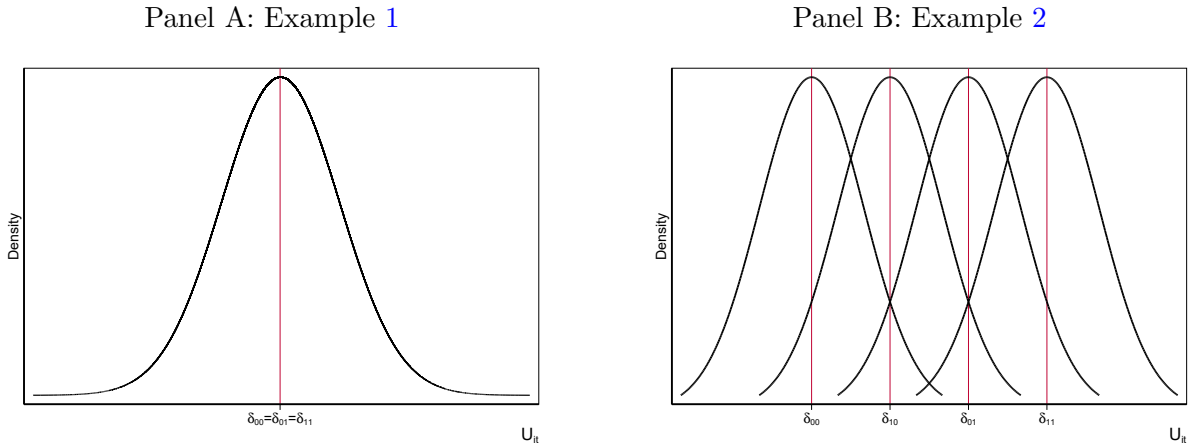
it allows for both treatment-only and control-only responders, but it assumes that these “inconsistent” types have the same distribution of  $(U_{i0}, U_{i1})$ . While strong in general, this assumption may be more realistic in experiments with two treatments. If coupled with equal attrition rates, exchangeability implies 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 the assumption in Proposition 3(i) holds,  $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$ . Furthermore, there is a group of individuals for whom it is too costly to respond if they are in the control group. This group will only respond if assigned to the treatment group (treatment-only responders), and thereby  $p_{01} > 0$ . Panel A of Figure 3 illustrates that these assumptions imply that the three response types in this example have the same distribution of  $U_{it}$ . Under random assignment,  $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1)) \Rightarrow (U_{i0}, U_{i1})|T_i, R_i \stackrel{d}{=} (U_{i0}, U_{i1})$ , which implies IV-P. Due to the presence of treatment-only responders,  $P(R_i = 0|T_i = 1) = p_{00}$ , and  $P(R_i = 0|T_i = 0) = p_{00} + p_{01}$ . As a result, we have differential attrition rates across the treatment and control groups, even though we not only have IV-R but also IV-P.

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



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

**Example 2.** (*Equal Attrition Rates & Violation of Internal Validity*)

Assume that potential response violates monotonicity, such that there are treatment-only and

control-only responders,<sup>27</sup> but their proportions are equal ( $p_{10} = p_{01} > 0$ ), which yields equal attrition rates across treatment and control groups.<sup>28</sup> 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 (17)-(18) 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 the differential attrition rate in empirical practice is that we cannot use it to test whether the IV-P assumption holds, 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).

### 3.3 Implications for Empirical Practice

Our results clarify the interpretation of attrition tests in the field experiment literature. First, the most commonly used test, the differential attrition rate test, is not based on a necessary condition of IV-R. The selective attrition tests used in the literature are mean implications of the joint distributional tests, in general. The joint test of selective attrition, which is used in 12% of the papers in our review, is based on the mean implication of the

---

<sup>27</sup>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 applications where monotonicity is implausible and establishes identification of a local average treatment effect under an alternative assumption.

<sup>28</sup>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 B in the online appendix for a derivation.



sharp testable restriction of the IV-R assumption in Proposition 1(a.ii).<sup>29</sup> The most common test of selective attrition, however, is the simple test using respondents only. Thus, it does not use all of the testable implications of the IV-R assumption. A large minority of authors do implement a determinants of attrition test. Since it tests differences across respondents and attriters, however, by itself it is not a test of IV-P or IV-R.

An important question that arises in empirical practice is whether covariates should be used in testing the identifying assumption in question. Suppose that the researcher has *a priori* information that establishes that there are covariates determined by the same unobservables as the outcome  $Y_{it}$ , specifically  $W_{it} = \nu_t(U_{it})$  for  $t = 0, 1$ . Then, the sharp testable restrictions of the identifying assumptions in Proposition 1 would be imposed on the joint distribution of  $Z_{i0} = (Y_{i0}, W'_{i0})'$  and not solely on the marginal distribution of  $Y_{i0}$ . If baseline outcome data are not available, the testable restrictions on  $W_{i0}$  can be used to test the IV-R and IV-P assumptions. However, if this *a priori* information is false and  $W_{it}$  also depends on unobservables that affect response,  $V_i$ , then the testable restrictions on  $W_{i0}$  may be violated even if the identifying assumption in question holds. Thus, the choice of covariates is an important consideration if a researcher decides to include them in testing the IV-R or IV-P assumption.

Finally, our theoretical analysis underscores the importance of the object of interest in determining the required identifying assumption and its testable restriction. Hence, explicitly stating the object of interest, whether it is the ATE-R, ATE, CATE-R or CATE, is important to determine whether an attrition test is appropriate in a given setting.

## 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. 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.<sup>30</sup> 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). While we focus on distributional statistics in this section, the randomization procedures we propose can be used to obtain  $p$ -values for  $t$ -tests and other statistics that test the equality of distributions.

---

<sup>29</sup>We note that while implementation of joint selective attrition test in practice typically falls under the IV-R category, some of the estimation strategies used to implement it could be used to test the IV-P assumption depending on the null hypothesis.

<sup>30</sup>We maintain the cross-sectional i.i.d. assumption to simplify the presentation. The randomization procedures proposed here remain valid under suitable exchangeability assumptions.



We first outline a general randomization procedure that we adapt to the different settings we consider.<sup>31</sup> 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  $\mathbf{Z} \stackrel{d}{=} g(\mathbf{Z})$  under  $H_0$  only.

**Procedure 1.** (*Randomization*)

1. For  $g_b$ , which is i.i.d.  $\text{Uniform}(\mathcal{G}_0)$ , compute  $\hat{T}_n(g_b) = T(g_b(\mathbf{Z}))$ ,
2. Repeat Step 1 for  $b = 1, \dots, B$  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  $\alpha$  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 attriters. Thus, the joint hypothesis is given by,

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

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

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

where for a random variable  $X_i$ ,  $F_{n, X_i}$  denotes the empirical cdf, i.e. the sample analogue of  $F_{X_i}$ , and  $\|\cdot\|$  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  $L^\infty$  norm

---

<sup>31</sup>See Lehmann and Romano (2005); Canay, Romano and Shaikh (2017) for a more detailed review.

over the sample points, i.e.  $\|f\|_{n,\infty} = \max_i |f(y_i)|$ , whereas the latter is obtained by using an  $L^2$ -type norm, i.e.  $\|f\|_{n,2} = \sum_{i=1}^n f(y_i)^2/n$ . In order to test the *joint* hypothesis in (7), the two following statistics that aggregate over  $T_{n,r}^1$  for  $r = 0, 1$  are standard choices in the literature (Imbens and Rubin, 2015),<sup>32</sup>

$$T_{n,m}^1 = \max\{T_{n,0}^1, T_{n,1}^1\},$$

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

Let  $\mathcal{G}_0^1$  denote the set of all permutations of individuals within respondent and attritor subgroups, for  $g \in \mathcal{G}_0^1$ ,  $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, R_{g(i)}) : R_{g(i)} = R_i, 1 \leq i \leq n\}$ . Under  $H_0^1$  and the cross-sectional i.i.d. assumption,  $\mathbf{Z} \stackrel{d}{=} g(\mathbf{Z})$  for  $g \in \mathcal{G}_0^1$ . Hence, we can obtain  $p$ -values for  $T_{n,m}^1$  and  $T_{n,p}^1$  under  $H_0^1$  by applying Procedure 1 using the set of permutations  $\mathcal{G}_0^1$ .

We now consider testing the restriction of 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  $(\tau_j, r_j)$  denote the  $j^{\text{th}}$  element of  $\mathcal{T} \times \mathcal{R}$ . Then, the joint hypothesis is given wlog by

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

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

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

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

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

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  $T_{n,m}^2$  or  $T_{n,p}^2$  under  $H_0^2$ .

---

<sup>32</sup>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 D of the online appendix.

## 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

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

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

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

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

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

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  $g \in \mathcal{G}_0^{1,\mathcal{S}}$ ,  $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, S_{g(i)}, R_{g(i)}) : S_{g(i)} = S_i, R_{g(i)} = R_i, 1 \leq i \leq n\}$ , where  $\mathbf{Z} = \{(Y_{i0}, T_i, S_i, R_i) : 1 \leq i \leq n\}$ . Under  $H_0^{1,\mathcal{S}}$  and the cross-sectional i.i.d. assumption within strata,  $\mathbf{Z} \stackrel{d}{=} g(\mathbf{Z})$  for  $g \in \mathcal{G}_0^{1,\mathcal{S}}$ . Hence, using  $\mathcal{G}_0^{1,\mathcal{S}}$ , 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

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

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

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

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

given fixed or random non-negative weights  $w_{j,s}$  for  $j = 1, \dots, |\mathcal{T} \times \mathcal{R}| - 1$  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  $g \in \mathcal{G}_0^{2,\mathcal{S}}$ ,  $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, S_{g(i)}, R_{g(i)}) : S_{g(i)} = S_i, 1 \leq i \leq n\}$ . 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 either respondents and attritors or four treatment-response groups. Then, treatment or treatment-response status is randomized at the individual level, respectively. This approach does not directly extend to cluster randomized experiments.<sup>33</sup> 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.

---

<sup>33</sup>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.

## 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.

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:	$T_i \overset{i.i.d.}{\sim} \text{Bernoulli}(0.5)$ , $D_{i0} = 0$ , $D_{i1} = T_i$ .			
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:	$\begin{cases} U_{it} = (\alpha_i, \eta_{it})', t = 0, 1, \\ \alpha_i   R_i(0), R_i(1) \overset{i.i.d.}{\sim} \begin{cases} N(\delta_{00}, 1) \text{ if } (R_i(0), R_i(1)) = (0, 0), \\ N(\delta_{01}, 1) \text{ if } (R_i(0), R_i(1)) = (0, 1), \\ N(\delta_{10}, 1) \text{ if } (R_i(0), R_i(1)) = (1, 0), \\ N(\delta_{11}, 1) \text{ if } (R_i(0), R_i(1)) = (1, 1). \end{cases} \\ \eta_{i1} = 0.5\eta_{i0} + \epsilon_{i0}, (\eta_{i0}, \epsilon_{i0})' \overset{i.i.d.}{\sim} N(0, 0.5I_2) \end{cases}$			

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  $P(R_i = 0|T_i = 0) = p_{00} + p_{10}$ .

### 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  $p_{r_0 r_1}$  for  $(r_0, r_1) \in \{0, 1\}^2$ . The unobservable,  $U_{it}$ , has time-varying and time-invariant components. The time-varying unobservable,  $\eta_{i1}$ , 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,  $\alpha_i$ , and potential response by allowing the means of the conditional distributions to differ for each response type (i.e.  $\delta_{r_0 r_1}$  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,  $\alpha_i$  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  $\beta_2$  be non-zero allows for the ATE-R to differ from the ATE. The ATE always equals  $\beta_1$ , however, since  $E[\alpha_i] = 0$ .

We conduct simulations using four variants of this simulation design, which are summarized in Panel B of Table 4.<sup>34</sup> Design I demonstrates the case in which the differential attrition rate test would in fact detect a violation of internal validity. This case requires both monotonicity in the response equation as well as dependence between the unobservables that affect the outcome and the potential response ( $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$ ). 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 maintaining the dependence between  $U_{it}$  and potential response.

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.

---

<sup>34</sup>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).

## 5.2 Differential Attrition Rates and Tests of Internal Validity

Table 5 reports simulation rejection probabilities 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 outcome between treatment and control respondents in the follow-up period ( $t = 1$ ),

$$\bar{Y}_1^{TR} - \bar{Y}_1^{CR} = \frac{\sum_{i=1}^n Y_{i1} D_{i1} R_i}{\sum_{i=1}^n D_{i1} R_i} - \frac{\sum_{i=1}^n Y_{i1} (1 - D_{i1}) R_i}{\sum_{i=1}^n (1 - D_{i1}) R_i}, \quad (11)$$

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),

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

where the subscript  $\mathcal{M}$  denotes the *mean* implication of the relevant distributional restriction.  $H_{0,\mathcal{M}}^{1,1}$  ( $H_{0,\mathcal{M}}^{1,2}$ ) tests the implication of the IV-R assumption for respondents (attritors) only. We present the tests of these two hypotheses since they are conceptually similar to widely used tests in the literature. The mean implication of the sharp testable restriction in Proposition 1(a.ii),  $H_{0,\mathcal{M}}^1$ , is a joint hypothesis of  $H_{0,\mathcal{M}}^{1,1}$  and  $H_{0,\mathcal{M}}^{1,2}$ . 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  $\chi^2$  statistics.<sup>35</sup> We also examine the finite-sample performance of the KS statistic of the sharp testable restriction of the IV-R assumption in (7). The reported p-values of the KS statistic defined below are obtained using the randomization procedure to test  $H_0^1$  from Section 4,

$$KS_{n,m}^1 = \max\{KS_{n,0}^1, KS_{n,1}^1\}, \text{ where for } r = 0, 1$$

$$KS_{n,r}^1 = \max_{i: R_i=r} \left| \sqrt{n} (F_{n,Y_{i0}}(y_{i0}|T_i = 1, R_i = r) - F_{n,Y_{i0}}(y_{i0}|T_i = 0, R_i = r)) \right|. \quad (13)$$

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 substantially higher than the test based on the difference between the treatment and control respondents (CR-TR).<sup>36</sup>

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 (8). Its mean version is defined as follows

$$H_{0,\mathcal{M}}^2 : E[Y_{i0}|T_i = \tau_j, R_i = r_j] = E[Y_{i0}|T_i = \tau_{j+1}, R_i = r_{j+1}] \text{ for } j = 1, \dots, |\mathcal{T} \times \mathcal{R}| - 1, \quad (14)$$

where  $(\tau_j, r_j)$  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  $\chi^2$  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

---

<sup>35</sup>To implement the test in R, we use the *linearHypothesis* command in the AER package.

<sup>36</sup>This may be because the treatment-only responders are proportionately larger in the control attritor subgroup than in the treatment respondent subgroup.



given below

$$KS_n^2 = \max_{j=1,2,3} KS_{n,j}^2, \text{ where} \quad (15)$$

$$KS_{n,j}^2 = \max_{i:(T_i, R_i) \in \{(\tau_j, r_j), (\tau_{j+1}, r_{j+1})\}} \left| \sqrt{n} (F_{n, Y_{i0} | T_i = \tau_j, R_i = r_j} - F_{n, Y_{i0} | T_i = \tau_{j+1}, R_i = r_{j+1}}) \right|.$$

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  $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$ , 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 D in the online appendix.

## 6 Empirical Applications

To complement the simulations presented above, we apply the proposed tests 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. In this case, the data comes from a limited selection of articles with both high attrition rates and publicly available data that includes attritors. Thus, the exercise is not intended to draw inference about implications of applying various attrition tests to a representative sample of published field experiments.<sup>37</sup>

For this application, we identified 47 articles that had publicly available analysis files from the 88 articles in our review (see Section 2). In order to select the five articles that had

---

<sup>37</sup>It is worth noting that, 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.

the highest attrition rates from that group, we reviewed the data files for twelve articles. We were unable to include field experiments for a variety of reasons that would not, in the majority of cases, affect the ability of the authors to implement our tests.<sup>38</sup> In keeping with our findings from Section 2, even within these five articles for which attrition bias is likely to have received some additional scrutiny given high attrition rates, we find that there is heterogeneity in the application of attrition tests. Two of the articles reported only a differential attrition rate test, while three also reported some type of selective attrition test.

Across the five selected articles, 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.<sup>39</sup> For each outcome, the approach to implementing the tests depends on the outcome and the type of randomization used in the article. 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.<sup>40</sup> For continuous outcomes in non-clustered experiments, we report p-values of the KS distributional tests using the appropriate randomization procedure.<sup>41</sup> For binary outcomes and also for all outcomes from clustered experiments, we apply regression-based mean tests (see Section B).

In addition to the tests of the restrictions in Propositions 1 and 2, we also apply a version of the tests commonly used in the literature, including: the differential attrition rate test, the IV-R test for respondents only and the IV-R test for attritors only. In the case of the IV-R tests for respondents and attritors only, we apply the same approaches to handling stratification and continuous outcomes as we do in implementing our proposed joint tests. This ensures that the three IV-R tests are directly comparable, but it also means that this exercise is not intended to be a replication of the attrition tests that are used in published field experiments. 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.

---

<sup>38</sup>Of the seven experiments that were excluded: two did not provide the data sets along with the analysis files due to confidentiality restrictions, two provided the data sets but did not include attritors, and one did not provide sufficient information to identify the attritors. In two cases, an exceptionally high number of missing values at baseline was the limiting factor since the attrition rate at follow-up conditional on baseline response was lower than the attrition rate reported in the paper.

<sup>39</sup>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.

<sup>40</sup>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.

<sup>41</sup>We apply the [Dufour \(2006\)](#) randomization procedure to accommodate the possibility of ties.

Table 6 reports the p-values of the attrition tests for the applications.<sup>42</sup> For the differential attrition rate test, we find p-values smaller than 5% for 9 out of 33 outcomes. This is 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).

Turning to the proposed joint IV-R test, however, all of the reported p-values are larger than 5%. For any of the outcomes reported here, a researcher using this test at the 5% significance level would 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 attriters 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.

Finally, we consider the results of our proposed IV-P test. For 8 out of the 33 outcomes, the differential attrition rate test has a p-value smaller than 5%, whereas the p-value of the IV-P test is substantially larger than 5%. These empirical cases are consistent with the testable implications of *Example 1*. This provides suggestive evidence that the theoretical conditions under which the differential attrition rate test does not control size are empirically relevant. More broadly, the p-values for the IV-P test are larger than 5% for a majority of outcomes in this exercise, specifically 25 out of 33. This surprising result has 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 objects 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 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

---

<sup>42</sup>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.

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 bias (Lee, 2009; Behagel et al., 2015; Millán and Macours, 2017), 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. 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. This paper also contributes to the ongoing debate about the value of collecting baseline data in field experiments. 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 Outcome between Treatment & Control Respondents ( $\bar{y}_i^{TR} - \bar{y}_i^{CR}$ )			
			Mean Tests			KS Test			Mean Test	KS Test	Mean	SD	$\hat{p}_{0.05}$	
			CR-TR	CA-TA	Joint	Joint	Joint	Joint						
														(4)
C	T	$\hat{p}_{0.05}$												
(1)	(2)	(3)	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	0.997	
	0.10	0.05	0.995	0.076	0.719	0.635	0.582	0.792	0.787	0.282	0.058	0.998	0.998	
	0.15	0.10	0.935	0.072	0.631	0.542	0.483	0.995	0.980	0.288	0.061	0.997	0.997	
	0.20	0.15	0.867	0.072	0.532	0.442	0.412	1.000	1.000	0.296	0.063	0.996	0.996	
	0.30	0.20	1.000	0.141	0.894	0.851	0.801	1.000	1.000	0.334	0.066	0.999	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	0.993	
	0.10	0.10	0.053	0.043	0.045	0.045	0.056	1.000	0.999	0.262	0.060	0.991	0.991	
	0.15	0.15	0.052	0.043	0.049	0.052	0.055	1.000	1.000	0.271	0.062	0.992	0.992	
	0.20	0.20	0.049	0.045	0.047	0.050	0.050	1.000	1.000	0.280	0.064	0.990	0.990	
	0.30	0.30	0.048	0.053	0.044	0.046	0.043	1.000	1.000	0.303	0.068	0.991	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	0.990	
	0.10	0.05	0.995	0.055	0.050	0.055	0.046	0.053	0.055	0.248	0.059	0.985	0.985	
	0.15	0.10	0.935	0.057	0.052	0.053	0.045	0.053	0.059	0.247	0.061	0.983	0.983	
	0.20	0.15	0.867	0.058	0.047	0.053	0.046	0.048	0.048	0.247	0.063	0.974	0.974	
	0.30	0.20	1.000	0.057	0.053	0.052	0.043	0.049	0.048	0.248	0.066	0.964	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	0.997	
	0.10	0.10	0.013	0.131	0.708	0.653	0.577	0.708	0.582	0.302	0.059	0.999	0.999	
	0.15	0.15	0.007	0.248	0.873	0.855	0.758	0.888	0.792	0.333	0.061	0.999	0.999	
	0.20	0.20	0.004	0.422	0.934	0.951	0.859	0.970	0.913	0.367	0.063	0.999	0.999	
	0.30	0.30	0.001	0.797	0.990	0.997	0.974	0.999	0.998	0.452	0.067	1.000	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 (12) ((14)). The KS statistics of the IV-R (IV-P) assumption are given in (13) ((15)), 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 (11). All tests are conducted using  $\alpha = 0.05$ . Additional details of the design are provided in Table 4.

<sup>†</sup> (\*) 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
	Control (%)	Differential (percentage points)		CR-TR	CA-TA	Joint	Joint
1	[10 - 30]	(10 - 20]	0.025	0.567	0.948	0.832	0.563
2	[10 - 30]	(0 - 5]	0.887	0.514	0.546	0.571	0.600
3	[10 - 30]	(0 - 5]	0.109	0.834	0.751	0.879	0.956
4	[10 - 30]	(0 - 5]	0.486	0.351	0.701	0.576	0.000
5	[10 - 30]	(0 - 5]	0.100	0.421	0.526	0.668	0.755
6	[10 - 30]	(0 - 5]	0.086	0.392	0.098	0.187	0.313
7	[10 - 30]	(0 - 5]	0.056	0.315	0.575	0.490	0.652
8	[10 - 30]	(0 - 5]	0.027	0.359	0.381	0.537	0.679
9	[10 - 30]	(0 - 5]	0.129	0.190	0.532	0.312	0.008
10	[30 - 50]	(0 - 5]	0.301	0.202	0.191	0.198	0.002
11	[10 - 30]	(0 - 5]	0.030	0.688	0.966	0.917	0.979
12	[10 - 30]	(0 - 5]	0.955	0.120	0.114	0.250	0.000
13	[10 - 30]	(10 - 20]	0.039	0.827	0.120	0.277	0.441
14	[10 - 30]	(0 - 5]	0.788	0.861	0.194	0.423	0.525
15	[10 - 30]	(10 - 20]	0.048	0.682	0.558	0.800	0.609
16	[10 - 30]	(0 - 5]	0.798	0.802	0.180	0.404	0.590
17	[10 - 30]	(10 - 20]	0.037	0.685	0.428	0.711	0.843
18	[10 - 30]	(0 - 5]	0.784	0.833	0.169	0.384	0.546
19	[30 - 50]	(0 - 5]	0.127	0.700	0.494	0.690	0.010
20	[30 - 50]	(0 - 5]	0.241	0.605	0.476	0.720	0.697
21	[10 - 30]	(0 - 5]	0.084	0.796	0.261	0.518	0.671
22	[30 - 50]	(0 - 5]	0.218	0.748	0.183	0.385	0.022
23	[30 - 50]	(0 - 5]	0.128	0.328	0.632	0.615	0.053
24	[30 - 50]	(0 - 5]	0.134	0.133	0.976	0.337	0.528
25	[30 - 50]	(0 - 5]	0.118	0.718	0.510	0.707	0.029
26	[30 - 50]	(0 - 5]	0.348	0.663	0.370	0.691	0.807
27	[30 - 50]	(0 - 5]	0.217	0.883	0.768	0.858	0.423
28	[10 - 30]	(0 - 5]	0.061	0.218	0.986	0.518	0.609
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
31	[30 - 50]	(10 - 20]	0.047	0.144	0.440	0.526	0.692
32	[10 - 30]	(0 - 5]	0.867	0.580	0.509	0.798	0.720
33	[10 - 30]	(5 - 10]	0.437	0.421	0.887	0.683	0.447

*Notes:* The table reports  $p$ -values for the differential attrition rate test as well as tests of the IV-R and IV-P assumptions.  $CR - TR$  ( $CA - TA$ ) 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.

## 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, Liu and Shi \(2019\)](#), we show that the testable restriction is sharp by showing that if  $(Y_{i0}, Y_{i1}, T_i, R_i)$  satisfy  $Y_{i0}|T_i = 0, R_i = r \stackrel{d}{=} Y_{i0}|T_i = 1, R_i = r$  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  $\mu_t$ , 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  $F_{Y_{i0}(1), Y_{i1}(0), Y_{i1}(1)|Y_{i0}, T_i=0, R_i=1} = F_{Y_{i0}(1), Y_{i1}(0), Y_{i1}|Y_{i0}, T_i=1, R_i=1}$  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  $d = 0, 1$ , 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  $\mu_t(d, \cdot)$  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.  $\square$

*Proof.* (Proposition 2) The proof is immediate from the proof of Proposition 1 by conditioning all statements on  $S_i$ .  $\square$

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

$$F_{U_i|T_i, R_i(0), R_i(1)} = F_{U_i|T_i, \xi(0, V_i), \xi(1, V_i)} = F_{U_i|\xi(0, V_i), \xi(1, V_i)} = F_{U_i|R_i(0), R_i(1)}. \quad (16)$$

As a result,

$$F_{U_i|T_i=1, R_i=1} = \frac{p_{01}F_{U_i|(R_i(0), R_i(1))=(0,1)} + p_{11}F_{U_i|(R_i(0), R_i(1))=(1,1)}}{P(R_i = 1|T_i = 1)}, \quad (17)$$

$$F_{U_i|T_i=0, R_i=1} = \frac{p_{10}F_{U_i|(R_i(0), R_i(1))=(1,0)} + p_{11}F_{U_i|(R_i(0), R_i(1))=(1,1)}}{P(R_i = 1|T_i = 0)}. \quad (18)$$

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 (17) and (18) 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,

$$\begin{aligned} F_{U_i|T_i=1, R_i=1} &= \frac{p_{01}F_{U_i|(R_i(0), R_i(1))=(0,1)} + p_{11}F_{U_i|(R_i(0), R_i(1))=(1,1)}}{P(R_i = 1|T_i = 1)} \\ &= \frac{p_{10}F_{U_i|(R_i(0), R_i(1))=(1,0)} + p_{11}F_{U_i|(R_i(0), R_i(1))=(1,1)}}{P(R_i = 1|T_i = 0)} = F_{U_i|T_i=0, R_i=1}. \end{aligned} \quad (19)$$



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.  $\square$

## B Regression Tests of Internal Validity

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

### B.1 Completely and Clustered Randomized Experiments

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

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

### B.2 Stratified Randomized Experiments

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

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

$$\begin{aligned} H_{0,\mathcal{M}}^{1,s} : \gamma_{11}^s &= \gamma_{01}^s \text{ \& } \gamma_{10}^s = \gamma_{00}^s, \\ H_{0,\mathcal{M}}^{2,s} : \gamma_{11}^s &= \gamma_{01}^s = \gamma_{10}^s = \gamma_{00}^s. \end{aligned}$$

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

## References

- Abadie, Alberto, Matthew M. Chingos, and Martin R. West.** 2018. “Endogenous Stratification in Randomized Experiments.” *Review of Economics and Statistics*, 100(4): 567–580.
- Ahn, Hyungtaik, and James L. Powell.** 1993. “Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism.” *Journal of Econometrics*, 58(1): 3–29.
- Altonji, Joseph, and Rosa Matzkin.** 2005. “Cross-section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors.” *Econometrica*, 73(3): 1053–1102.
- Andrews, Isaiah, and Emily Oster.** 2019. “A simple approximation for evaluating external validity bias.” *Economics Letters*, 178: 58 – 62.
- Angrist, Joshua D.** 1997. “Conditional Independence in Sample Selection Models.” *Economics Letters*, 54(2): 103 – 112.
- Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin.** 1996. “Identification of Causal Effects Using Instrumental Variables.” *Journal of the American Statistical Association*, 91(434): 444–455.
- Athey, S., and G.W. Imbens.** 2017. “Chapter 3 - The Econometrics of Randomized Experiments.” In *Handbook of Field Experiments*. Vol. 1 of *Handbook of Economic Field Experiments*, , ed. Abhijit Vinayak Banerjee and Esther Duflo, 73 – 140. North-Holland.
- Athey, Susan, Dean Eckles, and Guido W. Imbens.** 2018. “Exact p-Values for Network Interference.” *Journal of the American Statistical Association*, 113(521): 230–240.
- Azzam, Tarek, Michael Bates, and David Fairris.** 2018. “Do Learning Communities Increase First Year College Retention? Testing the External Validity of Randomized Control Trials.” Unpublished.
- Baird, Sarah, J. Aislinn Bohren, Craig McIntosh, and Berk Özler.** 2018. “Optimal Design of Experiments in the Presence of Interference.” *Review of Economics and Statistics*, 100(5): 844–860.
- Behagel, Luc, Bruno Crépon, Marc Gurgand, and Thomas Le Barbanchon.** 2015. “Please Call Again: Correcting Nonresponse Bias in Treatment Effect Models.” *Review of Economics and Statistics*, 97: 1070–1080.
- Bester, C. Alan, and Christian Hansen.** 2009. “Identification of Marginal Effects in a Nonparametric Correlated Random Effects Model.” *Journal of Business and Economic Statistics*, 27(2): 235–250.
- Brownstone, David.** 1998. “Multiple Imputation Methodology For Missing Data, Non-Random Response, And Panel Attrition.” In *Theoretical Foundations of Travel Choice Modeling*, , ed. T. Gärling, T. Laitila and K. Westin, 421–450. Elsevier.

- Bruhn, Miriam, and David McKenzie.** 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics*, 1(4): 200–232.
- Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh.** 2018. "Inference Under Covariate-Adaptive Randomization." *Journal of the American Statistical Association*, 113(524): 1784–1796.
- Canay, Ivan A., Joseph P. Romano, and Azeem M. Shaikh.** 2017. "Randomization Tests Under an Approximate Symmetry Assumption." *Econometrica*, 85(3): 1013–1030.
- Chernozhukov, Victor, Ivan Fernandez-Val, Jinyong Hahn, and Whitney Newey.** 2013. "Average and Quantile Effects in Nonseparable Panel Data Models." *Econometrica*, 81(2): pp.535–580.
- Das, Mitali, Whitney K. Newey, and Francis Vella.** 2003. "Nonparametric Estimation of Sample Selection Models." *Review of Economic Studies*, 70(1): 33–58.
- de Chaisemartin, Clément.** 2017. "Tolerating Defiance? Local Average Treatment Effects Without Monotonicity." *Quantitative Economics*, 8(2): 367–396.
- de Chaisemartin, Clément, and Luc Behaghel.** 2018. "Estimating the Effect of Treatments Allocated by Randomized Waiting Lists." arXiv.org Papers 1511.01453.
- Dufour, Jean-Marie.** 2006. "Monte Carlo Tests with Nuisance Parameters: A General approach to Finite-Sample Inference and Nonstandard Asymptotics." *Journal of Econometrics*, 133(2): 443 – 477.
- Dufour, Jean-Marie, Abdeljelil Farhat, Lucien Gardiol, and Lynda Khalaf.** 1998. "Simulation-based Finite Sample Normality Tests in Linear Regressions." *Econometrics Journal*, 1(1): 154–173.
- Ghanem, Dalia.** 2017. "Testing Identifying Assumptions in Nonseparable Panel Data Models." *Journal of Econometrics*, 197: 202–217.
- Glennerster, Rachel, and Kudzai Takavarasha.** 2013. *Running Randomized Evaluations: A Practical Guide*. . Student Edition ed., Princeton University Press.
- Hausman, Jerry A., and David A. Wise.** 1979. "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment." *Econometrica*, 47(2): 455–473.
- Heckman, James J.** 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." In *Annals of Economic and Social Measurement*. Vol. 5, , ed. Sanford V. Berg, 475–492. National Bureau of Economic Research.
- Heckman, James J.** 1979. "Sample Selection Bias as A Specification Error." *Econometrica*, 47(1): 153–161.

- Hirano, Keisuke, Guido W. Imbens, Geert Ridder, and Donald B. Rubin.** 2001. "Combining Panel Data Sets with Attrition and Refreshment Samples." *Econometrica*, 69(6): 1645–1659.
- Hoderlein, Stefan, and Halbert White.** 2012. "Nonparametric Identification of Non-separable Panel Data Models with Generalized Fixed Effects." *Journal of Econometrics*, 168(2): 300–314.
- Horowitz, Joel L., and Charles F. Manski.** 2000. "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data." *Journal of the American Statistical Association*, 95(449): 77–84.
- Hsu, Yu-Chin, Chu-An Liu, and Xiaoxia Shi.** 2019. "Testing Generalized Regression Monotonicity." *Econometric Theory*, 1 – 55.
- IES.** 2017. "What Works Clearinghouse. Standards Handbook Version 4.0." U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.
- Imbens, Guido W., and Donald B. Rubin.** 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.
- Imbens, Guido W., and Joshua D. Angrist.** 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.
- INSP.** 2005. "General Rural Methodology Note." Instituto Nacional de Salud Publica.
- Kitagawa, Toru.** 2015. "A Test for Instrument Validity." *Econometrica*, 83(5): 2043–2063.
- Kline, Patrick, and Andres Santos.** 2013. "Sensitivity to Missing Data Assumptions: Theory and An Evaluation of The U.S. Wage Structure." *Quantitative Economics*, 4(2): 231–267.
- Lee, David S.** 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies*, 76(3): 1071–1102.
- Lehmann, E. L., and Joseph P. Romano.** 2005. *Testing Statistical Hypotheses*. . Third ed., New York:Springer.
- Manski, Charles F.** 2005. "Partial Identification with Missing Data: Concepts and Findings." *International Journal of Approximate Reasoning*, 39(2): 151 – 165.
- McKenzie, David.** 2012. "Beyond Baseline and Follow-Up: The Case for More T in Experiments." *Journal of Development Economics*, 99(2): 210–221.
- McKenzie, David.** 2019. "Attrition Rates Typically Aren't that Different for The Control Group than The Treatment Group – Really? and Why?" *Development Impact Blog*. <https://blogs.worldbank.org/impactevaluations/attrition-rates-typically-arent-different-control-group-treatment-group-really-and-why>.

- Millán, Teresa Molina, and Karen Macours.** 2017. “Attrition in Randomized Control Trials: Using Tracking Information to Correct Bias.” IZA Discussion Paper No. 10711.
- Mourifié, Ismael, and Yuanyuan Wan.** 2017. “Testing Local Average Treatment Effect Assumptions.” *Review of Economics and Statistics*, 99(2): 305–313.
- Rubin, Donald B.** 1976. “Inference and Missing Data.” *Biometrika*, 63(3): 581–592.
- Skoufias, Emmanuel.** 2005. “PROGRESA and Its Impacts on The Welfare of Rural households in Mexico.” International Food Policy Research Institute (IFPRI) Research Report 139.
- Vella, Francis.** 1998. “Estimating Models with Sample Selection Bias: A Survey.” *Journal of Human Resources*, 33(1): 127–169.
- Wooldridge, Jeffrey M.** 1995. “Selection corrections for panel data models under conditional mean independence assumptions.” *Journal of Econometrics*, 68(1): 115 – 132.
- Wooldridge, Jeffrey M.** 2007. “Inverse Probability Weighted Estimation for General Missing Data Problems.” *Journal of Econometrics*, 141(2): 1281 – 1301.
- Young, Alwyn.** 2018. “Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results\*.” *Quarterly Journal of Economics*, 134(2): 557–598.

# Testing Attrition Bias in Field Experiments

## Online Appendix

*For Online Publication Only*

Dalia Ghanem  
UC Davis

Sarojini Hirshleifer  
UC Riverside

Karen Ortiz-Becerra  
UC Davis

August 6, 2019

### Contents

<b>A Details on the Review of the Field Experiment Literature</b>	<b>2</b>
A.1 Selection of Papers . . . . .	2
A.2 Attrition Tests . . . . .	3
A.2.1 Differential Attrition Rate Test . . . . .	4
A.2.2 Selective Attrition Test . . . . .	4
A.2.3 Determinants of Attrition Test . . . . .	4
<b>B Equal Attrition Rates with Multiple Treatment Groups</b>	<b>5</b>
<b>C Identification and Testing for the Multiple Treatment Case</b>	<b>5</b>
C.1 Identification and Sharp Testable Restrictions . . . . .	5
C.1.1 Completely Randomized Trials . . . . .	5
C.1.2 Stratified Randomized Trials . . . . .	7
C.2 Distributional Test Statistics . . . . .	7
C.2.1 Completely Randomized Trials . . . . .	8
C.2.2 Stratified Randomized Trials . . . . .	8
<b>D Extended Simulations for the Distributional Tests</b>	<b>9</b>
D.1 Comparing Different Statistics of the Distributional Hypotheses . . . . .	9
D.2 Additional Variants of the Simulation Designs . . . . .	10
<b>E List of Papers Included in the Review of Field Experiments</b>	<b>15</b>

### List of Tables

A.1 Distribution of Articles by Journal and Year of Publication . . . . .	2
D.1 Simulation Results on the KS & CM Randomization Test of IV-R . . . . .	12
D.2 Simulation Results on the KS & CM Randomization Test of IV-P . . . . .	13

### List of Figures

D.1 Additional Simulation Analysis for the KS Statistics of Internal Validity . . . . .	14
---	----

## A Details on the Review of the Field Experiment Literature

### A.1 Selection of Papers

In order to understand both the extent of attrition as well as how authors test for attrition bias in practice, we systematically reviewed articles that report the results of field experiments. We include articles that were published in the top five journals in economics, as well as four highly regarded applied economics journals: *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*.<sup>1</sup> By searching for *RCT*, *randomized controlled trial*, or *field experiment* in each journal’s website, we identified 191 articles that were published between 2009 and 2015.<sup>2</sup> From these 191 articles, we review those that satisfied three main criteria: i) the main goal of the article is to report the results of an intervention, ii) the study design is such that attrition is relevant, and iii) there is baseline data on at least one outcome.

Table A.1 displays the distribution of the 88 articles that satisfied the selection criteria by journal and year of publication. Since some of the articles report results for more than one intervention, these articles report the results of 91 field experiments. Of the included articles, 65% were published in the *Journal of Development Economics*, the *American Economic Journal: Applied Economics*, and the *Quarterly Journal of Economics*. Approximately 57% of our sample of articles was published in 2014 and 2015.

Table A.1: Distribution of Articles by Journal and Year of Publication

Journal	Year							Total
	2009	2010	2011	2012	2013	2014	2015	
AEJ: Applied	0	0	0	3	3	3	8	17
AER	0	1	1	2	0	2	2	8
EJ	0	0	1	2	0	5	0	8
Econometrica	1	0	0	0	0	1	0	2
JDE	0	0	1	1	3	11	6	22
JPE	0	0	1	0	0	0	0	1
QJE	1	1	4	3	2	4	3	18
REstat	2	0	2	1	1	1	3	10
REstud	0	0	0	0	1	1	0	2
Total	4	2	10	12	10	28	22	88

*Notes:* The 88 articles that we include in our review correspond to 91 field experiments. The two articles that reported more than one field experiment are published in the AER(2015) and the QJE(2011), respectively.

<sup>1</sup>We chose those four applied journals because they are important sources of published field experiments.

<sup>2</sup>Our initial search yielded 199 articles but 8 of those were not considered since they were observational studies exploiting some sort of quasi-experimental variation.

From the 103 articles that were excluded from our review, 14 had a primary goal that was different from reporting the impact of an intervention. Some articles developed new econometric techniques and used a field experiment in their empirical application, other used the random allocation of different surveys to test for the best approach to elicit information, and a couple used data from a previously studied experiment to test alternative counterfactual policies.<sup>3</sup> In addition to these articles, the other 23 were excluded because they did not have a study design for which attrition is relevant. Specifically, their primary data comes either from lab experiments in the field (that usually take place over very short period of time) or from repeated cross-sections.

The remaining 66 articles that were excluded did not have available baseline data for any of the outcomes reported in the abstract. Most of these studies did not collect baseline outcomes by design because they were irrelevant for the target population.<sup>4</sup> There were also some studies that either did not collect baseline data at all, had baseline data for variables that are not a good approximation of the endline outcome (i.e. scores of different tests that were not normalized), or had attrition rates at baseline above 50%.

One challenge that arose in our review was determining which attrition rates and attrition tests are most relevant, since the reported attrition rates usually vary across different data sources or different subsamples. We chose to focus on the results that are reported in the abstract in our analysis of attrition rates. But, since many authors do not report attrition tests for each of the abstract results, in our analysis of attrition tests we focus on whether authors report a test that is relevant to at least one abstract result.

## A.2 Attrition Tests

In order to classify the attrition tests that are conducted in the 88 articles that we review, we gathered information on the different econometric strategies that are carried out to test for attrition bias. In this section, we describe these empirical strategies and classify them into the three different types of tests. We use the following notation to facilitate the exposition of each strategy and the comparison across them:

- Let  $R_i$  take the value of 1 if individual  $i$  belongs to the follow-up sample.
- Let  $T_i$  take the value of 1 if individual  $i$  belongs to the treatment group.
- Let  $X_{i0}$  be a  $k \times 1$  vector of baseline variables.
- Let  $Y_{i0}$  be a  $l \times 1$  vector of outcomes collected at baseline.
- Let  $Z_{i0} = (X'_{i0}, Y'_{i0})'$ , and  $Z_{i0}^j$  be the  $j^{th}$  element of  $Z_{i0}$ .

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

---

<sup>3</sup>Our review does include articles that reported treatment effects of a program for the first time and used a structural model for additional analysis.

<sup>4</sup>A common example are those studies interested in take-up using a sample of individuals that did not have the product at baseline.



### A.2.1 Differential Attrition Rate Test

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

1.  $t$ -test of the equality of attrition rate by treatment group, i.e.  $H_0 : P(R_i = 0|T_i = 1) = P(R_i = 0|T_i = 0)$ .
2.  $R_i = \gamma + T_i\beta + U_i$ ; may include strata fixed effects.
3.  $R_i = \gamma + T_i\beta + X'_{i0}\theta + Y'_{i0}\alpha + U_i$ ; may include strata fixed effects.

### A.2.2 Selective Attrition Test

The *selective attrition test* determines whether, conditional on response status, the distribution of observable characteristics is the same across treatment and control groups. We identify two sub-types of selective attrition tests: i) a simple test conducted only on respondents or attritors, and ii) a joint test conducted on both respondents and attritors.

#### Simple tests:

1.  $t$ -test of baseline characteristics by treatment group among respondents, i.e.  $H_0^j : E[Z_{i0}^j|T_i = 1, R_i = 1] = E[Z_{i0}^j|T_i = 0, R_i = 1]$  for  $j = 1, 2, \dots, (l + k)$ .
2.  $T_i = \gamma + X'_{i0}\theta + Y'_{i0}\alpha + U_i$  if  $R_i = 1$ ; may include strata fixed effects.
3. Kolmogorov-Smirnov (KS) test of baseline characteristics by treatment group among respondents,  $H_0^j : F_{Z_{i0}^j|T_i, R_i=1} = F_{Z_{i0}^j|R_i=1}$  for  $j = 1, 2, \dots, (l + k)$ .
4.  $Z_{i0}^j = \gamma + T_i\beta^j + U_i^j$  if  $R_i = 1$ , for  $j = 1, 2, \dots, (l + k)$ ; may include strata fixed effects.
5.  $Z_{i0}^j = \gamma + T_i\beta^j + U_i^j$  if  $R_i = 0$ , for  $j = 1, 2, \dots, (l + k)$ ; may include strata fixed effects.

#### Joint tests:

1.  $Z_{i0}^j = \gamma^j + T_i\beta^j + (1 - R_i)\lambda^j + T_i(1 - R_i)\phi^j + U_i^j$ ; may include strata fixed effects.
2.  $R_i = \gamma + T_i\beta + X'_{i0}\theta + Y'_{i0}\alpha + T_iX'_{i0}\lambda_1 + T_iY'_{i0}\lambda_2 + U_i$ ; may include strata fixed effects.
3.  $t$ -test of the null hypothesis:  $E[Y_{i0}|T_i = 1, R_i = 1] - E[Y_{i0}|T_i = 1, R_i = 0] = E[Y_{i0}|T_i = 0, R_i = 1] - E[Y_{i0}|T_i = 0, R_i = 0]$ .

### A.2.3 Determinants of Attrition Test

The *determinants of attrition test* determines whether attritors are significantly different from respondents regardless of treatment assignment.

1.  $R_i = \gamma + T_i\beta + X'_{i0}\theta + Y'_{i0}\alpha + U_i$ ; may include strata fixed effects.
2.  $Z_{i0}^j = \gamma^j + (1 - R_i)\lambda^j + U_i^j$ ,  $j = 1, 2, \dots, (l + k)$ ; may include strata fixed effects.
3.  $R_i = \gamma + X'_{i0}\theta + Y'_{i0}\alpha + U_i$ ; may include strata fixed effects.

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

## B Equal Attrition Rates with Multiple Treatment Groups

In this section, we illustrate that once we have more than two treatment groups and violations of monotonicity, then equal attrition rates are possible without imposing the equality of proportions of certain subpopulations unlike Example 2 in the paper. Consider the case where we have three treatment groups, i.e.  $T_i \in \{0, 1, 2\}$ . For brevity, we use the notation  $P_i((r_0, r_1, r_2)) \equiv P((R_i(0), R_i(1), R_i(2)) = (r_0, r_1, r_2))$  for  $(r_0, r_1, r_2) \in \{0, 1\}^3$ . Hence,

$$\begin{aligned} P(R_i = 0|T_i = 0) &= P_i((0, 0, 0)) + P_i((0, 0, 1)) + P_i((0, 1, 0)) + P_i((0, 1, 1)) \\ P(R_i = 0|T_i = 1) &= P_i((0, 0, 0)) + P_i((0, 0, 1)) + P_i((1, 0, 0)) + P_i((1, 0, 1)) \\ P(R_i = 0|T_i = 2) &= P_i((0, 0, 0)) + P_i((1, 0, 0)) + P_i((0, 1, 0)) + P_i((1, 1, 0)) \end{aligned} \quad (1)$$

The equality of response rates across the three groups, i.e.  $P(R_i = 0|T_i = 0) - P(R_i = 0|T_i = 1) = P(R_i = 0|T_i = 0) - P(R_i = 0|T_i = 2) = 0$  implies the following equalities,

$$\begin{aligned} P_i((0, 1, 0)) + P_i((0, 1, 1)) &= P_i((1, 0, 0)) + P_i((1, 0, 1)) \\ P_i((0, 0, 1)) + P_i((0, 1, 1)) &= P_i((1, 0, 0)) + P_i((1, 1, 0)) \end{aligned} \quad (2)$$

which can occur without constraining the proportions of different subpopulations to be equal.

## C Identification and Testing for the Multiple Treatment Case

In this section, we present the generalization of Propositions 1 and 2 (Section C.1) as well as the distributional test statistics (Section C.2) in the paper to the case where the treatment variable has arbitrary finite-support. As in the paper, we provide results for completely and stratified randomized experiments. We maintain that  $D_{i0} = 0$  for all  $i$ , i.e. no treatment is assigned in the baseline period,  $D_{i1} \in \mathcal{D}$ , where wlog  $\mathcal{D} = \{0, 1, \dots, |\mathcal{D}| - 1\}$ ,  $|\mathcal{D}| < \infty$ .  $D_i \equiv (D_{i0}, D_{i1}) \in \{(0, 0), (0, 1), \dots, (0, |\mathcal{D}| - 1)\}$ . Let  $T_i$  denote the indicator for membership in the treatment group defined by  $D_i$ , i.e.  $T_i \in \mathcal{T} = \{0, 1, \dots, |\mathcal{D}| - 1\}$ , where  $T_i = D_{i1}$  and hence  $|\mathcal{T}| = |\mathcal{D}|$  by construction.

### C.1 Identification and Sharp Testable Restrictions

#### C.1.1 Completely Randomized Trials

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

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

- (i) (Identification)  $Y_{i1}|T_i = \tau, R_i = 1 \stackrel{d}{=} Y_{i1}(\tau)|R_i = 1$  for  $\tau \in \mathcal{T}$ .
- (ii) (Sharp Testable Restriction)  $Y_{i0}|T_i = \tau, R_i = r \stackrel{d}{=} Y_{i0}|T_i = \tau', R_i = r$  for  $r = 0, 1$ , for  $\tau, \tau' \in \mathcal{T}, \tau \neq \tau'$ .

(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  $\tau \in \mathcal{T}$ .

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

**Proof** (Proposition C.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  $d \in \mathcal{D}$ ,  $F_{Y_{it}(d) | T_i, R_i} = \int 1\{\mu_t(d, u) \leq \cdot\} dF_{U_{it} | T_i, R_i}(u) = \int 1\{\mu_t(d, u) \leq \cdot\} dF_{U_{it} | R_i}(u) = F_{Y_{it}(d) | R_i}$ . (i) follows by letting  $t = 1$  and  $d = \tau$ , while conditioning the left-hand side of the last equation on  $T_i = \tau$  and  $R_i = 1$  and the right-hand side on  $R_i = 1$ . The testable implication in (ii) follows by letting  $t = d = 0$  and conditioning the left-hand side on  $T_i = \tau$  and  $R_i = r$  and the right-hand side on  $T_i = \tau'$  and  $R_i = r$ , where  $\tau \neq \tau'$ .

Following Hsu, Liu and Shi (2019), we show that the testable restriction is sharp by showing that if  $(Y_{i0}, Y_{i1}, T_i, R_i)$  satisfy  $Y_{i0} | T_i = \tau, R_i = r \stackrel{d}{=} Y_{i0} | T_i = \tau', R_i = r$  for  $r = 0, 1, \tau, \tau' \in \mathcal{T}, \tau \neq \tau'$ , then there exists  $(U_{i0}, U_{i1})$  such that  $Y_{it}(d) = \mu_t(d, U_{it})$  for some  $\mu_t(d, \cdot)$  for  $d \in \mathcal{D}$  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  $\mu_t$ , we can let  $U'_{it} = \mathbf{Y}_{it}(\cdot) = (Y_{it}(0), Y_{it}(1), \dots, Y_{it}(|\mathcal{D}| - 1))$  and  $\mu_t(d, U_{it}) = \sum_{j=0}^{|\mathcal{D}|-1} 1\{j = d\} Y_{it}(j)$  for  $d \in \mathcal{D}, t = 0, 1$ . Note that  $Y_{i0} = Y_{i0}(0)$  since  $D_{i0} = 0$  w.p.1. Now we have to construct a distribution of  $U_i = (U'_{i0}, U'_{i1})$  that satisfies

$$F_{U_i | T_i, R_i} \equiv F_{\mathbf{Y}_{i0}(\cdot), \mathbf{Y}_{i1}(\cdot) | T_i, R_i} = F_{\mathbf{Y}_{i0}(\cdot), \mathbf{Y}_{i1}(\cdot) | R_i}$$

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 respondents in the different treatment groups

$$\begin{aligned} F_{U_i | T_i = \tau, R_i = 1} &= F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1}, \mathbf{Y}_{i1}(\cdot) | Y_{i0}, T_i = \tau, R_i = 1} F_{Y_{i0} | T_i = \tau, R_i = 1} \\ &= F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1}, \{Y_{i1}(d)\}_{d=0}^{\tau-1}, Y_{i1}, \{Y_{i1}(d)\}_{d=\tau+1}^{|\mathcal{D}|-1} | Y_{i0}, T_i = \tau, R_i = 1} F_{Y_{i0} | T_i = \tau, R_i = 1}. \end{aligned} \quad (3)$$

By construction,  $F_{Y_{i0} | T_i, R_i = 1} = F_{Y_{i0} | R_i = 1}$ . Now generating the above distribution for all  $\tau \in \mathcal{T}$  such that  $F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1}, \{Y_{i1}(d)\}_{d=0}^{\tau-1}, Y_{i1}, \{Y_{i1}(d)\}_{d=\tau+1}^{|\mathcal{D}|-1} | Y_{i0}, T_i = \tau, R_i = 1}$  which satisfies the following equality  $\forall \tau, \tau' \in \mathcal{T}, \tau \neq \tau'$ ,

$$\begin{aligned} &F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1}, \{Y_{i1}(d)\}_{d=0}^{\tau-1}, Y_{i1}, \{Y_{i1}(d)\}_{d=\tau+1}^{|\mathcal{D}|-1} | Y_{i0}, T_i = \tau, R_i = 1} \\ &= F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1}, \{Y_{i1}(d)\}_{d=0}^{\tau'-1}, Y_{i1}, \{Y_{i1}(d)\}_{d=\tau'+1}^{|\mathcal{D}|-1} | Y_{i0}, T_i = \tau', R_i = 1}, \end{aligned}$$

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,

$$F_{U_i | T_i = \tau, R_i = 0} = F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1}, \mathbf{Y}_{it}(\cdot) | Y_{i0}, T_i = \tau, R_i = 0} F_{Y_{i0} | T_i = \tau, R_i = 0} \quad (4)$$

Since  $F_{Y_{i0} | T_i, R_i = 0} = F_{Y_{i0} | R_i = 0}$  by construction, it remains to generate the above distribution for all  $\tau \in \mathcal{T}$  using the same  $F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1}, \mathbf{Y}_{it}(\cdot) | Y_{i0}, R_i = 0}$ . 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  $d \in \mathcal{D}$ ,  $F_{Y_{it}(d)|T_i, R_i}(\cdot) = \int 1\{\mu_t(d, u) \leq \cdot\} dF_{U_{it}|T_i, R_i}(u) = \int 1\{\mu_t(d, u) \leq \cdot\} dF_{U_{it}}(u) = F_{Y_{it}(d)}$ . (i) follows by letting  $d = \tau$  and  $t = 1$ , while conditioning the left-hand side of the last equation on  $T_i = \tau$  and  $R_i = 1$ , whereas (ii) follows by letting  $d = t = 0$  while conditioning on  $T_i = \tau$  and  $R_i = r$  for  $\tau \in \mathcal{T}$ ,  $r = 0, 1$ .

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  $\mu_t(d, \cdot)$  for  $d \in \mathcal{D}$  and  $t = 0, 1$  and  $(U_{i0}, U_{i1}) \perp (T_i, R_i)$ . Similar to (a.ii), we let  $U'_{it} = \mathbf{Y}_{it}(\cdot) = (Y_{it}(0), Y_{it}(1), \dots, Y_{it}(|\mathcal{D}| - 1))$  and  $\mu_t(d, U_{it}) = \sum_{j=0}^{|\mathcal{D}|-1} 1\{j = d\} Y_{it}(j)$  for  $d \in \mathcal{D}$ ,  $t = 0, 1$ . By construction,  $Y_{i0} = Y_{i0}(0)$ . Furthermore,  $F_{Y_{i0}|T_i, R_i} = F_{Y_{i0}}$  by assumption. It follows immediately that for all  $\tau \in \mathcal{T}$

$$\begin{aligned} F_{U_i|T_i=\tau, R_i=1} &= F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1}, \{Y_{i1}(d)\}_{d=0}^{\tau-1}, Y_{i1}, \{Y_{i1}(d)\}_{d=\tau+1}^{|\mathcal{D}|-1}|T_i=\tau, R_i=1} F_{Y_{i0}}, \\ F_{U_i|T_i=\tau, R_i=0} &= F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1}, \mathbf{Y}_{it}(\cdot)|Y_{i0}, T_i=\tau, R_i=0} F_{Y_{i0}}. \end{aligned}$$

Now constructing all of the above distributions using the same  $F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1}, \mathbf{Y}_{it}(\cdot)|Y_{i0}, T_i, R_i}$  that satisfies the above equalities for all  $\tau \in \mathcal{T}$  implies the result.

### C.1.2 Stratified Randomized Trials

**Proposition C.2** Assume  $(U_{i0}, U_{i1}, V_i) \perp T_i|S_i$ .

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

- (i) (Identification)  $Y_{i1}|T_i = \tau, S_i = s, R_i = 1 \stackrel{d}{=} Y_{i1}(\tau)|S_i = s, R_i = 1$ ,  
for  $\tau \in \mathcal{T}, s \in \mathcal{S}$ .
- (ii) (Sharp Testable Restriction)  $Y_{i0}|T_i = \tau, S_i = s, R_i = r \stackrel{d}{=} Y_{i0}|T_i = \tau', S_i = s, R_i = r$ ,  
 $\forall \tau, \tau' \in \mathcal{T}, \tau \neq \tau', s \in \mathcal{S}, r = 0, 1$ .

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

- (i) (Identification)  $Y_{i1}|T_i = \tau, S_i = s, R_i = 1 \stackrel{d}{=} Y_{i1}(\tau)|S_i = s$  for  $\tau \in \mathcal{T}, s \in \mathcal{S}$ .
- (ii) (Sharp Testable Restriction)  $Y_{i0}|T_i = \tau, S_i = s, R_i = r \stackrel{d}{=} Y_{i0}|S_i = s$  for  $\tau \in \mathcal{T}$ ,  
 $r = 0, 1, s \in \mathcal{S}$ .

**Proof** (Proposition C.2) The proof for this proposition follows in a straightforward manner from the proof for Proposition C.1 by conditioning all statements on  $S_i$ .

### C.2 Distributional Test Statistics

Next, we present the null hypotheses and distributional statistics for the multiple treatment case. For simplicity, we only present the joint statistics that take the maximum to aggregate over the individual statistics of each distributional equality implied by a given testable restriction.

### C.2.1 Completely Randomized Trials

The null hypothesis implied by Proposition C.1(a.ii) is given by the following,

$$H_0^{1,\mathcal{T}} : F_{Y_{i0}|T_i=\tau, R_i=r} = F_{Y_{i0}|T_i=\tau', R_i=r} \text{ for } \tau, \tau' \in \mathcal{T}, \tau \neq \tau', r = 0, 1. \quad (5)$$

Consider the following general form of the distributional statistic for the above null hypothesis is  $T_n^{1,\mathcal{T}} = \max_{r \in \{0,1\}} T_{n,r}^{1,\mathcal{T}}$ , where for  $r = 0, 1$ ,

$$T_{n,r}^{1,\mathcal{T}} = \max_{(\tau, \tau') \in \mathcal{T}^2: \tau \neq \tau'} \left\| \sqrt{n} (F_{n,Y_{i0}|T_i=\tau, R_i=r} - F_{n,Y_{i0}|T_i=\tau', R_i=r}) \right\|.$$

The randomization procedure proposed in the paper using the transformations  $\mathcal{G}_0^1$  can be used to obtain p-values for the above statistic under  $H_0^{1,\mathcal{T}}$ .

Let  $(\tau, r) \in \mathcal{T} \times \mathcal{R}$ , where  $\mathcal{R} = \{0, 1\}$ . Let  $(\tau_j, r_j)$  denote the  $j^{\text{th}}$  element of  $\mathcal{T} \times \mathcal{R}$ , then the null hypothesis implied by Proposition C.1(b.ii) is given by the following:

$$H_0^{2,\mathcal{T}} : F_{Y_{i0}|T_i=\tau_j, R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1}, R_i=r_{j+1}} \text{ for } j = 1, \dots, |\mathcal{T} \times \mathcal{R}| - 1. \quad (6)$$

the test statistic for the above *joint* hypothesis is given by

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

The randomization procedure proposed in the paper using the transformations  $\mathcal{G}_0^2$  can be used to obtain p-values for the above statistic under  $H_0^{2,\mathcal{T}}$ .

### C.2.2 Stratified Randomized Trials

The null hypothesis implied by Proposition C.2(a.ii) is given by the following,

$$H_0^{1,\mathcal{S},\mathcal{T}} : F_{Y_{i0}|T_i=\tau, S_i=s, R_i=r} = F_{Y_{i0}|T_i=\tau', S_i=s, R_i=r} \text{ for } \tau, \tau' \in \mathcal{T}, \tau \neq \tau', s \in \mathcal{S}, r = 0, 1. \quad (7)$$

Consider the following general form of the distributional statistic for the above null hypothesis is  $T_n^{1,\mathcal{S},\mathcal{T}} = \max_{s \in \mathcal{S}} \max_{r \in \{0,1\}} T_{n,r,s}^{1,\mathcal{T}}$ , where for  $s \in \mathcal{S}$  and  $r = 0, 1$ ,

$$T_{n,r,s}^{1,\mathcal{T}} = \max_{(\tau, \tau') \in \mathcal{T}^2: \tau \neq \tau'} \left\| \sqrt{n} (F_{n,Y_{i0}|T_i=\tau, S_i=s, R_i=r} - F_{n,Y_{i0}|T_i=\tau', S_i=s, R_i=r}) \right\|.$$

The randomization procedure proposed in the paper using the transformations  $\mathcal{G}_0^{1,\mathcal{S}}$  can be used to obtain p-values for  $T_n^{1,\mathcal{S},\mathcal{T}}$  under  $H_0^{1,\mathcal{S},\mathcal{T}}$ .

Let  $(\tau, r) \in \mathcal{T} \times \mathcal{R}$ . Let  $(\tau_j, r_j)$  denote the  $j^{\text{th}}$  element of  $\mathcal{T} \times \mathcal{R}$ , then the null hypothesis implied by Proposition C.2(b.ii) is given by the following:

$$H_0^{2,\mathcal{S},\mathcal{T}} : F_{Y_{i0}|T_i=\tau_j, S_i=s, R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1}, S_i=s, R_i=r_{j+1}} \text{ for } j = 1, \dots, |\mathcal{T} \times \mathcal{R}| - 1, s \in \mathcal{S}. \quad (8)$$

the test statistic for the above *joint* hypothesis is given by

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

The randomization procedure proposed in the paper using the transformations  $\mathcal{G}_0^{2,\mathcal{S}}$  can be used to obtain p-values for the above statistic under  $H_0^{2,\mathcal{S},\mathcal{T}}$ .

## D Extended Simulations for the Distributional Tests

### D.1 Comparing Different Statistics of the Distributional Hypotheses

In this section, we examine the finite-sample performance of a wider variety of the distributional tests of the IV-R and IV-P assumptions provided in Section 5 of the paper. We specifically consider the Kolmogorov-Smirnov (KS) and Cramer-von-Mises (CM) statistics of the simple and joint hypotheses. For the joint hypotheses, we include the probability weighted statistic in addition to the version used in the paper.

For the IV-R assumption, consider the following hypotheses implied by Proposition 1(b.ii) in the paper

$$\begin{aligned} H_0^{1,1} : Y_{i0}|T_i = 1, R_i = 0 &\stackrel{d}{=} Y_{i0}|T_i = 0, R_i = 0, & (CA - TA) \\ H_0^{1,2} : Y_{i0}|T_i = 1, R_i = 1 &\stackrel{d}{=} Y_{i0}|T_i = 0, R_i = 1, & (CR - TR) \\ H_0^1 : H_0^{1,1} \text{ \& } H_0^{1,2}. & & (Joint) \end{aligned} \quad (9)$$

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

$$\begin{aligned} KS_{n,r}^1 &= \max_{i: R_i=r} \left| \sqrt{n} (F_{n,Y_{i0}}(y_{i0}|T_i = 1, R_i = r) - F_{n,Y_{i0}}(y_{i0}|T_i = 0, R_i = r)) \right|, \\ CM_{n,r}^1 &= \frac{\sum_{i: R_i=r} (\sqrt{n} (F_{n,Y_{i0}}(y_{i0}|T_i = 1, R_i = r) - F_{n,Y_{i0}}(y_{i0}|T_i = 0, R_i = r))^2}{\sum_{i=1}^n 1\{R_i = r\}} \end{aligned} \quad (10)$$

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

Table D.1 presents the simulation rejection probabilities of the aforementioned statistics of the IV-R assumption. For each simulation design and attrition rate, we report the rejection probabilities for the KS statistics of the simple hypotheses,  $KS_{n,0}^1$  and  $KS_{n,1}^1$ , using asymptotic critical values ( $KS(Asym.)$ ) as a benchmark for the KS ( $KS(R)$ ) and the CM ( $CM(R)$ ) statistics using the  $p$ -values obtained from the proposed randomization procedure to test  $H_0^1$  ( $B = 199$ ). The different variants of the KS and CM test statistics control size under Designs II and III, where IV-R holds. They also have non-trivial power in finite samples in Designs I and IV, when IV-R is violated. The simulation results for the distributional statistics also illustrate the potential power gains in finite samples from using the attritor subgroup in testing the IV-R assumption. In testing the joint null hypothesis, we find that  $KS_{n,m}^1$  and  $CM_{n,m}^1$  (*Joint* ( $m$ )) exhibit better finite-sample power properties than  $KS_{n,p}^1$  and  $CM_{n,p}^1$  (*Joint* ( $p$ )). We also note that the randomization procedure yields rejection probabilities for the two-sample KS statistics,  $KS_{n,0}^1$  and  $KS_{n,1}^1$ , that are very similar to those obtained from the asymptotic critical values. In addition, in our simulation design, the CM statistics generally have better finite-sample power properties than their respective KS statistics, while maintaining comparable size control.

We then examine the finite-sample performance of the distributional statistics of the IV-P assumption. Proposition 1(b.ii) in the paper implies the three simple null hypotheses

as well as their joint hypothesis below,

$$\begin{aligned}
H_0^{2,1} : Y_{i0}|T_i = 0, R_i = 0 &\stackrel{d}{=} Y_{i0}|T_i = 0, R_i = 1, & (CA - CR) \\
H_0^{2,2} : Y_{i0}|T_i = 0, R_i = 1 &\stackrel{d}{=} Y_{i0}|T_i = 1, R_i = 0, & (CR - TA) \\
H_0^{2,3} : Y_{i0}|T_i = 1, R_i = 0 &\stackrel{d}{=} Y_{i0}|T_i = 1, R_i = 1, & (TA - TR) \\
H_0^2 : H_0^{2,1} \ \& \ H_0^{2,2} \ \& \ H_0^{2,3}. & (Joint)
\end{aligned} \tag{11}$$

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

$$\begin{aligned}
KS_{n,j}^2 &= \max_{i:(T_i, R_i) \in \{(\tau_j, r_j), (\tau_{j+1}, r_{j+1})\}} \left| \sqrt{n} \left( F_{n, Y_{i0}|T_i=\tau_{j-1}, R_i=r_{j-1}} - F_{n, Y_{i0}|T_i=\tau_j, R_i=r_j} \right) \right|, \\
CM_{n,j}^2 &= \frac{\sum_{i:(T_i, R_i) \in \{(\tau_j, r_j), (\tau_{j+1}, r_{j+1})\}} \left( \sqrt{n} \left( F_{n, Y_{i0}|T_i=\tau_{j-1}, R_i=r_{j-1}} - F_{n, Y_{i0}|T_i=\tau_j, R_i=r_j} \right) \right)^2}{\sum_{i=1}^n 1 \{ (T_i, R_i) \in \{(\tau_j, r_j), (\tau_{j+1}, r_{j+1})\} \}}, \tag{12}
\end{aligned}$$

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

In Table D.2, we report the simulation rejection probabilities for distributional tests of the IV-P assumption. In addition to the aforementioned statistics whose p-values are obtained using the proposed randomization procedure to test  $H_0^2$  ( $B = 199$ ), the table also reports the simulation results for the KS statistics of the simple hypotheses using the asymptotic critical values. Under Designs I, II and IV, IV-P is violated, the rejection probabilities for all the test statistics we consider tend to be higher than the nominal level, as we would expect. The joint KS and CM test statistics behave similarly in this design and have comparable finite-sample power properties to the test statistic of the simple hypothesis (TA-TR), which has the best finite-sample power properties in our simulation design. Finally, in Design III, where IV-P holds, our simulation results illustrate that the test statistics we consider control size.

## D.2 Additional Variants of the Simulation Designs

To illustrate the relative power properties of using the simple vs joint tests of internal validity, we present additional results using variants of the simulation designs. We show the results of the KS tests for the case where  $P(R_i = 0|T_i = 0) = 0.15$ .<sup>5</sup> For the joint hypotheses, we report the simulation results for the KS statistic that takes the maximum over the individual statistics.

Panel A in Figure D.1 displays the simulation rejection probabilities of the tests of the IV-R assumption while Panel B displays the simulation rejection probabilities of the tests of the IV-P assumption. We present these rejection probabilities for alternative parameter values of the designs we consider in Section 5 in the paper. *Design II to I* depicts the case in which we vary the proportion of treatment-only responders,  $p_{01}$ , from zero to  $0.9 \times P(R_i =$

---

<sup>5</sup>We use an attrition rate of 15% in the control group as reference since that is the average attrition rate in our review of field experiments. See Section 2 in the paper for details.

$0|T_i = 0)$ , where  $p_{01} = 0$  corresponds to Design II and  $p_{01} > 0$  to variants of Design I. *Design III to I* depicts the case in which we vary the correlation parameter between the unobservables in the outcome equation and the unobservables in the response equation,  $\rho$ , from zero to one. Hence,  $\rho = 0$  corresponds to Design III while  $\rho > 0$  corresponds to different versions of Design I. Finally, the results under *Design II to IV* are obtained by fixing  $p_{01} = p_{10}$  and varying them from zero to  $0.9 \times P(R_i = 0|T_i = 0)$ . Design II corresponds to the case in which  $p_{01} = p_{10} = 0$  and  $p_{01} = p_{10} > 0$  corresponds to different versions of Design IV.

Overall, the simulation results illustrate that the *joint* tests that we propose in Section 4 in the paper have better finite-sample power properties relative to the statistics of the simple null hypotheses. Most notably, the results under *Design II to I* in Panel A of Figure [D.1](#) show that when IV-R does not hold (i.e.  $p_{01} > 0$ ), the simulation rejection probabilities of the joint test are generally above the simulation rejection probabilities of the simple test that only uses the respondents.



Table D.1: Simulation Results on the KS & CM Randomization Test of IV-R

Design	Att. Rate		KS ( <i>Asym.</i> )				KS ( <i>R</i> )				CM( <i>R</i> )			
	C	T	CR-TR	CA-TA	CR-TR	CA-TA	Joint (m)	Joint (p)	CR-TR	CA-TA	Joint (m)	Joint (p)		
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$														
I	0.050	0.025	0.058	0.316	0.058	0.324	0.324	0.081	0.058	0.353	0.353	0.285		
	0.100	0.050	0.066	0.589	0.071	0.582	0.582	0.157	0.072	0.636	0.636	0.568		
	0.150	0.100	0.067	0.460	0.067	0.483	0.483	0.167	0.069	0.544	0.544	0.460		
	0.200	0.150	0.070	0.392	0.073	0.412	0.412	0.180	0.069	0.462	0.462	0.385		
	0.300	0.200	0.111	0.790	0.123	0.801	0.801	0.502	0.135	0.855	0.855	0.803		
Equal Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))^\dagger$														
II	0.050	0.050	0.052	0.059	0.053	0.062	0.062	0.052	0.054	0.056	0.056	0.061		
	0.100	0.100	0.049	0.054	0.053	0.056	0.056	0.050	0.054	0.054	0.054	0.053		
	0.150	0.150	0.044	0.049	0.049	0.055	0.055	0.051	0.049	0.054	0.054	0.055		
	0.200	0.200	0.052	0.044	0.052	0.050	0.050	0.058	0.052	0.049	0.049	0.052		
	0.300	0.300	0.051	0.043	0.051	0.042	0.043	0.053	0.049	0.047	0.048	0.057		
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$ ( <i>Example 1</i> )*														
III	0.050	0.025	0.049	0.051	0.054	0.052	0.052	0.056	0.048	0.051	0.051	0.049		
	0.100	0.050	0.047	0.042	0.050	0.046	0.046	0.047	0.053	0.047	0.047	0.043		
	0.150	0.100	0.047	0.038	0.052	0.045	0.045	0.047	0.049	0.049	0.049	0.048		
	0.200	0.150	0.054	0.031	0.053	0.036	0.036	0.047	0.055	0.036	0.036	0.044		
	0.300	0.200	0.050	0.043	0.050	0.043	0.043	0.050	0.051	0.042	0.042	0.050		
Equal Attrition Rates + Violation of Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$ ( <i>Example 2</i> )														
IV	0.050	0.050	0.059	0.332	0.065	0.329	0.329	0.093	0.067	0.375	0.375	0.302		
	0.100	0.100	0.102	0.569	0.102	0.577	0.577	0.230	0.116	0.663	0.663	0.593		
	0.150	0.150	0.178	0.740	0.190	0.758	0.758	0.465	0.211	0.816	0.816	0.805		
	0.200	0.200	0.313	0.854	0.319	0.859	0.859	0.709	0.368	0.917	0.916	0.910		
	0.300	0.300	0.683	0.970	0.680	0.972	0.974	0.974	0.760	0.985	0.991	0.996		

*Notes:* The above table presents the rejection probabilities of the KS and CM tests for the simple and joint null hypotheses in (9). We use the nominal level  $\alpha = 0.05$ , 2,000 simulation replications and  $n = 2,000$ .  $C$  denotes the control group,  $T$  denotes the treatment group.  $KS(Asym.)$  refers to the two-sample KS test using the asymptotic critical values.  $KS(R)$  and  $CM(R)$  refer to the randomization KS and CM tests, respectively, for the simple and joint hypotheses. *Joint (m)* and *Joint (p)* denote the randomization procedure applied to  $KS_{n,m}^1$  ( $CM_{n,m}^1$ ) and  $KS_{n,p}^1$  ( $CM_{n,p}^1$ ), respectively. Additional details of the design are provided in Table 4 in the paper.

<sup>†</sup> (\*) indicates IV-R only (IV-P).

Table D.2: Simulation Results on the KS & CM Randomization Test of IV-P

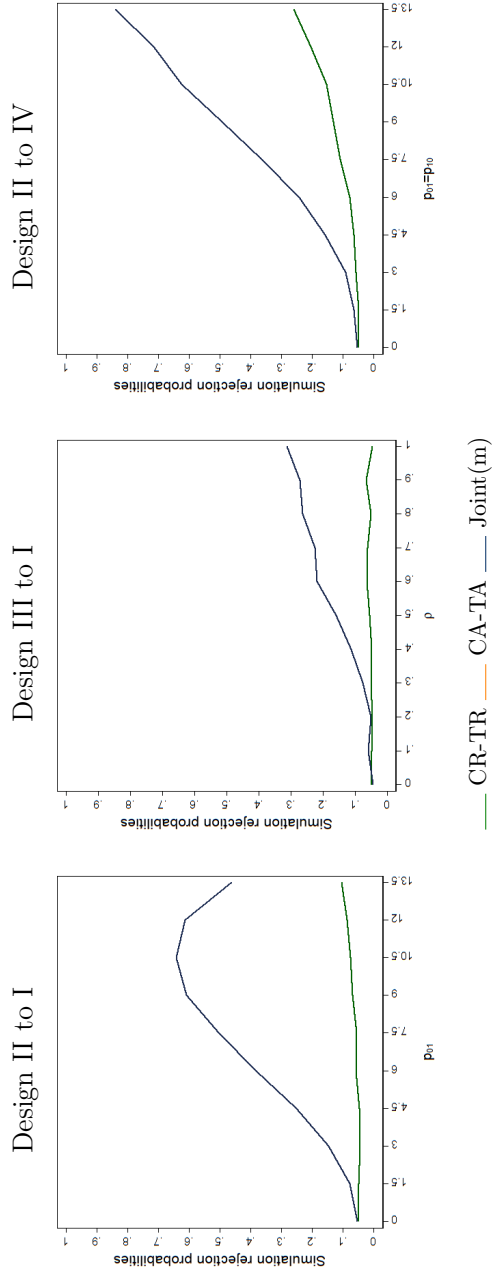
Design	Att. Rate		KS ( <i>Asym.</i> )				KS ( <i>R</i> )				CM( <i>R</i> )			
	C	T	CA-CR	CR-TA	TA-TR	CA-CR	CR-TA	TA-TR	Joint ( <i>m</i> )	CA-CR	CR-TA	TA-TR	Joint ( <i>m</i> )	
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$														
I	0.050	0.025	0.051	0.451	0.456	0.064	0.482	0.485	0.476	0.053	0.492	0.497	0.497	0.483
	0.100	0.050	0.053	0.746	0.787	0.055	0.763	0.801	0.787	0.058	0.806	0.837	0.837	0.824
	0.150	0.100	0.414	0.970	0.980	0.420	0.969	0.978	0.980	0.463	0.983	0.986	0.986	0.989
	0.200	0.150	0.865	0.999	0.998	0.870	0.998	0.998	1.000	0.902	1.000	0.999	0.999	1.000
	0.300	0.200	0.774	1.000	1.000	0.771	1.000	1.000	1.000	0.825	1.000	1.000	1.000	1.000
Equal Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))^\dagger$														
II	0.050	0.050	0.772	0.788	0.788	0.780	0.797	0.804	0.902	0.831	0.840	0.841	0.841	0.939
	0.100	0.100	0.984	0.983	0.980	0.985	0.981	0.981	0.999	0.994	0.989	0.986	0.986	1.000
	0.150	0.150	1.000	1.000	0.998	1.000	1.000	0.998	1.000	1.000	1.000	0.999	0.999	1.000
	0.200	0.200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.300	0.300	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$ ( <i>Example 1</i> )*														
III	0.050	0.025	0.040	0.042	0.043	0.044	0.050	0.051	0.050	0.047	0.053	0.053	0.053	0.054
	0.100	0.050	0.051	0.041	0.048	0.058	0.052	0.052	0.055	0.056	0.050	0.057	0.057	0.056
	0.150	0.100	0.040	0.051	0.052	0.046	0.056	0.057	0.059	0.047	0.054	0.055	0.055	0.059
	0.200	0.150	0.037	0.040	0.045	0.041	0.046	0.050	0.048	0.046	0.045	0.054	0.054	0.050
	0.300	0.200	0.048	0.044	0.044	0.050	0.049	0.046	0.048	0.049	0.044	0.051	0.051	0.054
Equal Attrition Rates + Violation of Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$ ( <i>Example 2</i> )														
IV	0.050	0.050	0.075	0.325	0.361	0.082	0.350	0.384	0.311	0.097	0.363	0.407	0.407	0.342
	0.100	0.100	0.113	0.548	0.668	0.125	0.558	0.681	0.582	0.152	0.605	0.742	0.742	0.661
	0.150	0.150	0.169	0.683	0.854	0.180	0.694	0.858	0.792	0.220	0.756	0.908	0.908	0.861
	0.200	0.200	0.234	0.759	0.947	0.239	0.762	0.950	0.913	0.288	0.822	0.974	0.974	0.952
	0.300	0.300	0.371	0.805	0.999	0.376	0.813	0.999	0.998	0.440	0.875	1.000	1.000	1.000

*Notes:* The above table presents the rejection probabilities of the KS and CM tests for the simple and joint null hypotheses in (11). We use the nominal level  $\alpha = 0.05$ , 2,000 simulation replications and  $n = 2,000$ .  $C$  denotes the control group,  $T$  denotes the treatment group.  $KS(Asym.)$  refers to the two-sample test using the asymptotic critical values.  $KS(R)$  and  $CM(R)$  refer to the randomization KS and CM tests, respectively, for the simple and joint hypotheses.  $Joint(m)$  denotes the randomization procedure applied to  $KS_{n,m}^2$  ( $CM_{n,m}^2$ ). Additional details of the design are provided in Table 4 in the paper.

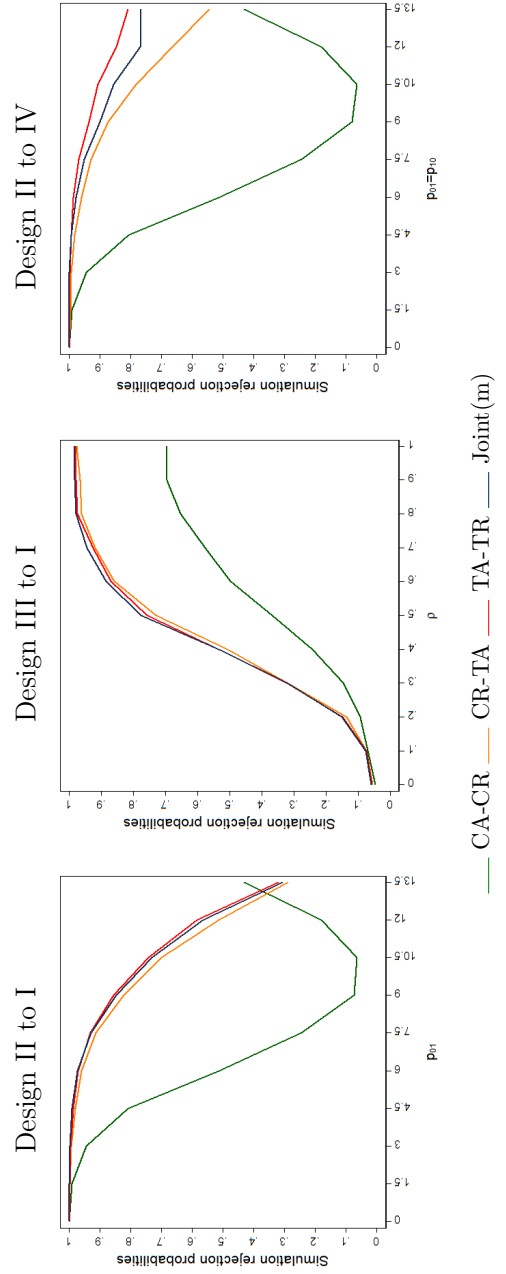
<sup>†</sup> (\*) indicates IV-R only (IV-P).

Figure D.1: Additional Simulation Analysis for the KS Statistics of Internal Validity

Panel A. Internal Validity for Respondents



Panel B. Internal Validity for the Study Population



## E List of Papers Included in the Review of Field Experiments

Abdulkadiroğlu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., & Pathak, P. A. (2011). Accountability and Flexibility in Public Schools: Evidence from Boston's Charters And Pilots. *Quarterly Journal of Economics*, 126(2), 699-748.

Aker, J. C., Ksoll, C., & Lybbert, T. J. (2012). Can Mobile Phones Improve Learning? Evidence from a Field Experiment in Niger. *American Economic Journal: Applied Economics*, 4(4), 94-120.

Ambler, K. (2015). Don't tell on me: Experimental evidence of asymmetric information in transnational households. *Journal of Development Economics*, 113, 52-69.

Ambler, K., Aycinena, D., & Yang, D. (2015). Channeling Remittances to Education: A Field Experiment among Migrants from El Salvador. *American Economic Journal: Applied Economics*, 7(2), 207-232.

Anderson, E. T., & Simester, D. I. (2010). Price Stickiness and Customer Antagonism. *Quarterly Journal of Economics*, 125(2), 729-765.

Ashraf, N., Aycinena, D., Martínez A., C., & Yang, D. (2015). Savings in Transnational Households: A Field Experiment among Migrants from El Salvador. *Review of Economics and Statistics*, 97(2), 332-351.

Ashraf, N., Berry, J., & Shapiro, J. M. (2010). Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia. *American Economic Review*, 100(5), 2383-2413.

Attanasio, O., Augsburg, B., De Haas, R., Fitzsimons, E., & Harmgart, H. (2015). The Impacts of Microfinance: Evidence from Joint-Liability Lending in Mongolia. *American Economic Journal: Applied Economics*, 7(1), 90-122.

Augsburg, B., De Haas, R., Harmgart, H., & Meghir, C. (2015). The Impacts of Microcredit: Evidence from Bosnia and Herzegovina. *American Economic Journal: Applied Economics*, 7(1), 183-203.

Avvisati, F., Gurgand, M., Guyon, N., & Maurin, E. (2014). Getting Parents Involved: A Field Experiment in Deprived Schools. *Review of Economic Studies*, 81(1), 57-83.

Baird, S., McIntosh, C., & Özler, B. (2011). Cash or Condition? Evidence from a Cash Transfer Experiment. *Quarterly Journal of Economics*, 126(4), 1709-1753.

Barham, T. (2011). A healthier start: The effect of conditional cash transfers on neonatal and infant mortality in rural Mexico. *Journal of Development Economics*, 94(1), 74-85.

Barton, J., Castillo, M., & Petrie, R. (2014). What Persuades Voters? A Field Experiment on Political Campaigning. *Economic Journal*, 124(574), F293-F326.

Basu, K., & Wong, M. (2015). Evaluating seasonal food storage and credit programs in east Indonesia. *Journal of Development Economics*, 115, 200-216.

Bauchet, J., Morduch, J., & Ravi, S. (2015). Failure vs. displacement: Why an innovative anti-poverty program showed no net impact in South India. *Journal of Development Economics*, 116, 1-16.

Bengtsson, N., & Engström, P. (2014). Replacing Trust with Control: A Field Test of Motivation Crowd Out Theory. *Economic Journal*, 124(577), 833-858.

Bettinger, E. P. (2012). Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores. *Review of Economics and Statistics*, 94(3), 686-698.

- Beuermann, D. W., Cristia, J., Cueto, S., Malamud, O., & Cruz-Aguayo, Y. (2015). One Laptop per Child at Home: Short-Term Impacts from a Randomized Experiment in Peru. *American Economic Journal: Applied Economics*, 7(2), 53-80.
- Bianchi, M., & Bobba, M. (2013). Liquidity, Risk, and Occupational Choices. *Review of Economic Studies*, 80(2), 491-511.
- Björkman, M., & Svensson, J. (2009). Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda. *Quarterly Journal of Economics*, 124(2), 735-769.
- Blattman, C., Fiala, N., & Martinez, S. (2014). Generating Skilled Self-Employment in Developing Countries: Experimental Evidence from Uganda. *Quarterly Journal of Economics*, 129(2), 697-752.
- Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., & Roberts, J. (2013). Does Management Matter? Evidence from India. *Quarterly Journal of Economics*, 128(1), 1-51.
- Bloom, N., Liang, J., Roberts, J., & Ying, Z. J. (2015). Does Working from Home Work? Evidence from a Chinese Experiment. *Quarterly Journal of Economics*, 130(1), 165-218.
- Bobonis, G. J., & Finan, F. (2009). Neighborhood Peer Effects in Secondary School Enrollment Decisions. *Review of Economics and Statistics*, 91(4), 695-716.
- Bruhn, M., Ibarra, G. L., & McKenzie, D. (2014). The minimal impact of a large-scale financial education program in Mexico City. *Journal of Development Economics*, 108, 184-189.
- Bryan, G., Chowdhury, S., & Mobarak, A. M. (2014). Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh. *Econometrica*, 82(5), 1671-1748.
- Cai, H., Chen, Y., Fang, H., & Zhou, L.-A. (2015). The Effect of Microinsurance on Economic Activities: Evidence from a Randomized Field Experiment. *Review of Economics and Statistics*.
- Charness, G., & Gneezy, U. (2009). Incentives to Exercise. *Econometrica*, 77(3), 909-931.
- Chetty, R., & Saez, E. (2013). Teaching the Tax Code: Earnings Responses to an Experiment with EITC Recipients. *American Economic Journal: Applied Economics*, 5(1), 1-31.
- Collier, P., & Vicente, P. C. (2014). Votes and Violence: Evidence from a Field Experiment in Nigeria. *Economic Journal*, 124(574), F327-F355.
- Crépon, B., Devoto, F., Duflo, E., & Parienté, W. (2015). Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco. *American Economic Journal: Applied Economics*, 7(1), 123-150.
- Cunha, J. M. (2014). Testing Paternalism: Cash versus In-Kind Transfers. *American Economic Journal: Applied Economics*, 6(2), 195-230.
- De Grip, A., & Sauermann, J. (2012). The Effects of Training on Own and Co-worker Productivity: Evidence from a Field Experiment. *Economic Journal*, 122(560), 376-399.
- de Mel, S., McKenzie, D., & Woodruff, C. (2014). Business training and female enterprise start-up, growth, and dynamics: Experimental evidence from Sri Lanka. *Journal of Development Economics*, 106, 199-210.
- De Mel, S., McKenzie, D., & Woodruff, C. (2012). Enterprise Recovery Following Natural Disasters. *Economic Journal*, 122(559), 64-91.

- de Mel, S., McKenzie, D., & Woodruff, C. (2013). The Demand for, and Consequences of, Formalization among Informal Firms in Sri Lanka. *American Economic Journal: Applied Economics*, 5(2), 122-150.
- Dinkelman, T., & Martínez A., C. (2014). Investing in Schooling In Chile: The Role of Information about Financial Aid for Higher Education. *Review of Economics and Statistics*, 96(2), 244-257.
- Doi, Y., McKenzie, D., & Zia, B. (2014). Who you train matters: Identifying combined effects of financial education on migrant households. *Journal of Development Economics*, 109, 39-55.
- Duflo, E., Dupas, P., & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5), 1739-1774.
- Duflo, E., Greenstone, M., Pande, R., & Ryan, N. (2013). Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India. *Quarterly Journal of Economics*, 128(4), 1499-1545.
- Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives Work: Getting Teachers to Come to School. *American Economic Review*, 102(4), 1241-1278.
- Dupas, P., & Robinson, J. (2013). Savings Constraints and Microenterprise Development: Evidence from a Field Experiment in Kenya. *American Economic Journal: Applied Economics*, 5(1), 163-192.
- Edmonds, E. V., & Shrestha, M. (2014). You get what you pay for: Schooling incentives and child labor. *Journal of Development Economics*, 111, 196-211.
- Fafchamps, M., McKenzie, D., Quinn, S., & Woodruff, C. (2014). Microenterprise growth and the flypaper effect: Evidence from a randomized experiment in Ghana. *Journal of Development Economics*, 106(Supplement C), 211-226.
- Fafchamps, M., & Vicente, P. C. (2013). Political violence and social networks: Experimental evidence from a Nigerian election. *Journal of Development Economics*, 101(Supplement C), 27-48.
- Ferraro, P. J., & Price, M. K. (2013). Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment. *Review of Economics and Statistics*, 95(1), 64-73.
- Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Baicker, K. (2012). The Oregon Health Insurance Experiment: Evidence from the First Year. *Quarterly Journal of Economics*, 127(3), 1057-1106.
- Fryer, J. R. G. (2011). Financial Incentives and Student Achievement: Evidence from Randomized Trials. *Quarterly Journal of Economics*, 126(4), 1755-1798.
- Fryer, J. R. G. (2014). Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments. *Quarterly Journal of Economics*, 129(3), 1355-1407.
- Gertler, P. J., Martinez, S. W., & Rubio-Codina, M. (2012). Investing Cash Transfers to Raise Long-Term Living Standards. *American Economic Journal: Applied Economics*, 4(1), 164-192.
- Giné, X., Goldberg, J., & Yang, D. (2012). Credit Market Consequences of Improved Personal Identification: Field Experimental Evidence from Malawi. *American Economic Review*, 102(6), 2923-2954.

Giné, X., & Karlan, D. S. (2014). Group versus individual liability: Short and long term evidence from Philippine microcredit lending groups. *Journal of Development Economics*, 107, 65-83.

Hainmueller, J., Hiscox, M. J., & Sequeira, S. (2015). Consumer Demand for Fair Trade: Evidence from a Multistore Field Experiment. *Review of Economics and Statistics*, 97(2), 242-256.

Hanna, R., Mullainathan, S., & Schwartzstein, J. (2014). Learning Through Noticing: Theory and Evidence from a Field Experiment. *Quarterly Journal of Economics*, 129(3), 1311-1353.

Hidrobo, M., Hoddinott, J., Peterman, A., Margolies, A., & Moreira, V. (2014). Cash, food, or vouchers? Evidence from a randomized experiment in northern Ecuador. *Journal of Development Economics*, 107, 144-156.

Jackson, C. K., & Schneider, H. S. (2015). Checklists and Worker Behavior: A Field Experiment. *American Economic Journal: Applied Economics*, 7(4), 136-168.

Jacob, B. A., Kapustin, M., & Ludwig, J. (2015). The Impact of Housing Assistance on Child Outcomes: Evidence from a Randomized Housing Lottery. *Quarterly Journal of Economics*, 130(1), 465-506.

Jensen, R. (2012). Do Labor Market Opportunities Affect Young Women's Work and Family Decisions? Experimental Evidence from India. *Quarterly Journal of Economics*, 127(2), 753-792.

Jensen, R. T., & Miller, N. H. (2011). Do Consumer Price Subsidies Really Improve Nutrition? *Review of Economics and Statistics*, 93(4), 1205-1223.

Karlan, D., Osei, R., Osei-Akoto, I., & Udry, C. (2014). Agricultural Decisions after Relaxing Credit and Risk Constraints. *Quarterly Journal of Economics*, 129(2), 597-652.

Karlan, D., & Valdivia, M. (2011). Teaching Entrepreneurship: Impact of Business Training on Microfinance Clients and Institutions. *Review of Economics and Statistics*, 93(2), 510-527.

Kazianga, H., de Walque, D., & Alderman, H. (2014). School feeding programs, intra-household allocation and the nutrition of siblings: Evidence from a randomized trial in rural Burkina Faso. *Journal of Development Economics*, 106, 15-34.

Kendall, C., Nannicini, T., & Trebbi, F. (2015). How Do Voters Respond to Information? Evidence from a Randomized Campaign. *American Economic Review*, 105(1), 322-353.

Kling, J. R., Mullainathan, S., Shafir, E., Vermeulen, L. C., & Wrobel, M. V. (2012). Comparison Friction: Experimental Evidence from Medicare Drug Plans. *Quarterly Journal of Economics*, 127(1), 199-235.

Kremer, M., Leino, J., Miguel, E., & Zwane, A. P. (2011). Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions. *Quarterly Journal of Economics*, 126(1), 145-205.

Labonne, J. (2013). The local electoral impacts of conditional cash transfers: Evidence from a field experiment. *Journal of Development Economics*, 104, 73-88.

Lalive, R., & Cattaneo, M. A. (2009). Social Interactions and Schooling Decisions. *Review of Economics and Statistics*, 91(3), 457-477.

Macours, K., Schady, N., & Vakis, R. (2012). Cash Transfers, Behavioral Changes, and Cognitive Development in Early Childhood: Evidence from a Randomized Experiment.

*American Economic Journal: Applied Economics*, 4(2), 247-273.

Macours, K., & Vakis, R. (2014). Changing Households' Investment Behaviour through Social Interactions with Local Leaders: Evidence from a Randomised Transfer Programme. *Economic Journal*, 124(576), 607-633.

Meredith, J., Robinson, J., Walker, S., & Wydick, B. (2013). Keeping the doctor away: Experimental evidence on investment in preventative health products. *Journal of Development Economics*, 105, 196-210.

Muralidharan, K., & Sundararaman, V. (2011). Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy*, 119(1), 39-77.

Muralidharan, K., & Sundararaman, V. (2015). The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India. *Quarterly Journal of Economics*, 130(3), 1011-1066.

Olken, B. A., Onishi, J., & Wong, S. (2014). Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia. *American Economic Journal: Applied Economics*, 6(4), 1-34.

Pallais, A. (2014). Inefficient Hiring in Entry-Level Labor Markets. *American Economic Review*, 104(11), 3565-3599.

Pomeranz, D. (2015). No Taxation without Information: Deterrence and Self-Enforcement in the Value Added Tax. *American Economic Review*, 105(8), 2539-2569.

Powell-Jackson, T., Hanson, K., Whitty, C. J. M., & Ansah, E. K. (2014). Who benefits from free healthcare? Evidence from a randomized experiment in Ghana. *Journal of Development Economics*, 107, 305-319.

Pradhan, M., Suryadarma, D., Beatty, A., Wong, M., Gaduh, A., Alisjahbana, A., & Artha, R. P. (2014). Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia. *American Economic Journal: Applied Economics*, 6(2), 105-126.

Prina, S. (2015). Banking the poor via savings accounts: Evidence from a field experiment. *Journal of Development Economics*, 115, 16-31.

Royer, H., Stehr, M., & Sydnor, J. (2015). Incentives, Commitments, and Habit Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company. *American Economic Journal: Applied Economics*, 7(3), 51-84.

Seshan, G., & Yang, D. (2014). Motivating migrants: A field experiment on financial decision-making in transnational households. *Journal of Development Economics*, 108, 119-127.

Stutzer, A., Goette, L., & Zehnder, M. (2011). Active Decisions and Prosocial Behaviour: a Field Experiment on Blood Donation. *Economic Journal*, 121(556), F476-F493.

Szabó, A., & Ujhelyi, G. (2015). Reducing nonpayment for public utilities: Experimental evidence from South Africa. *Journal of Development Economics*, 117, 20-31.

Tarozzi, A., Mahajan, A., Blackburn, B., Kopf, D., Krishnan, L., & Yoong, J. (2014). Micro-loans, insecticide-treated bednets, and malaria: Evidence from a randomized controlled trial in Orissa, India. *American Economic Review*, 104, 1909-41.

Thornton, R. L. (2012). HIV testing, subjective beliefs and economic behavior. *Journal of Development Economics*, 99(2), 300-313.

Valdivia, M. (2015). Business training plus for female entrepreneurship? Short and medium-term experimental evidence from Peru. *Journal of Development Economics*, 113,



33-51.

Vicente, P. C. (2014). Is Vote Buying Effective? Evidence from a Field Experiment in West Africa. *Economic Journal*, 124(574), F356-F387.

Walters, C. R. (2015). Inputs in the Production of Early Childhood Human Capital: Evidence from Head Start. *American Economic Journal: Applied Economics*, 7(4), 76-102.

Wilson, N. L., Xiong, W., & Mattson, C. L. (2014). Is sex like driving? HIV prevention and risk compensation. *Journal of Development Economics*, 106, 78-91.

## References

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