Testing Attrition Bias in Field Experiments*

Dalia Ghanem — UC Davis
Sarojini Hirshleifer — UC Riverside
Karen Ortiz-Becerra — UC Davis

February 18, 2021

Abstract

We approach attrition in field experiments with baseline 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 propose randomization and regression-based procedures to test their sharp implications. We then relate our proposed tests to current empirical practice, and demonstrate that the most commonly used test can over-reject internal validity. Simulations and applications further support the relevance of our analysis.

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

*E-mail: dghanem@ucdavis.edu, sarojini.hirshleifer@ucr.edu, kaortizb@ucdavis.edu.
We thank Alberto Abadie, Josh Angrist, Stephen Boucher, Federico Bugni, Pamela Jakiela, Tae-hwy Lee, Jia Li, Aprajit Mahajan, Matthew Masten, Craig McIntosh, David McKenzie, Adam Rosen, Monica Singhal and Aman Ullah for helpful discussions. The Stata ado file to implement the attrition tests proposed in this paper is available at https://github.com/daghanem/ATTRITIONTESTS.

1 Introduction

Randomized control trials (RCTs) are an increasingly important tool of applied economics since, when properly designed and implemented, they can produce internally valid estimates of causal impact.[1] Non-response on outcome measures at endline, however, is an unavoidable threat to the internal validity of many carefully implemented trials. Long-distance migration can make it prohibitively expensive to follow members of an evaluation sample. Conflict, intimidation or natural disasters sometimes make it unsafe to collect complete response data. In high-income countries, survey response rates are often low and may be declining.[2] The recent, increased focus on the long-term impacts of interventions has also made non-response especially relevant. Thus, researchers often face the question: How much of a threat is attrition to the internal validity of a given study?

In this paper, we approach attrition in field experiments with baseline data as an identification problem in a nonseparable panel model. We focus on two identification questions generated by attrition in this setting. First, does the difference in mean outcomes between treatment and control respondents identify the average treatment effect for the respondent subpopulation (ATE-R)? Second, is this estimand equal to the average treatment effect for the study population (ATE)?[3] To answer these questions, we examine the testable implications of the relevant identifying assumptions and propose procedures to test them. Our results provide insights that are relevant to current empirical practice.

We first conduct a systematic review of 96 recent field experiments with baseline outcome data in order to document attrition rates and understand how authors test for attrition bias. Attrition and attrition tests are both common in published field experiments. Although we find wide variation in the choice and implementation of attrition tests in the literature, we are able to identify two main types: (i) a *differential attrition rate test* that determines if

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

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

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

attrition rates are different across treatment and control groups, and (ii) a *selective attrition test* that attempts to determine if the mean of baseline observable characteristics differs across the treatment and control groups conditional on response status. While authors report a differential attrition rate test for 79% of field experiments, they report a selective attrition test only 60% of the time. In addition, for a substantial minority of field experiments (36%), 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 outcome data. Specifically, we establish the identifying assumptions in the presence of attrition for two cases that are likely to be of interest to the researcher. For the first case, in which the researcher's objective is internal validity for the respondent subpopulation (IV-R), the identifying assumption is random assignment conditional on response status (IV-R assumption). This implies that the difference in the mean outcome across the treatment and control respondents identifies the ATE-R, a local average treatment effect for the respondents.[4] In the second case, where internal validity for the study population (IV-P) is of interest, the identifying assumption is that the unobservables that affect response and outcome are independent in addition to the initial random assignment of the treatment (IV-P assumption). If this identifying assumption holds, the ATE for the study population is identified. This second case is especially relevant in settings where the study population is representative of a larger population.

We then derive testable restrictions for each of the above identifying assumptions. If treatment effects for the respondents are the researchers' object of interest, they can implement a test of the IV-R assumption. The null hypothesis of the IV-R test consists of two equality restrictions on the baseline outcome distribution; specifically, for treatment and control respondents as well as treatment and control attritors. Alternatively, if the researchers are interested in treatment effects for the study population, they can test the restriction

[4]For brevity, we use a "difference in means" to refer to a "difference in population means". To distinguish it from its sample analogue, we refer to the latter as a "difference in sample means".

of the IV-P assumption. The hypothesis of the IV-P test is the equality of the baseline outcome distribution across all four treatment/response subgroups. We show that these testable restrictions are sharp, meaning that they are the strongest implications that we can test given the available data.[5] We also propose randomization procedures to test the sharp distributional restrictions implied by each identifying assumption as well as regression-based procedures to test their mean counterparts.

In a motivating example, we apply our proposed attrition tests to the randomized evaluation of the *Progresa* program in which the study population is representative of a broader population of interest. We focus on two main outcomes, school enrollment and adult employment. The IV-R test does not reject for either of these outcomes, which is promising for the identification of the treatment effects for the respondent subpopulation. Interestingly, the IV-P test rejects for school enrollment, but it does not reject for adult employment. Thus, for school enrollment only, we reject the internal validity of its treatment effects for the study population. This application illustrates that attrition can have differential implications for the interpretation of treatment effects for different outcomes, even those collected in the same survey. An important takeaway from our analysis is that researchers should consider an outcome-specific approach to testing for attrition bias.

Given their relevance to current empirical practice, we also provide a formal treatment of the differential attrition rate test and the use of covariates. In order to understand the role of differential attrition rates for internal validity, we apply the framework of partial compliance from the local average treatment effect (LATE) literature to potential response.[6] We demonstrate that even though equal attrition rates are sufficient for IV-R under additional assumptions, they are not a necessary condition for internal validity in general. We illustrate using an analytical example and simulations that it is possible to have differences

[5]Sharp testable restrictions are the restrictions for which there are the smallest possible set of cases such that the testable restriction holds even though the identifying assumption does not. The concept of sharpness of testable restrictions was previously developed and applied in Kitagawa (2015), Hsu et al. (2019), and Mourifié and Wan (2017).

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

in attrition rates across treatment and control groups while internal validity holds not only for the respondent subpopulation but also the study population. Next, we examine the use of covariates in testing the IV-R or IV-P assumption, which is useful for settings where data on the outcome is not available at baseline. We note two types of covariates that may be included: (i) determinants of the outcome, and (ii) "proxy" variables which are determined by the same variables as the outcome in question. We caution that using covariates that do not fulfill either of these criteria can lead to a false rejection of the IV-R or IV-P assumption.

Finally, we illustrate the empirical relevance of our results by applying our tests to five published field experiments with high attrition rates.[7] A particularly notable result is that, for two-thirds of the outcomes, we neither reject the IV-R nor the IV-P assumption, which ensures the identification of treatment effects for the study population. This is promising for field experiments where the study population is of interest. For the remaining outcomes, however, our tests reject the IV-P but not the IV-R assumption. In other words, for those outcomes, the researcher would reject the internal validity of the corresponding treatment effect for the study population, but would not reject the assumption that ensures the internal validity of the treatment effect for the respondent subpopulation. When we consider the authors' attrition tests, we find heterogeneity in the choice of tests as well as their implementation, consistent with the findings from our review. Furthermore, our empirical results support the limitations of the differential attrition rate test highlighted by the theoretical analysis. For about one-quarter of the outcomes, our test results are consistent with the conditions under which this test would not control size as a test of internal validity.

This paper has several implications for current empirical practice. First, our theoretical and empirical results imply that the most widely used test in the literature, the differential attrition rate test, may overreject internal validity in practice. The second most widely used test, the selective attrition test, is implemented using a variety of approaches. Most such tests constitute IV-R tests, although those typically use respondents only. Our theoretical

[7] We choose the five published field experiments from our review that have the highest attrition rates subject to data availability.

results indicate, however, that the implication of the relevant identifying assumption is a joint test that uses all of the available information in the baseline data, and thus includes both respondents and attritors. In addition, while the majority of testing procedures pertain to IV-R and not IV-P, the use of determinants of attrition tests suggests that some researchers may be interested in implications of the estimated treatment effects for the study population. Finally, we note that authors do not typically correct for multiple hypothesis testing in the implementation of selective attrition tests, even when these tests are performed on a non-trivial number of baseline variables. This is another possible source of overrejection of internal validity in the literature. More generally, this paper highlights the importance of understanding the implications of attrition for a broader population when interpreting field experiment results for policy.[8]

This paper contributes to a growing literature that considers methodological questions relevant to field experiments.[9] Given the wide use of attrition tests, we formally examine the testing problem here. Our focus complements a thread in this literature that outlines various approaches to correcting attrition bias in field experiments (Horowitz and Manski, 2000; Lee, 2009; Huber, 2012; Behagel et al., 2015; Millán and Macours, 2019).[10] These corrections build on the vast sample selection literature in econometrics going back to Heckman (1976, 1979).[11] While the latter literature is broadly concerned with population objects, work that is

[8]External validity can be assessed in a number of ways (see, for example, Andrews and Oster (2019) and Azzam et al. (2018)). In our setting, we note that if IV-R holds but not IV-P, we may be able to draw inference from the local average treatment effect for respondents to a broader population.

[9]Bruhn and McKenzie (2009) compare the performance of different randomization methods; McKenzie (2012) discusses the power trade-offs of the number of follow-up samples in the experimental design; Baird et al. (2018) propose an optimal method to design field experiments in the presence of interference; de Chaisemartin and Behaghel (2018) present how to estimate treatment effects in the context of randomized wait lists; Abadie et al. (2018) propose alternative estimators that reduce the bias resulting from endogenous stratification in field experiments; Muralidharan et al. (2019) examine empirical practice in analyzing experiment with factorial design and analyze the trade-off between power and correct inference in this setting; Kasy and Sautmann (2020) propose a treatment assignment algorithm to choose the best among a set of policies at the end of an experiment; Vazquez-Bare (2020) examines the identification and estimation of spillover effects in randomized experiments.

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

[11]Nonparametric Heckman-style corrections have been proposed for linear and nonparametric outcome

relevant to program evaluation proposes corrections for objects pertaining to subpopulations (e.g. Lee, 2009; Huber, 2012; Chen and Flores, 2015). Our paper provides tests of identifying assumptions emphasizing the distinction between the (study) population and the respondent subpopulation. Finally, the randomization tests we propose contribute to recent work that examines the potential use of randomization tests in analyzing field experiment data (Young, 2018; Athey and Imbens, 2017; Athey et al., 2018; Bugni et al., 2018).

We also build on other strands of the econometrics literature. Recent work on nonparametric identification in nonseparable panel data models informs our approach (Altonji and Matzkin, 2005; Bester and Hansen, 2009; Chernozhukov et al., 2013; Hoderlein and White, 2012; Ghanem, 2017). Specifically, the identifying assumptions in this paper fall under the nonparametric correlated random effects category (Altonji and Matzkin, 2005). Furthermore, we build on the literature on randomization tests for distributional statistics (Dufour, 2006; Dufour et al., 1998).

The paper proceeds as follows. Section 2 presents the review of the field experiment literature. Section 3 formally presents the identifying assumptions and their sharp testable restrictions. It also includes a formal treatment of differential attrition rates and of the role of covariates in testing internal validity. Section 4 presents simulation experiments to illustrate the theoretical results. Section 5 presents the results of the empirical application exercise. Section 6 concludes. Sections A and B present the randomization and regression-based procedures, respectively, to test the IV-R and IV-P assumptions for completely, stratified and cluster randomized experiments.

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

2 Attrition in the Field Experiment Literature

We systematically reviewed 93 recent articles published in economics journals that report the results of 96 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.[12] Our categorization imposes some structure on the variety of different estimation strategies used to test for attrition bias in the literature.[13] 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.[14]

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, 45% of reviewed field experiments have an attrition rate higher than the average of 15%.[15] 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: 63% have a differential rate that is less than 2 percentage points, and only 11% have a

[12]We included articles from 2009 to 2015 that were published in the top five journals in economics as well as five 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 Human Resources*, *Journal of Political Economy*, *Review of Economics and Statistics*, *Review of Economic Studies*, and *Quarterly Journal of Economics*. Section SA1.1 in the online appendix includes additional details on the selection of papers and relevant attrition rates. Section SA6 in the online appendix contains a list of all the papers included in the review.

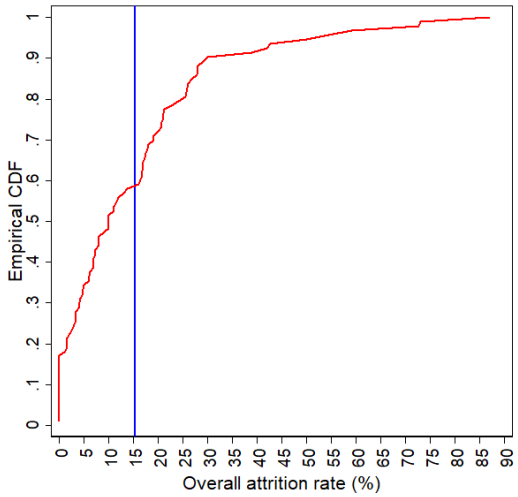
[13]We identify fifteen estimation strategies used to conduct attrition tests (see Section SA2 in the online appendix).

[14]We exclude 64 field experiments that were published during that time period, since they lack baseline data for any outcome mentioned in the abstract. Of those, slightly less than half (44%) are experiments for which the baseline outcome is the same for everyone by design and hence is not informative (see Section SA1.1 in the online appendix).

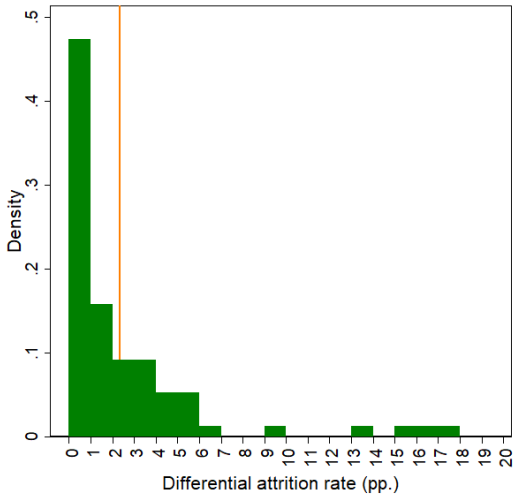
[15]To understand the extent of attrition that is relevant to the main outcomes in the paper, we focus on attrition rates that are relevant to outcomes reported in the abstract (i.e. "abstract results"). Most papers report attrition rates at the level of the data source or subsample, rather than at the level of the outcome. Since the number of data sources and/or subsamples that are relevant to the abstract results vary by experiment, we include one attrition rate per field experiment for consistency. Specifically, we report the highest attrition rate relevant to an abstract result. Authors do not in general report attrition rates conditional on baseline response. A noteworthy finding from Table SA2 in the online appendix is that attrition rates are higher on average for experiments in high-income countries.

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

Panel A. Overall Attrition Rate



Panel B. Differential Attrition Rate



Notes: We report one observation per field experiment. Specifically, the highest attrition rate relevant to a result reported in the abstract of the article. The *Overall* rate is the attrition rate for the full sample, which is composed of the treatment and control groups. The *Differential* rate is the absolute value of the difference in attrition rates across treatment and control groups. The blue (orange) line depicts the average overall (differential) attrition rate in our sample of field experiments. Panel A includes 93 field experiments and Panel B includes 76 experiments since the relevant attrition rates are not reported in some articles.

differential attrition rate that is greater than 5 percentage points.[16]

We then study how authors test for attrition bias. Notably, attrition tests are widely used in the literature: 92% of field experiments with an attrition rate of at least 1% for an outcome with baseline data conduct at least one attrition test. We first identify two main types of tests that aim to determine the impact of attrition on internal validity: (i) a ***differential attrition rate test***, and (ii) a ***selective attrition test***. A ***differential attrition rate test*** determines whether the rates of attrition are statistically significantly different across treatment and control groups. In contrast, a ***selective attrition test*** aims to determine whether, conditional on being a respondent and/or attritor, the mean of observable characteristics is the same across treatment and control groups. We find that there is no consensus on whether to conduct a differential attrition rate test or a selective attrition test, however (Panel A in

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

Table 1). In the field experiments that we reviewed, the differential attrition rate test is substantially more common (79%) than the selective attrition test (60%). In fact, 30% of the articles that conducted a differential attrition rate do not conduct a selective attrition test.[17]

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%	21%
	Yes	30%	49%	79%
	Total	40%	60%	100%

Panel B: Types of Selective Attrition Test

Conditional on conducting a selective attrition test:	
Test using respondents and attritors	29%
Test using respondents only	67%
Test using attritors only	4%
Total†	100%

Panel C: Determinants of Attrition Tests

Proportion of field experiments that conduct:	Determinants of attrition test		
	Yes	*No*	*Total*
Differential attrition rate test only	12%	18%	30%
Selective attrition test only	1%	9%	10%
Differential & selective attrition tests	21%	28%	49%
No differential & no selective attrition test	1%	9%	10%
Total	36%	64%	100%

Notes: Panel A and C include 77 field experiments that have an attrition rate of at least 1% for an outcome with baseline data. Panel B includes 46 of those experiments that conducted a selective attrition test (†). For details on the classification of the empirical strategies, see Section SA2 in the online appendix.

We further consider if selective attrition tests include both respondents and attritors or if they include either only respondents or only attritors (Panel B in Table 1). Conditional on having conducted any type of selective attrition test, authors include both respondents and

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

attritors in only 29% of those field experiments. Instead, authors conduct a selective attrition test on the sample of respondents in most cases (67%). Although our review is limited to experiments in which baseline outcome data is available, covariates are typically included in attrition tests along with the baseline outcome. In particular, 98% of field experiments that report a selective attrition test include more than one baseline variable in that test.[18] A key issue that arises with the inclusion of covariates is how to approach the issue of multiple testing. We find that 75% of the experiments that implement a selective attrition test conduct it on an average of 16 variables, and none of those implement a multiple testing correction (Table SA3 in online appendix). Only a minority of authors conduct a joint test across all of the baseline variables included in the test (25%).

Another important aspect of testing for attrition bias is testing for differences in the distributions of respondents and attritors. Such tests can illustrate the implications of the main results of the experiment for the study population. We define a *determinants of attrition test* as a test of whether baseline outcomes and covariates correlate with response status and find that authors conduct such a test in approximately one-third of field experiments (Panel C of Table 1). Table 1 illustrates that conducting the determinants of attrition test does not have a one-to-one relationship with either conducting a differential attrition rate test or conducting a selective attrition test.[19]

3 Testing Attrition Bias Using Baseline Data

This section presents a formal treatment of attrition in field experiments with baseline outcome data. First, we motivate the problem with an example from the *Progresa* evaluation. Then, we present the identifying assumptions in the presence of non-response and show their

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

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

sharp testable implications when baseline outcome data is available for both completely and stratified randomized experiments. We further examine the role of the widely-used differential attrition rate test and discuss the implications of our theoretical analysis for empirical practice.

3.1 Motivating Example

To illustrate the problem of attrition in field experiments, we use data collected for the randomized evaluation of *Progresa*, a social program in Mexico that provides cash to eligible poor households on the condition that children attend school and family members visit health centers regularly (Skoufias, 2005). The evaluation of *Progresa* relied on the cluster-level 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.[20] The surveys conducted for the experiment include a baseline and three follow-up rounds collected 5, 13, and 18 months after the program began.[21] 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.

In Table 2, we report the initial sample size and summary statistics for each outcome by treatment group at baseline and follow-up. The failure to reject the null hypothesis of the equality of means across the treatment and control groups at baseline is suggestive evidence that the randomization of localities into treatment and control was implemented correctly. In the context of treatment randomization and absence of attrition, the difference in a mean outcome across treatment and control groups at follow-up would identify the average

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

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

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
Panel A. School Enrollment (6-16 years old)								
Baseline	24353	0.824	0.007	0.455				
Pooled					0.183	0.793	0.046	0.000
1st					0.142	0.814	0.043	0.000
2nd					0.234	0.829	0.046	0.000
3rd					0.174	0.740	0.047	0.000
Panel B. Employment Last Week (18+ years old)								
Baseline	31237	0.471	-0.006	0.546				
Pooled					0.161	0.464	0.014	0.002
1st					0.096	0.460	0.016	0.016
2nd					0.196	0.459	0.009	0.138
3rd					0.192	0.472	0.018	0.001

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

treatment effect for the study population.[22] Pooling data from the three follow-up rounds, we would conclude that the impact of *Progesa* on school enrollment (adult employment) 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 these treatment effect estimates are unbiased for at least one of two objects of interest: (i) the average treatment effect for the respondent subpopulation (ATE-R) or (ii) the average treatment effect for the entire study population (ATE).

In order to understand whether attrition affects the internal validity of this experiment, we inspect the mean baseline outcomes across the four treatment-response subgroups. For the outcome of school enrollment, there are two distinct patterns. First, baseline school enrollment is similar across treatment and control respondents as well as treatment and control attritors. Second, we find meaningful differences when we compare respondents and attritors: baseline school enrollment is around 87% for the respondents and 61% for the attritors in the pooled follow-up sample. Taken together, these two patterns suggest that

[22]Here we follow our convention of referring to a "difference in population means" as a "difference in means."

while the unobservables that affect the outcome are correlated with response, they are still independent of the treatment *within* respondents and *within* attritors. As we formalize in the next section, independence between treatment status and the unobservables that affect the outcome conditional on response status constitutes the identifying assumption of internal validity for the respondents (IV-R assumption). We show that the IV-R assumption implies the identification of treatment effects for the respondent subpopulation and that its testable implication is that the distribution of a baseline outcome is identical across treatment and control respondents as well as treatment and control attritors. Applying this test to school enrollment in Column 7 of Table 3, we do not reject the IV-R assumption.[23] If the IV-R assumption does hold for this outcome, then the difference in means across treatment and control respondents at follow-up identifies an average treatment effect for the respondents (ATE-R).

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

Follow-up Sample	Attrition Rate		Mean Baseline Outcome by Group				Test of IV-R	Test of IV-P
	C	Differential	TR	CR	TA	CA	p-value	p-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. School Enrollment (6-16 years old)								
Pooled	0.187	-0.007	0.878	0.874	0.615	0.605	0.836	0.000
1st	0.150	-0.013	0.875	0.871	0.550	0.554	0.810	0.000
2nd	0.244	-0.017	0.901	0.897	0.590	0.595	0.824	0.000
3rd	0.168	0.009	0.859	0.856	0.697	0.663	0.217	0.000
Panel B. Employment Last Week (18+ years old)								
Pooled	0.157	0.007	0.463	0.468	0.472	0.486	0.698	0.132
1st	0.100	-0.007	0.464	0.471	0.472	0.473	0.825	0.860
2nd	0.195	0.001	0.463	0.465	0.474	0.496	0.566	0.058
3rd	0.175	0.027	0.463	0.469	0.471	0.481	0.769	0.503

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

Next, we examine the second outcome, adult employment, as observed at baseline. In

[23]Note that the two outcomes we examine here are binary, so the equality of means is equivalent to a distributional equality. 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.

contrast to school enrollment, adult employment is similar across all four treatment-response subgroups. This pattern indicates that the unobservables that determine the outcome are independent of treatment and response status. This is consistent with the identifying assumption for internal validity for the study population (the IV-P assumption), which we formally define in the next section. We then show that under random assignment the IV-P assumption implies the identification of treatment effects for the study population and its testable implication is indeed that the distribution of baseline outcome is identical across all four treatment-response subgroups. When we formally test the implication of the IV-P assumption for adult employment, we do not reject it (Column 8 of Table 3). Thus, we do not reject the assumption that ensures that the difference in mean employment rates between treatment and control respondents at follow-up identifies not only the ATE-R but also the average treatment effect (ATE). For the outcome of school enrollment, however, we do reject the IV-P assumption (Column 8 of Table 3), and thus the estimated treatment effect cannot be interpreted as internally valid for the study population. This is consistent with our previous observation that the children that are observed in the follow-up data are substantially different at baseline from those that are not.

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

3.2 Internal Validity in the Presence of Attrition

In this section, we derive the testable implications of our distributional and mean identifying assumptions. We also present the extension of the results to stratified randomization and

heterogeneous treatment effects, formally defined as conditional average treatment effects.

3.2.1 Internal Validity and its Testable Restrictions

In a field experiment with baseline outcome data, we observe individuals $i = 1, \ldots, n$ over two time periods, $t = 0, 1$. We will refer to $t = 0$ as the baseline period, and $t = 1$ as the follow-up period. Individuals are randomly assigned in the baseline period to the treatment and control groups. We use D_{it} to denote treatment status for individual i in period t, where $D_{it} \in \{0, 1\}$.[24] Hence, the treatment and control groups can be characterized by $D_i \equiv (D_{i0}, D_{i1}) = (0, 1)$ and $D_i = (0, 0)$, respectively. For notational brevity, we let an indicator variable T_i denote the group membership. Specifically, $T_i = 1$ if individual i belongs to the treatment group and $T_i = 0$ if individual i belongs to the control group.

For each period $t = 0, 1$, we observe an outcome Y_{it}, which is determined by the treatment status and a $d_U \times 1$ vector of time-invariant and time-varying variables, $U_{it} \equiv (\alpha_i^{\prime}, \eta_{it}^{\prime})^{\prime}$,

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

Given this structural function, we can define the potential outcomes $Y_{it}(d) = \mu_t(d, U_{it})$ for $d = 0, 1$. We use 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. To simplify illustration, we postpone the discussion of covariates to Section 3.4.1.

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. $Y_{i1}(0)|T_i = 1 \overset{d}{=} Y_{i1}|T_i = 0$, where $\overset{d}{=}$ denotes the equality in distribution. In this case, any difference in the outcome distribution

[24] The extension to the multiple treatment case is in Section SA4 of the online appendix.

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

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

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

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

where V_i denotes a vector of unobservables that determine response status and potential response can be defined as $R_i(\tau) = R_i(\tau, V_i)$ for $\tau = 0, 1$. If individual i responds, then $R_i = 1$, otherwise it is zero. As a result, 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. Random assignment in the presence of attrition, $(U_{i0}, U_{i1}, V_i) \perp T_i$, does not ensure that comparisons between treatment and control respondents are solely attributable to the treatment, since these comparisons are conditional on being able to observe individuals at follow-up ($R_i = 1$).[26]

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

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

[26] We use a random assignment condition similar to Lee (2009). Using potential outcome and response notation, we can express the random assignment condition as $(Y_{i0}(0), Y_{i0}(1), Y_{i1}(0), Y_{i1}(1), R_i(0), R_i(1)) \perp T_i$ which is similar to Lee (2009).

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

The next proposition provides sufficient conditions to obtain each of the aforementioned equalities as well as their respective sharp testable restrictions. Restrictions are sharp when they are the strongest implications that can be tested given the available data (see Figure 4). Part *a* (*b*) of the following proposition refers to the case where we can obtain valid estimands for the respondent subpopulation (study population). The proof of the proposition is given in Section C.

Proposition 1. *Assume* $(U_{i0}, U_{i1}, V_i) \perp T_i$.[27]

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

(i) (Identi cation) $Y_{i1}|T_i = 0, R_i = 1 \overset{d}{=} Y_{i1}(0)|T_i = 1, R_i = 1$

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

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

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

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

Proposition 1(a) relies on the assumption of random assignment conditional on response status (IV-R assumption). This assumption implies that the outcome distributions of treatment and control *respondents* at endline would have been the same if the treatment status had never been assigned. We refer to this equality (a.i) as *internal validity for the respondent subpopulation* (IV-R). When IV-R holds, the difference in means between treatment and control respondents identifies the ATE-R. IV-R cannot be tested directly, however, since

[27] The random assignment condition can be expressed as $(Y_{i0}(0), Y_{i0}(1), Y_{i1}(0), Y_{i1}(1), R_i(0), R_i(1)) \perp T_i$ in potential outcome and response notation.

treatment was in fact assigned. Thus, we derive a sharp testable restriction (a.ii) of the IV-R assumption, which exploits the information in the baseline data.[28] This restriction implies that the appropriate attrition test (when the object of interest is the treatment effect on the respondent subpopulation) is a *joint test* of the equality of the baseline outcome distribution between treatment and control respondents as well as treatment and control attritors.[29]

The assumption in Proposition 1(b), under random assignment, implies that treatment and response status are jointly independent of the unobservables in the outcome equation.[30] As a result, in the absence of treatment, all four treatment-response sub-groups would have the same outcome distribution. We refer to this case as *internal validity for the study population* (IV-P) and the assumption in (b) as the IV-P assumption. When IV-P holds, the ATE is identified, and so are quantile and other distributional treatment effects for the study population. The sharp testable restriction of the IV-P assumption under random assignment is given in (b.ii).

3.2.2 Mean Tests of Internal Validity

The vast majority of selective attrition tests implemented in the literature are based on restrictions on the mean of the baseline variables in question. The IV-R and IV-P assumptions

[28] While it is *theoretically* possible for identification to hold while the testable restriction is violated, it is not an interesting case empirically. If a field experimentalist finds violations of the testable implication of the IV-R (or IV-P) assumption at baseline, it is highly unlikely that they will discount this evidence and argue that identification of the ATE-R (or ATE) remains possible from a simple difference of means between treatment and control respondents.

[29] If IV-R is of interest, a natural question is whether one should simply test the implication of $(U_{i0}, U_{i1}) \perp T_i | R_i = 1$ in lieu of the IV-R assumption ($(U_{i0}, U_{i1}) \perp T_i | R_i$). This would be empirically relevant if it is plausible that $(U_{i0}, U_{i1}) \perp T_i | R_i = 1$ holds while $(U_{i0}, U_{i1}) \perp T_i | R_i = 0$ is violated. Using the subgroups defined by potential response status, we note that a primitive condition for this to hold is $(U_{i0}, U_{i1}) | (R_i(0), R_i(1)) \stackrel{d}{=} (U_{i0}, U_{i1}) | max\{R_i(0), R_i(1)\}$. This condition is not empirically plausible since it implies that the unobservable distribution is the same for always-responders, treatment-only and control-only responders, but different for the never-responders.

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

we present above ensure the identification of distributional treatment effects in addition to average treatment effects. In some experiments, however, researchers may be solely interested in average treatment effects. Here, we discuss the weaker conditions required to identify these objects and their sharp testable implications. Section B presents regression-based tests for these restrictions.

If the researcher is interested in mean impacts for the respondent subpopulation, then the IV-R assumption in Proposition 1(a), while sufficient, is stronger than required. A weaker condition that ensures that the average potential outcome without the treatment is identical for treatment and control respondents as well as treatment and control attritors, specifically

$$E[Y_{it}(0)|T_i, R_i] = E[Y_{it}(0)|R_i],\ t = 0, 1, \quad \text{(Mean IV-R Assumption)} \tag{4}$$

implies the identification of the ATE-R. Its sharp testable implication is the mean version of the testable restriction in Proposition 1(a.ii),

$$E[Y_{i0}|T_i, R_i] = E[Y_{i0}|R_i], \tag{5}$$

so it also includes testable restrictions on attritors and respondents. We will refer to a test of the mean equality restrictions in (5) as a mean IV-R test.

Similarly, if the object of interest is the ATE for the study population, then the relevant identifying assumption is

$$E[Y_{it}(d)|T_i, R_i] = E[Y_{it}(d)],\ d = 0, 1,\ t = 0, 1, \qquad \text{(Mean IV-P Assumption)} \tag{6}$$

which ensures that the average potential outcomes are identical across the four treatment-response subgroups. The sharp testable restriction of this assumption,

$$E[Y_{i0}|T_i, R_i] = E[Y_{i0}], \tag{7}$$

involves all treatment-response subgroups as its distributional version in Proposition 1(b.ii). We will refer to a test based on (7) as a mean IV-P test.

3.2.3 Heterogeneous Treatment Effects and Stratified Randomization

In this section, we extend our analysis to discuss heterogeneous treatment effects and stratified randomization. Heterogeneous treatment effects, more formally referred to as conditional average treatment effects (CATE), are of interest in many experiments. Stratified randomization is also common in empirical practice. Sometimes it is a necessity of the design, such as when the study is randomized within roll-out waves or locations. At other times, it is included in the experimental design with the aim of increasing precision and reducing bias of both average and heterogeneous treatment effects. The results in this section are relevant both for stratified randomized experiments and for completely randomized experiments that estimate heterogeneous treatment effects.[31]

In the following, let S_i denote the stratum of individual i which has support $\mathcal{S}$, where $|\mathcal{S}| < \infty$.[32] To exclude trivial strata, we assume that $P(S_i = s) > 0$ for all $s \in \mathcal{S}$ throughout the paper. In a stratified randomized experiment, random assignment is defined by $(U_{i0}, U_{i1}, V_i) \perp T_i | S_i$, whereas in a completely randomized experiment this conditional independence assumption holds as an implication of simple randomization ($(S_i, U_{i0}, U_{i1}, V_i) \perp T_i$). As a result, the following proposition applies to both completely and stratified randomized experiments.

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

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

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

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

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

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

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

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

$$\begin{aligned} &E[Y_{i1}(1) - Y_{i1}(0)|T_i = 1, S_i = s, R_i = 1] \\ =&E[Y_{i1}|T_i = 1, S_i = s, R_i = 1] - E[Y_{i1}|T_i = 0, S_i = s, R_i = 1]. \text{ (CATE-R)} \end{aligned} \tag{8}$$

Alternatively, the ATE-R can then be identified by averaging over S_i, i.e. $\sum_{s2\mathcal{S}} P(S_i = s|R_i = 1)\left(E[Y_{i1}|T_i = 1, S_i = s, R_i = 1] - E[Y_{i1}|T_i = 0, S_i = s, R_i = 1]\right)$. The testable restriction in (a.ii) is the identity of the distribution of baseline outcome for treatment and control groups conditional on response status *and* stratum. In other words, the equality of the outcome distribution for treatment and control respondents (as well as for treatment and control attritors) conditional on stratrum is the sharp testable restriction of the IV-R assumption in the case of block randomization. The results in part (b) of the proposition refer to IV-P in the context of block randomization. Thus, they are also conditional versions of the results in Proposition 1(b).

Randomization tests of the restrictions in Proposition 2(a.ii) and (b.ii) are provided in Section A, respectively. The key distinction between the randomization tests for stratified and completely randomized experiments is that in the former permutations are performed within strata.

3.3 Differential Attrition Rates and Internal Validity

The differential attrition rate test is the most widely used according to our review. Thus, we examine the relationship between internal validity and differential attrition rates ($P(R_i = 0|T_i = 1) \neq P(R_i = 0|T_i = 0)$). Our goal in this section is to formally understand the properties of the differential attrition rate test as a test of internal validity.

We first adapt the LATE framework (Imbens and Angrist, 1994; Angrist et al., 1996) to potential response. Specifically, in order to understand how treatment and control respondents and attritors consist of different response types, we modify the four types from the LATE literature: never-takers, always-takers, compliers and defiers. We establish four similar types as shown in Figure 2: 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 ($(R_i(0), R_i(1)) = (1, 0)$).

Figure 2: Respondent and Attritor Subgroups

	Control ($T_i = 0$)	Treatment ($T_i = 1$)
Attritors ($R_i = 0$)	Treatment-only responders Never responders	Control-only responders Never responders
Respondents ($R_i = 1$)	Control-only responders Always responders	Treatment-only responders Always responders

We can now examine the attrition rates in the treatment and control group and how they relate to the different response types. By random assignment, the distribution of response types is identical across treatment and control groups, $(R_i(0), R_i(1)) \perp T_i$. In other words, the treatment and control groups consist of the same proportion of never responders, treatment-only responders, control-only responders and always responders, which we denote by p_{00}, p_{01}, p_{10} and p_{11}, respectively. With the aid of Figure 2, we note that the attrition rate in the control group equals the proportion of never-responders and treatment-only responders, whereas the attrition rate in the treatment group equals the proportion of

never-responders and control-only responders, specifically

$$P(R_i = 0|T_i = 0) = p_{00} + p_{01}, \quad P(R_i = 0|T_i = 1) = p_{00} + p_{10}. \tag{9}$$

The difference in attrition rates across groups depends on 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, attrition rates are equal if the proportions of treatment-only and control-only responders are equal.

Next, we illustrate the relationship between differential attrition rates and the IV-R assumption (Proposition 1(a)), $(U_{i0}, U_{i1}) \perp T_i|R_i$. The proof of the proposition is given in Section C.

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 *⊥ 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 main takeaway from the above proposition is that equal attrition rates alone do not constitute a sufficient condition for internal validity. Proposition 3(i) provides a case in which equal attrition rates are not necessary for internal validity. The 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 the two other cases, (ii) and (iii), equal attrition rates together with an additional assumption imply the IV-R assumption. The monotonicity assumption in (ii) is from Lee (2009) and rules out control-only responders.

The exchangeability restriction allows for both treatment-only and control-only responders, but it assumes that these two types have the same distribution of (U_{i0}, U_{i1}). This assumption may be plausible in experiments with two treatments.

Using these insights, we now provide two simple examples that illustrate that differential attrition rates can coincide with internal validity (*Example 1*) and that equal attrition rates can coincide with a violation of internal validity (*Example 2*). In Section 4, we design simulation experiments that mimic both examples to illustrate these points numerically. Furthermore, we find several empirical cases in Section 5 that are consistent with the theoretical conditions of *Example 1*.

Example 1. *(Internal Validity & Di erential Attrition Rates)*

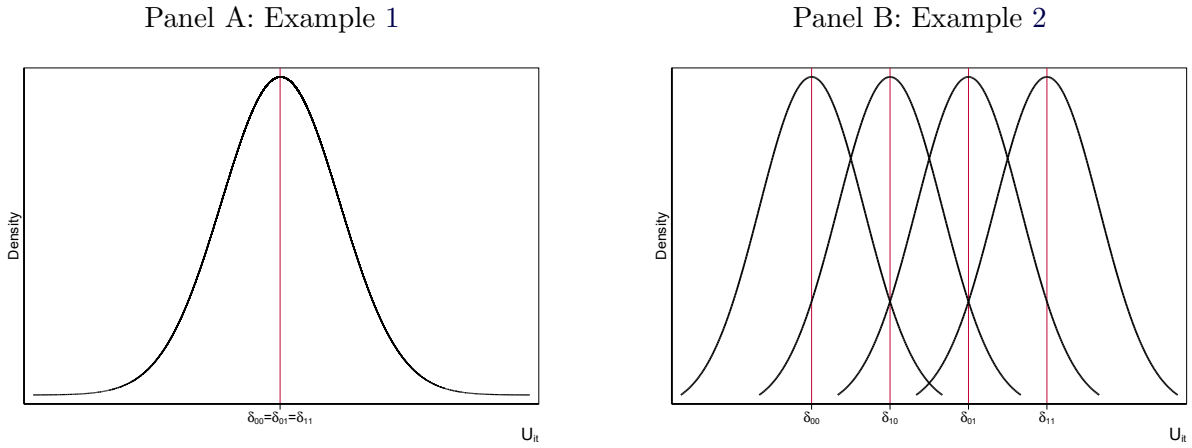
Assume that potential response satis es monotonicity, i.e. $p_{10} = 0$*, and all response types have the same unobservable distribution,* $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$*. Panel A of Figure 3 illustrates the resulting distribution of* U_{it}*. By the above proposition, IV-P holds under random assignment, since* $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1)) \Rightarrow (U_{i0}, U_{i1})|T_i, R_i \stackrel{d}{=} (U_{i0}, U_{i1})$*. Suppose that there is a group of individuals for whom it is too costly to respond if they are in the control group, so they only respond if assigned the treatment. Due to the presence of these treatment-only responders (*$p_{01} > 0$*), the attrition rates in the treatment and control groups are not equal, speci cally* $P(R_i = 0|T_i = 1) = p_{00}$*, and* $P(R_i = 0|T_i = 0) = p_{00} + p_{01}$*. This example thereby provides a case where we have di erential attrition rates even though not only IV-R but also IV-P holds. Under these conditions, the di erential attrition rate test would not control size as a test of internal validity as we illustrate in the simulation section.*

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

[33] 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

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



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

attrition rates across treatment and control groups.[34] *If* $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$*, then the di erent response types will have di erent 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 de ned in* (23)-(24) *will be di erent and hence IV-R is violated.*

A further limitation of the focus on the differential attrition rate test in empirical practice is that we cannot use it to test IV-P, even in cases where the differential attrition rate test is a valid test of IV-R. For instance, consider the case in which monotonicity holds and the attrition rates are equal across groups. We can then identify the ATE-R, since the respondent subpopulation is composed solely of always-responders as pointed out above. If the researchers are interested in identifying the treatment effect for the study population,

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.

[34] 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 SA3 in the online appendix for a derivation.

however, they 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).

3.4 Implications for Empirical Practice

Our theoretical analysis has multiple implications for empirical practice. For one, it underscores the importance of the object of interest in determining the appropriateness of an attrition test. Hence, explicitly stating the object of interest, whether it is the ATE-R, ATE, CATE-R or CATE, is important to justify a particular attrition test.

Our results further clarify the interpretation of attrition tests in the field experiment literature. The differential attrition rate test, which is implemented in 79% of papers in our review, is not based on a necessary condition of IV-R, and is not designed to test IV-P. Turning to the selective attrition tests, used in 60% of the papers, the null hypotheses are largely implications of the IV-R assumption (see Section SA2.2 in the online appendix). The most common version of this test (40% of all papers) uses respondents only; and hence, it does not exploit all the information in the baseline sample, specifically the attritors. Seventeen percent of papers do implement a selective attrition test that includes both respondents and attritors, suggesting that some authors are aware of the value of this information. Several of the null hypotheses they use, however, do not constitute IV-R or IV-P tests. This is perhaps unsurprising given the wide range of null hypotheses tested. Although authors do not in general conduct a direct test of IV-P, the inclusion of respondents and attritors in some selective attrition tests as well as the use of determinants of attrition tests suggest that some authors are likely interested in internally valid estimates for the study population. As we discuss in the empirical applications of Section 5, our results are promising for field experiments where treatment effects for the study population are of interest.

3.4.1 The Role of Covariates

An important question that arises in empirical practice is whether to include covariates in attrition tests. In our review of field experiments, we find that most authors use covariates in attrition tests regardless of the design of the study. While we restrict our review to experiments with baseline outcome data, there are settings where using covariates may be the only way to test attrition bias. In particular, some experiments target a population for which the baseline outcome always takes on the same value by design (e.g. if a job training program is targeted to unemployed people and employment is the main outcome). In other field experiments, baseline outcome data may not be available. We therefore provide a formal discussion of the role of covariates in attrition tests in this section.

Suppose that there is a set of covariates that are functions of the same determinants as the outcome, formally

$$W_{it} = \nu_t(U_{it}) \text{ for } t = 0, 1. \tag{10}$$

This definition pins down two types of covariates: (i) covariates that are themselves determinants of the outcome, i.e. $W_{it}^k = U_{it}^j$ for some k, j, $k = 1, \ldots, d_W$, $j = 1, \ldots, d_U$, or (ii) "proxy" variables, which are covariates determined by the same factors as the outcome Y_{it}. If this *a priori* information is true, the testable restrictions of the IV-R and IV-P assumptions would be on the joint distribution of $Z_{i0} = (Y_{i0}, W_{i0}^0)^0$.[35] However, if this *a priori* information is false, then including covariates may lead to a false rejection of the identifying assumption in question. In addition, we note that studies that implement the selective attrition tests on all baseline variables, $\mathcal{Z}_{i0} = (Y_{i0}, W_{i0}^0, X_{i0}^0)^0$, are testing the IV-R assumption for all variables in the survey as opposed to the outcome in question only. This IV-R assumption is a much stronger condition that may be violated, even if the IV-R assumption for the outcome in question holds.[36]

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

[36] Formally, the IV-R assumption relevant to all variables in the survey is $(\mathcal{E}_{i0}, \mathcal{E}_{i1}) \perp T_i | R_i$, where $\mathcal{Z}_{it} =$

Thus, if the baseline survey contains determinants of the outcome or proxy variables (W_{i0}), then they can be included in tests of the IV-R or IV-P assumption for the outcome in question. Our results suggest, however, that the inclusion of covariates that are not determined solely by the same unobservables as the outcome (X_{i0}) may lead to false rejection of the IV-R or IV-P assumption. This outcome-specific approach to including other variables in attrition tests is further supported by our *Progresa* example, which illustrates empirically that attrition may affect internal validity differently for two different outcomes collected in the same survey. Another reason for potential over-rejection of internal validity in the literature is that a substantial proportion of the implementation of selective attrition tests consists of individual tests for each baseline variable without correcting for multiple testing.

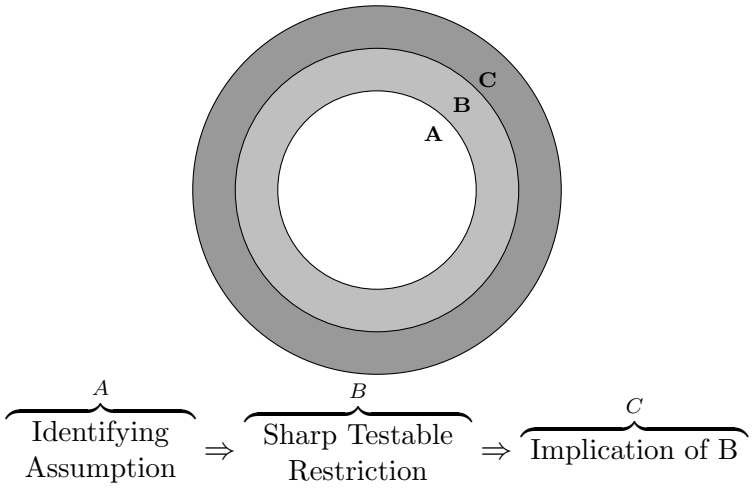
The implications of our analysis for empirical practice resonate with existing recommendations in the literature regarding the random assignment method used to ensure the similarity of treatment and control groups in terms of baseline observables in a given sample (i.e. "balance"). In seminal work on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

3.4.2 Attrition Tests as Identification Tests

Our approach emphasizes that attrition tests are identification tests. While rejection of such tests is clear evidence against the identifying assumption in question, it is possible to fail to reject such tests when the assumption is in fact violated. This is because in general we can only test identifying assumptions by implication. In other words, their testable restrictions

$\xi_t(\mathcal{E}_{it})$ and $\mathcal{E}_{it} = (U'_{it}, \eta'_{it})'$. However, the IV-R assumption that ensures identification of treatment effects for the outcome in question is weaker, since it imposes the conditional random assignment restriction on the unobservables relevant to that outcome only, U_{it}.

Figure 4: Graphical Illustration of Sharp Testable Restriction



are necessary, but not sufficient for the identifying assumption to hold.[37] Figure 4 graphically presents this issue. The light gray area represents cases where the identifying assumption is violated yet the sharp testable restriction holds true.

Figure 4 also illustrates that the sharp testable restriction is the strongest testable implication of the identifying assumption. Basing a test of the identifying assumption on another implication (C) leads to more cases where the implication holds yet the identifying assumption fails, represented by the dark gray area. Using sharp testable restrictions eliminates the cases in the dark gray area. The cases in the light gray area, which are unavoidable in general, complicate the interpretation of non-rejection of any identification test. Fortunately, our framework allows us to characterize the set of conditions under which this may or may not be a concern.

For both the IV-R and IV-P assumptions, there is a set of conditions in our setup under which identification holds if and only if the testable implication holds. These conditions consist of time homogeneity of the structural function and the unobservable distribution for the different treatment-response subpopulations (Chernozhukov et al., 2013).[38] This

[37] In Footnote 28, we elaborate on why the theoretical case where the testable restriction is violated while identification holds is not empirically relevant in our setting.

[38] Formally, $\mu_0(d,u) = \mu_1(d,u)$ and $U_{i0}|T_i, R_i \overset{d}{=} U_{i1}|T_i, R_i$.

assumption may be plausible in some field experiments where researchers do not expect the structural function or the determinants of the outcome to vary between the baseline and follow-up surveys. To provide a simple example, suppose that the outcome equation is determined by ability (U_i^1) and the opportunity cost of time (U_i^2), where the super-script is an index for the unobservables. We assume that both unobservables are time-invariant here to simplify notation. For a more general example with time-varying variables, see Section C.1. Now suppose that ability fulfils the IV-R assumption ($U_i^1 \perp T_i|R_i$), whereas the cost of time does not ($U_i^2 \not\perp T_i|R_i$). If ability ***and*** the cost of time both enter the baseline and follow-up outcomes, for instance,

$$Y_{i0} = U_i^1 + U_i^2$$
$$Y_{i1} = U_i^1 + U_i^2 + T_i(U_i^1 + U_i^2)$$

then comparisons between treatment and control respondents at follow-up would not be solely attributable to the treatment. Baseline outcome data would allow us to detect a violation of internal validity by comparing treatment and control respondents as well as treatment and control attritors.

Now let us consider a case where baseline outcome data would not help us detect such a violation of internal validity. This would require baseline outcome to only be a function of ability and not the cost of time, which only determines the outcome in the follow-up period,

$$Y_{i0} = U_i^1$$
$$Y_{i1} = U_i^1 + U_i^2 + T_i(U_i^1 + U_i^2).$$

Since ability fulfils the IV-R assumption, when comparing baseline outcome data of treatment and control respondents as well as treatment and control attritors, we would not detect any substantial differences between these subgroups, even though internal validity is violated.[39]

[39] An interesting case that we illustrate in Section C.1 is that if the cost of time only interacts with the

While we focus the example on the IV-R assumption, similar arguments can be made for the IV-P assumption.

A practical implication of our analysis is that when interpreting non-rejection of tests of the IV-R or IV-P assumptions, practitioners should consider whether the relationship between the outcome and its determinants may have changed over the time span between baseline and follow-up periods.

4 Simulation Study

We illustrate the theoretical results in the paper using a numerical study. The simulations examine 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.

4.1 Simulation Design and Test Statistics

The data-generating process (DGP) is described in Panel A of Table 4. We assign individuals to one of the four response types: always-responders, never-responders, control-only responders, and treatment-only responders. The unobservables that determine the outcome consist of time-invariant and time-varying components. We introduce dependence between the unobservables in the outcome equation and potential response by allowing the means of the time-invariant component to differ for each response type. We also allow for heterogeneous treatment effects, so that the ATE-R can differ from the ATE.

We conduct simulations using four variants of this simulation design that feature different cases of IV-R and IV-P as summarized in Panel B of Table 4.[40] Designs I and II present

treatment, the difference in mean outcome between treatment and control respondents identifies an internally valid estimand that is not equal to the ATE-R.

[40] 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)).

cases where the differential rate test would have desirable properties as a test of IV-R.[41] Both designs allow for dependence between the unobservables in the outcome equation and potential response and impose monotonicity in the response equation by ruling out control-only responders. Design I allows for non-zero proportions of treatment-only responders and thereby a violation of IV-R. Design II rules out treatment-only responders and, as a result, we have IV-R, but not IV-P.

Table 4: Simulation Design

Panel A. Data-Generating Process

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

Panel B. Variants of the Design

Design	I	II	III	IV
Monotonicity in the Response Equation	Yes	Yes	Yes	No
Equal Attrition Rates	No	Yes	No	Yes
IV-R Assumption	No	Yes	Yes	No
IV-P Assumption ($(U_{i0}, U_{i1}) \perp R_i$)	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 $P(R_i = 0|T_i = 0) = p_{00} + p_{10}$.

Designs III and IV illustrate *Examples 1* and *2* in Section 3.3, respectively. Design III

[41] To be precise, in these designs, the differential attrition rate test would have non-trivial power when IV-R is violated while controlling size when IV-R holds.

demonstrates a setting in which we have differential attrition rates and IV-P. It imposes monotonicity and differential attrition rates as in Design I, but allows the unobservables in the outcome equation and potential response to be independent. Finally, Design IV follows *Example 2* in demonstrating a case in which there are equal attrition rates and a violation of internal validity. Here, we allow for a violation of monotonicity and dependence between the unobservables in the outcome equation and potential response. We impose that the proportion of treatment-only and control-only responders is identical and, as a result, the design features equal attrition rates.

In all four designs, 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. To illustrate the implication of the designs for estimated mean effects, we report the simulation mean and standard deviation of the estimated difference in mean outcomes for the treatment and control respondents in the follow-up period ($\bar{Y}_1^{TR} - \bar{Y}_1^{CR}$).

The primary goal of our simulation analysis is to compare the performance of the differential attrition rate test as well as the mean and distributional IV-R and IV-P tests using a 5% level of significance. The differential attrition rate test is a two-sample t-test of the equality of attrition rates between the treatment and control group, $P(R_i = 0|T_i) = P(R_i = 0)$. The hypotheses of the mean IV-R and IV-P tests (denoted with an $\mathcal{M}$ subscript) are given by:

$$Y_{i0} = \gamma_{11} T_i R_i + \gamma_{01}(1 - T_i)R_i + \gamma_{10} T_i(1 - R_i) + \gamma_{00}(1 - T_i)(1 - R_i) + \epsilon_i \tag{11}$$

$$H_{0,M}^{1,1} : \quad \gamma_{10} = \gamma_{00}, \qquad (CR\text{-}TR)$$

$$H_{0,M}^{1,2} : \quad \gamma_{11} = \gamma_{01}, \qquad (CA\text{-}TA)$$

$$H_{0,M}^{1} : \quad \gamma_{10} = \gamma_{00} \;\&\; \gamma_{11} = \gamma_{01}, \qquad (IV\text{-}R) \tag{12}$$

$$H_{0,M}^{2} : \; \gamma_{11} = \gamma_{01} = \gamma_{10} = \gamma_{00}, \qquad (IV\text{-}P) \tag{13}$$

$H_{0,M}^{1,1}$ ($H_{0,M}^{1,2}$) tests the significance of mean differences between the treatment and control

respondents (attritors) only. These two hypotheses are similar to widely used tests in the literature and are both implications of the IV-R assumption. $H^1_{0,M}$ ($H^2_{0,M}$) are the hypotheses of the mean IV-R (IV-P) tests in Section 3.2.2, which we implement using Wald statistics and asymptotic χ^2 critical values. To implement the distributional IV-R and IV-P tests, we use Kolmogorov-Smirnov-type (KS) statistics of their respective hypotheses,

$$H^1_0: \quad Y_{i0}|T_i, R_i = r \overset{d}{=} Y_{i0}|R_i = r, \text{ for } r = 0, 1, \tag{14}$$

$$H^2_0: \quad Y_{i0}|T_i, R_i \overset{d}{=} Y_{i0}. \tag{15}$$

We formally define the KS statistics for the above hypotheses in Section A.1, where we also describe the randomization procedures we use to obtain their p-values.

4.2 Simulation Results

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 assumptions for Designs I-IV. 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. Designs I and II, which obey monotonicity and allow for dependence between the unobservables in the outcome equation and potential response, illustrate the typical cases in which the differential attrition rate test can be viewed as a test of IV-R. In Design I, where internal validity is violated, the test rejects above 5%, while in Design II, where IV-R holds, the test controls size. 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 differential attrition rate test rejects at a frequency higher than 5% simply because the attrition rates are different even though IV-P holds. In Design IV, however, the differential attrition rate test does not reject above 5% when internal validity is violated because attrition rates are

equal.

Next, we examine the performance of the IV-R tests, which are given in Columns 4 through 7 of Table 5. As expected, where IV-R holds (Designs II and III), the tests control size. Similarly, where IV-R is violated (Designs I and IV), the tests reject above 5%. In general, the relative power of the test statistics may differ depending on the DGP. In our simulation design, however, the rejection probabilities of the attritors-only test (CA-TA) and the joint tests ($Mean$ and KS) are significantly higher than the test based on the difference between the treatment and control respondents (CR-TR).[42]

The test statistics of the IV-P assumption (Columns 8 and 9 in Table 5) also behave according to our theoretical predictions. In Designs I, II and IV, where there is dependence between the unobservables in the outcome equation and potential response, the IV-P test rejects above 5%. Of particular interest is Design II, since internal validity holds for the respondents, but not for the population (i.e. IV-R holds, but IV-P does not). Thus, although the IV-P test does reject, the IV-R test does not reject above 5%. In this case, the difference in mean outcomes between treatment and control respondents (i.e. the estimated treatment effect) is not unbiased for the ATE (0.25), but 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 Kolmogorov-Smirnov and Cramer-von-Mises statistics, see Section SA5 in the online appendix.

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

5 Empirical Applications

To complement our simulation analysis, we apply the proposed tests of attrition bias to five published field experiments. The data comes from field experiments with both high attrition rates and publicly available data that includes attritors.[43] Thus, the exercise is not intended to draw inference about implications of applying various attrition tests to a representative sample of published field experiments. In addition, field experiments that are published in prestigious journals may not to be representative of all field experiment data, especially if perceptions of attrition bias had an impact on publication.

5.1 Implementation of Attrition Tests

Across the five selected articles included in this exercise, we conduct attrition tests for a total of 33 outcomes. This includes all outcomes with baseline data that are reported in the abstracts as well as all other unique outcomes with baseline data.[44] For each outcome included in this exercise, the appropriate attrition test depends on the type of outcome and the approach to randomization used in the experiment. For fully randomized experiments, we apply the 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.[45] For binary outcomes and also for all outcomes from clustered experiments, we apply regression-based mean tests (see Section B). For continuous outcomes in non-clustered experiments, we report p-values of the KS distributional tests using the appropriate randomization procedure.[46] For all tests, the results are presented in a way that is designed to preserve the anonymity of the

[43] We selected the articles with the five highest attrition rates for which the data required to implement the attrition tests is available (see Section SA1.2 in the online appendix for details).

[44] 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.

[45] 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.

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

results and papers. Thus, attrition rates are presented as ranges, the results are not linked to specific articles, and we randomize the order of the outcomes such that they are not listed by paper.

In addition to applying our proposed attrition tests, we also consider how those tests might compare to other approaches. Thus, we apply a version of the tests commonly used in the literature to the data, including: the differential attrition rate test, the IV-R test using the respondent subsample only, and the IV-R test using the attritor subsample only. We use the same approaches to handling stratification and continuous outcomes in all three IV-R tests to ensure they are directly comparable, but that also means that we do not necessarily replicate the exact tests that are used in the articles from which we drew data for this exercise. Instead, we indicate whether authors' attrition tests reject for the outcomes for which they are available.

In keeping with our findings from Section 2, there is heterogeneity in the application of attrition tests across these articles. Two of the articles only report a differential attrition rate test, one article only reports a selective attrition test and two report both. The differential attrition rate test used by authors is based on survey-level attrition rates. As for the selective attrition test, each of the three articles that conducts such a test relies on a different implementation. One article uses a selective attrition test that neither constitutes an IV-R nor an IV-P test. The two other articles examine experiments that are randomized within strata. One article includes strata fixed effects in its selective attrition test in line with the IV-R tests implied by our analysis, whereas the other does not, and thus does not account for the stratification of the experimental design.

5.2 Results of the Empirical Applications

Our IV-R and IV-P test results reported in Table 6 have promising implications for the internal validity of randomized experiments. The joint IV-R test does not reject for any of the 33 outcomes at the 5% level. The IV-R tests using only respondents or attritors yield

the same conclusion for all 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. The IV-P test also does not reject the IV-P assumption at the 5% level for 26 out of the 33 outcomes (28 when accounting for multiple hypothesis testing).[47] While keeping in mind the usual caveats regarding the power of any test in finite samples, our results suggest that a researcher interested in treatment effects for the respondent subpopulation would not reject the relevant identifying assumption for any of the outcomes in our analysis, even when exploiting all the information in the baseline sample (i.e. respondents and attritors). It is particularly notable that, for a majority of the outcomes we consider, a researcher would also not reject the identifying assumption that ensures the identification of the treatment effects for the study population.

Given its wide use in empirical practice, we also implement the differential attrition rate test. Using outcome-level attrition rates, we reject the null hypothesis of equal attrition rates at the 5% level for 9 of 33 outcomes (3 outcomes after correcting for multiple hypothesis testing).[48] For all 9 outcomes, the differential attrition rate test rejects the null hypothesis at the 5% level, whereas the IV-P assumption is not rejected at the 5% level using our test. These empirical cases are consistent with the testable implications of *Example 1*. Thus, according to our theoretical analysis, a researcher using the differential attrition rate test may falsely reject not only IV-R but also IV-P for these outcomes.

Next, we consider the results of the attrition tests reported by the authors (Table 6). The authors report a differential attrition rate test that is relevant to 30 out of the 33 outcomes and a selective attrition test for 8 outcomes. The reported differential attrition rate tests are rejected at the 5% level for 23 outcomes. The higher frequency of rejections of the authors' differential attrition rate test relative to ours is driven by their use of survey-

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

[48] The relatively high differential attrition rates we find in this exercise are perhaps not surprising, given that overall attrition rates and differential attrition rates seem to be correlated, and these outcomes have fairly high attrition rates (McKenzie, 2019).

level, as opposed to outcome-level, attrition rates. In the three articles in which the authors conduct a selective attrition test, they largely do not find evidence of selective attrition. They do, however, reject their version of the test at the 10% level for 2 of the 8 outcomes.

When we compare our test results with the authors', we note several differences. While we do not reject the IV-R assumption for any of the outcomes we consider, the authors reject their survey-level differential attrition rate test for 23 outcomes. Once we account for outcome-level attrition, we only reject equal attrition rates for 9 outcomes. As we note above, in all of these cases, our IV-P (or IV-R) test does not reject, which suggests that the differential attrition rate test is likely falsely rejecting internal validity for these outcomes. In addition, authors do not consistently account for the stratification of the experimental design in their selective attrition test, which may lead to a false rejection of internal validity.[49] Furthermore, one of the selective attrition tests used in the articles we examine does not constitute an IV-R or IV-P test. One limitation in comparing our results with the authors' is that, since they do not state their object of interest, it is not clear whether they intend to test for IV-R or IV-P.

Thus, we draw several conclusions from this empirical exercise. Our analysis illustrates that the differential attrition rate test may lead to over-rejection of internal validity in practice. Furthermore, our empirical analysis highlights the disadvantages of the lack of consensus in empirical practice. Selective attrition tests are not universally implemented. The heterogeneity in the implementation of selective attrition tests could lead empirical researchers to unnecessarily question the internal validity of their study. In contrast, for all outcomes we consider, the results of our proposed joint IV-R test would not reject the identifying assumption that allows them to interpret their treatment effects as internally valid

[49] To provide a simple example, consider a case where there are two strata (men and women). For simplicity, assume all men respond in the follow-up period. Now suppose 10% (5%) of women in the control (treatment) group do not respond to the follow-up survey, but the unobservables that affect outcome are independent of response. As a result, the treatment and control respondents consist of different proportions of men and women. It follows that, even though women in the different treatment-response subgroups have the same mean baseline outcome, the pooled treatment and control respondents may differ in that regard. Thus, a regression-based IV-R test that does not account for the stratification may falsely reject internal validity.

for the respondent subpopulation. If researchers were interested in the study population, our IV-P test results suggest that the data do not reject the identifying assumption in question for the majority of the outcomes in this exercise. Building on the *Progresa* example, our empirical exercise provides several additional cases where attrition impacts outcomes in the same experiment differently. These findings further highlight the advantages of our testing framework that allows the empirical researcher to align their attrition testing procedure with the outcome and population of interest.

6 Conclusion

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

While this paper is a step forward toward understanding current empirical practice and establishing a standard in testing attrition bias in field experiments, it opens several questions for future research. Despite the availability of several approaches to correct for attrition bias (Lee, 2009; Huber, 2012; Behagel et al., 2015; Millán and Macours, 2019), alternative approaches that exploit the information in baseline outcome data as in the framework here may require weaker assumptions and hence constitute an important direction for future work.

The extension of the analysis in this paper to the problem of attrition in the presence of partial compliance is another interesting direction. Furthermore, several practical aspects of the implementation of the proposed test may lead to pre-test bias issues. For instance, the proposed tests may be used in practice to inform whether an attrition correction is warranted or not in the empirical analysis. Empirical researchers may also be interested in first testing the identifying assumption for treatment effects for the respondent subpopulation and then testing their validity for the entire study population. Inference procedures that correct for these and other pre-test bias issues are a priority for future work.

Finally, this paper has several policy implications. Attrition in a given study is often used as a metric to evaluate the study's reliability to inform policy. For instance, *What Works Clearinghouse*, an initiative of the U.S. Department of Education, has specific (differential) attrition rate standards for studies (IES, 2017). Our results indicate an alternative approach to assessing potential attrition bias. Furthermore, questions regarding external validity of treatment effects measured from field experiments are especially important from a policy perspective. This paper points to the possibility that in the presence of response problems, the identified effect in a given field experiment may only be valid for the respondent subpopulation, and hence may not identify the ATE for the study population. This is an important issue to consider when synthesizing results of field experiments to inform policy.

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

Design	Attrition Rates		Differential Attrition Rate Test	Tests of the IV-R Assumption				Tests of the IV-P Assumption		Difference in Mean Outcomes between Treatment & Control Respondents ($\bar{Y}_1^{TR} - \bar{Y}_1^{CR}$)		
				Mean Tests			KS Test	Mean Test	KS Test			
	C	T	$\hat{p}_{0.05}$	CR-TR	CA-TA	Joint	Joint	Joint	Joint	Mean	SD	$\hat{p}_{0.05}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$												
I	0.05	0.025	0.866	0.049	0.446	0.353	0.324	0.452	0.476	0.265	0.057	0.997
	0.10	0.05	0.995	0.076	0.719	0.635	0.582	0.792	0.787	0.282	0.058	0.998
	0.15	0.10	0.935	0.072	0.631	0.542	0.483	0.995	0.980	0.288	0.061	0.997
	0.20	0.15	0.867	0.072	0.532	0.442	0.412	1.000	1.000	0.296	0.063	0.996
	0.30	0.20	1.000	0.141	0.894	0.851	0.801	1.000	1.000	0.334	0.066	0.999
Equal Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))^{\dagger}$												
II	0.05	0.05	0.049	0.046	0.044	0.053	0.062	0.981	0.902	0.255	0.058	0.993
	0.10	0.10	0.053	0.043	0.045	0.045	0.056	1.000	0.999	0.262	0.060	0.991
	0.15	0.15	0.052	0.043	0.049	0.052	0.055	1.000	1.000	0.271	0.062	0.992
	0.20	0.20	0.049	0.045	0.047	0.050	0.050	1.000	1.000	0.280	0.064	0.990
	0.30	0.30	0.048	0.053	0.044	0.046	0.043	1.000	1.000	0.303	0.068	0.991
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$ (*Example 1*)*												
III	0.05	0.025	0.866	0.055	0.051	0.056	0.052	0.065	0.050	0.248	0.058	0.990
	0.10	0.05	0.995	0.055	0.050	0.055	0.046	0.053	0.055	0.248	0.059	0.985
	0.15	0.10	0.935	0.057	0.052	0.053	0.045	0.053	0.059	0.247	0.061	0.983
	0.20	0.15	0.867	0.058	0.047	0.053	0.046	0.048	0.048	0.247	0.063	0.974
	0.30	0.20	1.000	0.057	0.053	0.052	0.043	0.049	0.048	0.248	0.066	0.964
Equal Attrition Rates + Violation of Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$ (*Example 2*)												
IV	0.05	0.05	0.012	0.067	0.429	0.337	0.329	0.360	0.311	0.273	0.058	0.997
	0.10	0.10	0.013	0.131	0.708	0.653	0.577	0.708	0.582	0.302	0.059	0.999
	0.15	0.15	0.007	0.248	0.873	0.855	0.758	0.888	0.792	0.333	0.061	0.999
	0.20	0.20	0.004	0.422	0.934	0.951	0.859	0.970	0.913	0.367	0.063	0.999
	0.30	0.30	0.001	0.797	0.990	0.997	0.974	0.999	0.998	0.452	0.067	1.000

Notes: The above table reports simulation summary statistics for $n = 2,000$ across $2,000$ simulation replications. C denotes the control group, T denotes the treatment group, and $\hat{p}_{0.05}$ denotes the simulation rejection probability of a 5% test. The *Mean* tests of the IV-R (IV-P) assumption refer to the regression tests (Section B) of the null hypothesis in (12) ((13)). The KS statistics of the IV-R (IV-P) assumption are given in (17) ((19)), and their p-values are obtained using the proposed randomization procedures in Section A.1 ($B = 199$). 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, $\bar{Y}_1^{TR} - \bar{Y}_1^{CR} = \frac{\sum_{i=1}^n Y_{i1} D_{i1} R_i}{\sum_{i=1}^n D_{i1} R_i} - \frac{\sum_{i=1}^n Y_{i1}(1-D_{i1}) R_i}{\sum_{i=1}^n (1-D_{i1}) R_i}$. All tests are conducted using $\alpha = 0.05$. Additional details of the design are provided in Table 4.
† (*) indicates IV-R only (IV-P).

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

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

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

A 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.[50] They can also be adapted to accommodate possibly discrete or mixed outcome distributions, which may result from rounding or censoring in the data collection, by applying the procedure in Dufour (2006). In this section, we focus on distributional statistics for the testable restrictions on the baseline outcome as in Propositions 1 and 2. The randomization procedures we propose, however, can be applied to test joint distributional hypotheses that include covariates as in Section 3.4.1.

We first outline a general randomization procedure that we adapt to the different settings we consider.[51] Given a dataset $\mathbf{Z}$ and a statistic $T_n = T(\mathbf{Z})$ that tests a null hypothesis H_0, we use the following procedure to provide a stochastic approximation of the exact p-value for the test statistic T_n exploiting invariant transformations $g \in \mathcal{G}_0$ (Lehmann and Romano, 2005, Chapter 15.2). Specifically, the transformations $g \in \mathcal{G}_0$ satisfy $\mathbf{Z} \overset{d}{=} g(\mathbf{Z})$ under H_0 only.

Procedure 1. *(Randomization)*

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

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

A.1 Completely Randomized Trials

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

$$H_0^1 : F_{Y_{i0}|T_i=0,R_i=r} = F_{Y_{i0}|T_i=1,R_i=r} \text{ for } r = 0, 1. \tag{16}$$

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

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

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

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

where for a random variable X_i, F_{n,X_i} denotes the empirical cdf, i.e. the sample analogue of F_{X_i}, and $\|.\|$ denotes some non-random or random norm. Different choices of the norm give rise to different statistics. For instance, the KS and CM statistics are the most widely known and used. The former is obtained by using the L^1 norm over the sample points, i.e. $\|f\|_{n,1} = \max_i |f(y_i)|$, whereas the latter is obtained by using an L^2 norm, i.e. $\|f\|_{n,2} = \sum_{i=1}^n f(y_i)^2/n$. In order to test the ***joint*** hypothesis in (16), the two following statistics that aggregate over $T^1_{n,r}$ for $r = 0, 1$ are standard choices in the literature (Imbens and Rubin, 2015),[52]

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

The joint KS statistic we use to test H^1_0 in the simulation and empirical section is given by

$$KS^1_{n,m} = \max\{KS^1_{n,0}, KS^1_{n,1}\}, \text{where for } r = 0, 1$$
$$KS^1_{n,r} = \max_{i:R_i=r} \left|\sqrt{n}\left(F_{n,Y_{i0}}(y_{i0}|T_i = 1, R_i = r) - F_{n,Y_{i0}}(y_{i0}|T_i = 0, R_i = r)\right)\right|. \tag{17}$$

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

We now consider testing the restriction of the IV-P assumption stated in Proposition 1(b.ii). This restriction implies that the distribution of the baseline outcome variable is identically distributed across all four subgroups defined by treatment and response status. Let $(T_i, R_i) = (\tau, r)$, where $(\tau, r) \in \mathcal{T} \times \mathcal{R} = \{(0,0), (0,1), (1,0), (1,1)\}$ and (τ_j, r_j) denote the j^{th} element of $\mathcal{T} \times \mathcal{R}$. Then, the joint hypothesis is given wlog by

$$H^2_0 : \; F_{Y_{i0}|T_i=\tau_j,R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1},R_i=r_{j+1}} \text{ for } j = 1, \ldots, |\mathcal{T} \times \mathcal{R}| - 1. \tag{18}$$

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

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

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

[52]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 SA5 of the online appendix.

simulation and empirical sections, we use the following KS statistic to test H_0^2

$$KS_n^2 = \max_{j=1,2,3} KS_{n,j}^2, \text{ where} \tag{19}$$
$$KS_{n,j}^2 = \max_i \left| \sqrt{n} \left(F_{n,Y_{i0}}(y_{i0}|T_i = \tau_j, R_i = r_j) - F_{n,Y_{i0}}(y_{i0}|T_i = \tau_{j+1}, R_i = r_{j+1}) \right) \right|.$$

and $\{\tau_j, r_j\}$ is the j^{th} element of $\mathcal{T} \times \mathcal{R} = \{(0,0),(0,1),(1,0),(1,1)\}$.

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.

A.2 Stratified Randomized Trials

As pointed out in Section 3.2.3, 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 A.1.

We first consider the restriction in Proposition 2(a.ii), which yields the following null hypothesis

$$H_0^{1,\mathsf{S}} : F_{Y_{i0}|T_i=0,S_i=s,R_i=r} = F_{Y_{i0}|T_i=1,S_i=s,R_i=r} \text{ for } r = 0,1,\ s \in \mathcal{S}. \tag{20}$$

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

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

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

$$T_{n,m}^{1,\mathsf{S}} = \max_{s2\mathsf{S}} T_{n,m,s}^{1,\mathsf{S}},$$
$$T_{n,p}^{1,\mathsf{S}} = \sum_{s2\mathsf{S}} p_n^s T_{n,p,s}^{1,\mathsf{S}}, \text{ where } p_n^s = \sum_{i=1}^n 1\{S_i = s\}/n \text{ for } s \in \mathcal{S}.$$

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

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

$$H_0^{2,\mathsf{S}} : F_{Y_{i0}\mathsf{j}T_i=\tau_j,S_i=s,R_i=r_j} = F_{Y_{i0}\mathsf{j}T_i=\tau_{j+1},S_i=s,R_i=r_{j+1}} \text{ for } j = 1,\ldots,|\mathcal{T}\times\mathcal{R}|-1,\ s\in\mathcal{S}. \tag{21}$$

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

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

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

$$T_{n,m}^{2,\mathsf{S}} = \max_{s2\mathsf{S}} T_{n,m,s}^{2,\mathsf{S}},$$
$$T_{n,p}^{2,\mathsf{S}} = \sum_{s2\mathsf{S}} w_s T_{n,p,s}^{2,\mathsf{S}},$$

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

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

The aforementioned subgroup-randomization procedures split the original sample into respondents and attritors or four treatment-response groups. This approach does not directly extend to cluster randomized experiments.[53] 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.

[53]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.

B Regression Tests of Internal Validity

In this section, we show how to implement the mean IV-R and IV-P tests using regression-based procedures. In completely and cluster randomized experiments, the null hypothesis of the IV-R test ($H^1_{0,M}$) consists of the equality of means across treatment and control responders as well as treatment and control attritors. Meanwhile, the null hypothesis of the IV-P test ($H^2_{0,M}$) consists of the equality of means across all treatment/respondent subgroups. In the stratified randomization case, the null hypotheses of the IV-R and IV-P tests consist of analogous restrictions *within* strata, $H^{1,S}_{0,M}$ and $H^{2,S}_{0,M}$, respectively. Here, we present these hypotheses as joint restrictions on linear regression coefficients, which are straightforward to test using the appropriate standard errors. The Stata ado file to implement those regression-based tests is available at `https://github.com/daghanem/ATTRITIONTESTS`.

B.1 Completely and Cluster Randomized Experiments

If the experiment is completely or cluster randomized and Y_{i0} is the baseline outcome, the practitioner may implement one of two equivalent approaches to conducting the mean tests. The first approach is given by:

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

The second approach allows for an intercept in the regression, which captures the mean baseline outcome for the control attritors:

$$\begin{aligned} Y_{i0} &= \alpha + \beta_{01} R_i + \beta_{10} T_i + \beta_{11} T_i R_i + \epsilon_i \\ H^1_{0,M} &: \beta_{10} = \beta_{11} = 0, \\ H^2_{0,M} &: \beta_{01} = \beta_{10} = \beta_{11} = 0. \end{aligned}$$

In some cases, the practitioner may have collected baseline data on determinants of (or proxies for) the outcome of interest, W_{i0} (as defined in Equation 10). If the practitioner chooses to include these determinants in testing for attrition bias, the regression-based procedure should test the joint hypotheses across the baseline outcome (if available) and the d_W baseline covariates that are relevant for such outcome, i.e. $Z_{i0} = (Y_{i0}, W^{0}_{i0})^{0}$, $\forall j = 1, \ldots, (d_W + 1)$.

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

As in the univariate case above, the null hypotheses in this multivariate case can also be tested using the specification that inlcudes an intercept. Note that if the researcher is interested instead in testing across multiple *outcomes* we recommend testing these individually

rather than jointly (as in Section 3.1), while accounting for multiple testing.

B.2 Stratified Randomized Experiments

As in Section B.1, we again present two equivalent formulations of the tests for stratified experiments. In these fully saturated models, the null hypotheses test the equality of means *within* strata. The first version of the test is given by:

$$Y_{i0} = \sum_{s2S} \left[\gamma_{11}^s T_i R_i + \gamma_{10}^s T_i (1 - R_i) + \gamma_{01}^s (1 - T_i) R_i + \gamma_{00}^s (1 - T_i)(1 - R_i)\right] 1\{S_i = s\} + \epsilon_i$$

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

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

In this case, the equivalent formulation uses a model with strata fixed effects and strata-specific coefficients,

$$\begin{aligned} Y_{i0} &= \sum_{s=1}^{S} \left\{\alpha^s + \beta_{01}^s R_i + \beta_{10}^s T_i + \beta_{11}^s T_i R_i\right\} 1\{S_i = s\} + \epsilon_i \\ H_{0,M}^{1,S} &: \beta_{10}^s = \beta_{11}^s = 0, \text{for all } s \in \mathcal{S}, \\ H_{0,M}^{2,S} &: \beta_{01}^s = \beta_{10}^s = \beta_{11}^s = 0, \text{for all } s \in \mathcal{S}. \end{aligned}$$

When the number of strata is large, however, testing the equality of means across groups *within* each stratum may result in high-dimensional inference issues. In that case, practitioners can instead test implications of $H_{0,M}^{1,S}$ and $H_{0,M}^{2,S}$ as follows:

$$\begin{aligned} Y_{i0} &= \sum_{s=1}^{S} (\alpha^s + \beta_{01}^s R_i) 1\{S_i = s\} + \pi_{10} T_i + \pi_{11} T_i R_i + \epsilon_i \\ H_{0,M}^{1',S} &: \pi_{10} = \pi_{11} = 0, \\ Y_{i0} &= \sum_{s=1}^{S} \alpha^s 1\{S_i = s\} + \pi_{01} R_i + \pi_{10} T_i + \pi_{11} T_i R_i + \epsilon_i \\ H_{0,M}^{2',S} &: \pi_{01} = \pi_{10} = \pi_{11} = 0. \end{aligned}$$

If the practitioner chooses to include baseline covariates for a stratified experiment, as in Section B.1, she should test the joint hypotheses across the baseline outcome and all relevant baseline covariates.

C Proofs

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

Following Hsu et al. (2019), we show that the testable restriction is sharp by showing that if $(Y_{i0}, Y_{i1}, T_i, R_i)$ satisfy $Y_{i0}|T_i = 0, R_i = r \overset{d}{=} Y_{i0}|T_i = 1, R_i = r$ for $r = 0, 1$, then there exists (U_{i0}, U_{i1}) such that $Y_{it}(d) = \mu_t(d, U_{it})$ for some $\mu_t(d, .)$ for $d = 0, 1$ and $t = 0, 1$, and $(U_{i0}, U_{i1}) \perp T_i|R_i$ that generate the observed distributions. By the arbitrariness of U_{it} and μ_t, we can let $U_{it} = (Y_{it}(0), Y_{it}(1))'$ and $\mu_t(d, U_{it}) = dY_{it}(1) + (1-d)Y_{it}(0)$ for $d = 0, 1$, $t = 0, 1$. Note that $Y_{i0} = Y_{i0}(0)$ since $D_{i0} = 0$ w.p.1. Now we need to construct a distribution of $U_i = (U_{i0}', U_{i1}')$ that satisfies

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

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

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

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

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

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

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

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

To show that the testable restriction is sharp, it remains to show that if $(Y_{i0}, Y_{i1}, T_i, R_i)$

satisfies $Y_{i0}|T_i, R_i \stackrel{d}{=} Y_{i0}(0)$, then there exists (U_{i0}, U_{i1}) such that $Y_{it}(d) = \mu_t(d, U_{it})$ for some $\mu_t(d,.)$ for $d = 0, 1$ and $t = 0, 1$, and $(U_{i0}, U_{i1}) \perp (T_i, R_i)$. Similar to (a.ii), we let $U_{it} = (Y_{it}(0), Y_{it}(1))^{\emptyset}$ 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, $F_{Y_{i0}jT_i,R_i} = F_{Y_{i0}}$ by construction and it follows immediately that

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

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

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

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

$$
F_{U_ijT_i,R_i(0),R_i(1)} = F_{U_ijT_i,\xi(0,V_i),\xi(1,V_i)}=F_{U_ij\xi(0,V_i),\xi(1,V_i)} = F_{U_ijR_i(0),R_i(1)}. \tag{22}
$$

As a result,

$$
F_{U_ijT_i=1,R_i=1} = \frac{p_{01}F_{U_ij(R_i(0),R_i(1))=(0,1)} + p_{11}F_{U_ij(R_i(0),R_i(1))=(1,1)}}{P(R_i=1|T_i=1)}, \tag{23}
$$

$$
F_{U_ijT_i=0,R_i=1} = \frac{p_{10}F_{U_ij(R_i(0),R_i(1))=(1,0)} + p_{11}F_{U_ij(R_i(0),R_i(1))=(1,1)}}{P(R_i=1|T_i=0)}. \tag{24}
$$

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

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

We can similarly show that $F_{U_ijT_i,R_i=0} = F_{U_i}$, 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 (23) and (24) both equal $F_{U_ij(R_i(0),R_i(1))=(1,1)}$. Similarly, it follows that $F_{U_ijT_i=1,R_i=0} = F_{U_ijT_i=0,R_i=0} = F_{U_ij(R_i(0),R_i(1))=(0,0)}$, which implies the result.

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

$$
\begin{aligned}
F_{U_ijT_i=1,R_i=1} &= \frac{p_{01}F_{U_ij(R_i(0),R_i(1))=(0,1)} + p_{11}F_{U_ij(R_i(0),R_i(1))=(1,1)}}{P(R_i=1|T_i=1)}\\
&= \frac{p_{10}F_{U_ij(R_i(0),R_i(1))=(1,0)} + p_{11}F_{U_ij(R_i(0),R_i(1))=(1,1)}}{P(R_i=1|T_i=0)} = F_{U_ijT_i=0,R_i=1}.
\end{aligned} \tag{25}
$$

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. □

C.1 Supplementary Example for Section 3.4.2

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

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

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

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

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.

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

$$
Y_{it} = U_{it}^1 + T_i(\beta_1 U_{it}^1 + \beta_2 U_{it}^2) \tag{26}
$$

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.

References

Abadie, Alberto, Matthew M. Chingos, and Martin R. West, "Endogenous Stratification in Randomized Experiments," *Review of Economics and Statistics*, 2018, *100* (4), 567–580.

Ahn, Hyungtaik and James L. Powell, "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 1993, *58* (1), 3–29.

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

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

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

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

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

Athey, Susan, Dean Eckles, and Guido W. Imbens, "Exact p-Values for Network Interference," *Journal of the American Statistical Association*, 2018, *113* (521), 230–240.

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

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

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

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

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

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

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

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh, "Inference Under Covariate-Adaptive Randomization," *Journal of the American Statistical Association*, 2018, *113* (524), 1784–1796.

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

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

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

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

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

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

Dufour, Jean-Marie, "Monte Carlo Tests with Nuisance Parameters: A General approach to Finite-Sample Inference and Nonstandard Asymptotics," *Journal of Econometrics*, 2006, *133* (2), 443 – 477.

_ **, Abdeljelil Farhat, Lucien Gardiol, and Lynda Khalaf**, "Simulation-based Finite Sample Normality Tests in Linear Regressions," *Econometrics Journal*, 1998, *1* (1), 154–173.

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

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

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

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

Heckman, James J, "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," in Sanford V. Berg, ed., *Annals of Economic and Social Measurement*, Vol. 5, National Bureau of Economic Research, 1976, pp. 475–492.

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

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

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

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

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

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

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

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

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

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

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

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

Kasy, Maximilian and Anja Sautmann, "Adaptive Treatment Assignment in Experiments For Policy Choice," *Econometrica*, 2020, *Forthcoming.*

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

Kline, Patrick and Andres Santos, "Sensitivity to Missing Data Assumptions: Theory and An Evaluation of The U.S. Wage Structure," *Quantitative Economics*, 2013, *4* (2), 231–267.

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

Lehmann, E. L. and Joseph P. Romano, *Testing Statistical Hypotheses*, third ed., New York: Springer, 2005.

Manski, Charles F., "Partial Identification with Missing Data: Concepts and Findings," *International Journal of Approximate Reasoning*, 2005, *39* (2), 151 – 165.

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

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

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

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

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

Muralidharan, Karthik, Mauricio Romero, and Kaspar Wüthrich, "Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments," Working Paper 26562, National Bureau of Economic Research December 2019.

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

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

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

Vazquez-Bare, Gonzalo, "Identification and Estimation of Spillover Effects in Randomized Experiments," 2020. Unpublished Manuscript.

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

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

Testing Attrition Bias in Field Experiments

Dalia Ghanem Sarojini Hirshleifer Karen Ortiz-Becerra

Online Appendix
February 18, 2021

SA1 Selection of Articles from the Field Experiment Literature

SA1.1 Selection of Articles for the Review

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 five highly regarded applied economics journals: *American Economic Review*, *American Economic Journal: Applied Economics*, *Econometrica*, *Economic Journal*, *Journal of Development Economics*, *Journal of Human Resources*, *Journal of Political Economy*, *Review of Economics and Statistics*, *Review of Economic Studies*, and *Quarterly Journal of Economics*.[54] By searching for *RCT*, *randomized controlled trial*, or *eld experiment* in each journal's website, we identified 202 articles that were published between 2009 and 2015.[55] From these 202 articles, we review those in which the main goal is to report the results of a field experiment and for which attrition is relevant given the experiment's study design. To be consistent with our panel approach in Section 3 in the paper, we only focus on those experiments with baseline data on at least one main outcome variable.

Table SA1 displays the distribution of the 93 articles that satisfied the selection criteria by journal and year of publication.[56] Of these 93 articles, 61% were published in the *Journal*

[54] We chose these four applied journals because they are important sources of published field experiments.

[55] Our initial search using these keywords yielded 235 articles but 33 of those papers were excluded since they were observational studies exploiting some sort of quasi-experimental variation.

[56] Some of the articles report results for more than one intervention. Thus, these 93 articles correspond to 96 field experiments.

of Development Economics, the *American Economic Journal: Applied Economics*, and the *Quarterly Journal of Economics*. Approximately 56% of our sample of articles were published in 2014 and 2015.

Table SA1: 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
JHR	0	0	0	1	1	1	2	5
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	13	11	29	24	93

Notes: The 93 articles that we include in our review correspond to 96 field experiments. The two articles that reported more than one field experiment are published in the AER(2015) and the QJE(2011), respectively.

We also exclude 64 articles that do not have available baseline data for any of the outcomes reported in the abstract. From these papers, 52% do not collect baseline outcome and 5% collect baseline data but have a baseline attrition above fifty percent. The remaining 28 papers that we exclude (43%) have the same baseline outcome for everyone by design. Some examples in this category include training interventions that target unemployed individuals and measure impacts on employment, and interventions that aim to estimate which of the multiple treatment arms has a higher impact on the take-up of a newly introduced product.

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.

SA1.2 Selection of Articles for the Empirical Applications

In order to conduct the empirical applications in Section 5, we identified 47 articles that had publicly available analysis files from the 93 articles in our review (see Section 2). To select the five articles that had the highest attrition rates from that group, we reviewed the data files for twelve articles. We excluded 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. 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.

SA2 Attrition Tests in the Field Experiment Literature

In order to classify the attrition tests that are conducted in the 93 articles that we review, we gathered information on the different econometric strategies that were carried out to test for attrition bias. In this section, we describe these empirical strategies and classify them into differential attrition rates test, selective attrition tests, and determinants of attrition test. We specify the null hypotheses of the selective attrition tests since this test is closely related to the tests that we propose. In contrast, we categorize the estimation strategies for the differential attrition rates test and the determinants of attrition test as broadly as possible and include any article that performs a regression under any of these two categories as performing the relevant test. Throughout this section, 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 $Z_{i0} = (X_{i0}^{\prime}, Y_{i0}^{\prime})^{\prime}$.

-For a vector w, w^j denotes the j^{th} element of w.

SA2.1 Differential Attrition Rates Test

The *differential attrition rates 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. $R_i = \gamma + T_i\beta + X_{i0}^{\prime}\theta + Y_{i0}^{\prime}\alpha + U_i$; may include strata fixed effects.

SA2.2 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 test that includes only respondents or attritors, and ii) a test that includes both respondents and attritors. We note that the selective attrition tests are usually conducted on both baseline outcomes and baseline covariates. Some authors conduct multiple tests for *individual* baseline variables while others test *all* baseline variables jointly (see Table SA3 for details). Thus, for each estimation strategy, we report the null hypotheses that are used in each case.

SA2.2.1 Tests that include only respondents or attritors

1. t-test of baseline characteristics by treatment group among respondents:

(a) *Multiple hypotheses for individual baseline variables:*

For each $j = 1, 2, \ldots, (l + k)$

$$H_0^j : E[Z_{i0}^j | T_i = 1, R_i = 1] = E[Z_{i0}^j | T_i = 0, R_i = 1].$$

(b) *Joint hypothesis for all baseline variables:*

$$H_0 : E[Z_{i0}^j | T_i = 1, R_i = 1] = E[Z_{i0}^j | T_i = 0, R_i = 1], \ \forall j = 1, \ldots, (l + k).$$

2. $T_i = \gamma + X_{i0}^{\prime}\theta + Y_{i0}^{\prime}\alpha + U_i$ if $R_i = 1$; may include strata fixed effects.

(a) *Joint hypothesis for all baseline variables:*

$$H_0 : \theta = \alpha = 0$$

3. Kolmogorov-Smirnov (KS) test of baseline characteristics by treatment group among respondents.

(a) *Multiple hypotheses for individual baseline variables:*

For each $j = 1, 2, \ldots, (l + k)$

$$H_0^j : F_{Z_{i0}^j | T_i, R_i = 1} = F_{Z_{i0}^j | R_i = 1}$$

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

(a) *Multiple hypotheses for individual baseline variables:*

For each $j = 1, 2, \ldots, (l + k)$

$$H_0^j : \beta^j = 0$$

(b) *Joint hypothesis for all baseline variables:*

$$H_0 : \beta^1 = \beta^2 = \cdots = \beta^{l+k} = 0$$

5. $Z_{i0}^j = \gamma + T_i\beta^j + U_i^j$ if $R_i = 0$, for $j = 1, 2, \ldots, (l+k)$; may include strata fixed effects.

   (a) *Multiple hypotheses for individual baseline variables:*

   For each $j = 1, 2, \ldots, (l+k)$

$$H_0^j : \beta^j = 0$$

SA2.2.2 Tests that include both respondents and attritors

1. $Z_{i0}^j = \gamma^j + T_i\beta^j + (1 - R_i)\lambda^j + T_i(1 - R_i)\phi^j + U_i^j$ for $j = 1, 2, \ldots, (l+k)$; may include strata fixed effects.

   (a) *Multiple hypotheses for individual baseline variables:*[57]

   For each $j = 1, 2, \ldots, (l+k)$

$$H_0^j : \beta^j = 0$$

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.

   (a) *Multiple hypotheses for individual baseline variables I:*

   For each $m = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, l$

$$H_0^{\theta,m} : \theta^m = 0 \quad , \quad H_0^{\alpha,j} : \alpha^j = 0 \quad , \quad H_0^{\lambda_{1,m}} : \lambda_1^m = 0 \quad , \quad H_0^{\lambda_{2,j}} : \lambda_2^j = 0$$

   (b) *Multiple hypotheses for individual baseline variables II:*

[57] Although this null hypothesis is testing for the equality of means for treatment and control respondents, we classify this strategy as one that includes both respondents and attritors given that the regression test is based on both samples.

For each $m = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, l$

$$H_0^{\lambda_1,m} : \lambda_1^m = 0 \quad , \quad H_0^{\lambda_2,j} : \lambda_2^j = 0$$

(c) *Joint hypothesis for all baseline variables I:*

$$H_0 : \beta = \theta = \alpha = \lambda_1 = \lambda_2 = 0$$

(d) *Joint hypothesis for all baseline variables II:*

$$H_0 : \lambda_1 = \lambda_2 = 0$$

3. t-test of the equality of the difference in baseline outcome between respondents and attritors across treatment groups.

   (a) *Multiple hypotheses for individual baseline outcomes:*

   For each $j = 1, 2, \ldots, l$

$$\begin{aligned} H_0^j : & E[Y_{i0}^j | T_i = 1, R_i = 1] - E[Y_{i0}^j | T_i = 1, R_i = 0] \\ & = E[Y_{i0}^j | T_i = 0, R_i = 1] - E[Y_{i0}^j | T_i = 0, R_i = 0] \end{aligned}$$

SA2.3 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. $Z_{i0}^j = \gamma^j + (1 - R_i)\lambda^j + U_i^j$, $j = 1, 2, \ldots, (l + k)$; may include strata fixed effects.
3. $R_i = \gamma + X_{i0}'\theta + Y_{i0}'\alpha + U_i$; may include strata fixed effects.

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

Table SA2: Overall Attrition Rate by Country's Income Group

Field Experiments in:	*N*	*Mean*	*SD*	*Min*	*Max*	*p25*	*p75*	Prop. of Experiments with Rate > 15%
High income countries	28	20.7	24.2	0	87	3	28	46%
Upper middle income countries	18	15.6	13.1	0	54	7	20	55%
Low and lower middle income countries	47	11.9	12.6	0	59	2	18	34%
All countries	93	15.3	17.2	0	87	3.3	21	42%

Notes: This table considers the highest overall attrition rate for each field experiment in our review and excludes one paper that does not report overall attrition rates. We classify countries by income group according to the official definition of the World Bank.

Table SA3: Number of Baseline Variables Included in The Selective Attrition Test

Category	**No. of Baseline Variables Included**					
	Mean	*SD*	*Min*	*Max*	*p25*	*p75*
All papers that conduct a selective attrition test	17.2	10.3	1	46	10	21
Papers that test on multiple baseline variables:						
Multiple hypotheses for individual variables (75%)	16.6	9.7	2	46	10	21
Joint hypothesis for all variables (25%)	20.3	11.3	4	44	13	23

Notes: Of the 46 experiments that conduct a selective attrition test, 44 test on multiple baseline variables. This table excludes one experiment that tests on multiple baseline variables but does not provide sufficient information for it to be categorized. Percentages are a proportion of the 44 experiments that test on multiple baseline variables.

SA3 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,

$$
\begin{aligned}
P(R_i = 0|T_i = 0) &= P_i((0,0,0)) + P_i((0,0,1)) + P_i((0,1,0)) + P_i((0,1,1)) \\
P(R_i = 0|T_i = 1) &= P_i((0,0,0)) + P_i((0,0,1)) + P_i((1,0,0)) + P_i((1,0,1)) \\
P(R_i = 0|T_i = 2) &= P_i((0,0,0)) + P_i((1,0,0)) + P_i((0,1,0)) + P_i((1,1,0))
\end{aligned} \tag{SA3.1}
$$

The equality of attrition 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,

$$
\begin{aligned}
P_i((0,1,0)) + P_i((0,1,1)) &= P_i((1,0,0)) + P_i((1,0,1)) \\
P_i((0,0,1)) + P_i((0,1,1)) &= P_i((1,0,0)) + P_i((1,1,0))
\end{aligned} \tag{SA3.2}
$$

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

SA4 Identification and Testing for the Multiple Treatment Case

In this section, we present the generalization of Propositions 1 and 2 (Section SA4.1) as well as the distributional test statistics (Section SA4.2) 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 $\mathcal{D} = \{0, 1, \ldots, |\mathcal{D}| - 1\}$, $|\mathcal{D}| < \infty$. $D_i \equiv (D_{i0}, D_{i1}) \in \{(0,0), (0,1), \ldots, (0, |\mathcal{D}| - 1)\}$. Let T_i denote the indicator for membership in the treatment group defined by D_i, i.e. $T_i \in \mathcal{T} = \{0, 1, \ldots, |\mathcal{D}| - 1\}$, where $T_i = D_{i1}$ and hence $|\mathcal{T}| = |\mathcal{D}|$ by construction.

SA4.1 Identification and Sharp Testable Restrictions

SA4.1.1 Completely Randomized Trials

Proposition 4. *Assume* $(U_{i0}, U_{i1}, V_i) \perp T_i$.

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

(i) (Identi cation) $Y_{i1}|T_i = \tau, R_i = 1 \stackrel{d}{=} Y_{i1}(\tau)|R_i = 1$ *for* $\tau \in \mathcal{T}$.

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

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

(i) (Identi cation) $Y_{i1}|T_i = \tau, R_i = 1 \stackrel{d}{=} Y_{i1}(\tau)$ *for* $\tau \in \mathcal{T}$, $r = 0, 1$.

(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 4) (a) Under the assumptions imposed it follows that $F_{U_{i0},U_{i1}jT_i,R_i} = F_{U_{i0},U_{i1}jR_i}$, which implies that for $d \in \mathcal{D}$, $F_{Y_{it}(d)jT_i,R_i} = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}jT_i,R_i}(u) = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}jR_i}(u) = F_{Y_{it}(d)jR_i}$. (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 $T_i = \tau^{\emptyset}$ and $R_i = r$, where $\tau \neq \tau^{\emptyset}$.

Following Hsu et al. (2019), we show that the testable restriction is sharp by showing that if $(Y_{i0}, Y_{i1}, T_i, R_i)$ satisfy $Y_{i0}|T_i = \tau, R_i = r \stackrel{d}{=} Y_{i0}|T_i = \tau^{\emptyset}, R_i = r$ for $r = 0, 1$, $\tau, \tau^{\emptyset} \in \mathcal{T}$, $\tau \neq \tau^{\emptyset}$, then there exists (U_{i0}, U_{i1}) such that $Y_{it}(d) = \mu_t(d, U_{it})$ for some $\mu_t(d, .)$ 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 μ_t, we can let $U_{it}^{\emptyset} = \mathbf{Y}_{it}(.) = (Y_{it}(0), Y_{it}(1), \ldots, Y_{it}(|\mathcal{D}| - 1))$ and $\mu_t(d, U_{it}) = \sum_{j=0}^{D} {}^{1} 1\{j = d\} Y_{it}(j)$ 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 $U_i = (U_{i0}^{\emptyset}, U_{i1}^{\emptyset})$ that satisfies

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

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

$$\begin{aligned} F_{U_i | T_i = \tau, R_i = 1} &= F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1}, \mathbf{Y}_{i1}(.) | Y_{i0}, T_i = \tau, R_i = 1} F_{Y_{i0} | T_i = \tau, R_i = 1} \\ &= F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1}, \{Y_{i1}(d)\}_{d=0}^{\tau-1}, Y_{i1}, \{Y_{i1}(d)\}_{d=\tau+1}^{|\mathcal{D}|-1} | Y_{i0}, T_i = \tau, R_i = 1} F_{Y_{i0} | T_i = \tau, R_i = 1}. \end{aligned} \tag{SA4.1}$$

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 $\forall \tau, \tau^{\emptyset} \in \mathcal{T}$, $\tau \neq \tau^{\emptyset}$,

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

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,

$$F_{U_i | T_i = \tau, R_i = 0} = F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1}, \mathbf{Y}_{it}(.) | Y_{i0}, T_i = \tau, R_i = 0} F_{Y_{i0} | T_i = \tau, R_i = 0} \tag{SA4.2}$$

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 $F_{\{Y_{i0}(d)\}_{d=1}^{|\mathcal{D}|-1}, \mathbf{Y}_{it}(.) | Y_{i0}, R_i = 0}$. 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 $d \in \mathcal{D}$, $F_{Y_{it}(d)|T_i,R_i}(.) = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}|T_i,R_i}(u) = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}}(u) = F_{Y_{it}(d)}$. (i) follows by letting $d = \tau$ 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 $\mu_t(d,.)$ 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 $U_{it}^{\emptyset} = \mathbf{Y}_{it}(.) = (Y_{it}(0), Y_{it}(1), \ldots, Y_{it}(|\mathcal{D}| - 1))$ 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 $\tau \in \mathcal{T}$

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

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

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

SA4.1.2 Stratified Randomized Trials

Proposition 5. *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) $Y_{i0}|T_i = \tau, S_i = s, R_i = r \stackrel{d}{=} Y_{i0}|T_i = \tau^{\emptyset}, S_i = s, R_i = r$, $\forall \tau, \tau^{\emptyset} \in \mathcal{T}, \tau \neq \tau, s \in \mathcal{S}, r = 0, 1$.

(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 5) The proof for this proposition follows in a straightforward manner from the proof for Proposition 4 by conditioning all statements on S_i. □

SA4.2 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.

SA4.2.1 Completely Randomized Trials

The null hypothesis implied by Proposition 4(a.ii) is given by the following,

$$H_0^{1,\mathcal{T}} : F_{Y_{i0}|T_i=\tau,R_i=r} = F_{Y_{i0}|T_i=\tau',R_i=r} \text{ for } \tau, \tau' \in \mathcal{T},\ \tau \neq \tau',\ r = 0, 1. \tag{SA4.3}$$

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 $r = 0, 1$,

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

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 (τ_j, r_j) denote the j^{th} element of $\mathcal{T} \times \mathcal{R}$, then

the null hypothesis implied by Proposition 4(b.ii) is given by the following:

$$H_0^{2,\mathcal{T}} \; : \; F_{Y_{i0}|T_i=\tau_j,R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1},R_i=r_{j+1}} \text{ for } j = 1,\ldots,|\mathcal{T}\times\mathcal{R}|-1. \tag{SA4.4}$$

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

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

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 $H_0^{2,\mathcal{T}}$.

SA4.2.2 Stratified Randomized Trials

The null hypothesis implied by Proposition 5(a.ii) is given by the following,

$$H_0^{1,\mathcal{S},\mathcal{T}} : F_{Y_{i0}|T_i=\tau,S_i=s,R_i=r} = F_{Y_{i0}|T_i=\tau',S_i=s,R_i=r} \text{ for } \tau,\tau' \in \mathcal{T},\ \tau\neq\tau',\ s\in\mathcal{S},\ r=0,1. \tag{SA4.5}$$

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 $r=0,1$,

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

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 (τ_j,r_j) denote the j^{th} element of $\mathcal{T}\times\mathcal{R}$, then the null hypothesis implied by Proposition 5(b.ii) is given by the following:

$$H_0^{2,\mathcal{S},\mathcal{T}} \; : \; F_{Y_{i0}|T_i=\tau_j,S_i=s,R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1},S_i=s,R_i=r_{j+1}} \text{ for } j = 1,\ldots,|\mathcal{T}\times\mathcal{R}|-1,\ s\in\mathcal{S}. \tag{SA4.6}$$

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

$$T_{n,m}^{2,S,T} = \max_{s2S} \max_{j=1,\ldots,jT \quad Rj \quad 1} \left\| \sqrt{n} \left(F_{n,Y_{i0}jT_i=\tau_j,S_i=s,R_i=r_j} - F_{n,Y_{i0}jT_i=\tau_{j+1},S_i=s,R_i=r_{j+1}} \right) \right\|,$$

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

SA5 Extended Simulations for the Distributional Tests

SA5.1 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 4 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

$$\begin{aligned}
H_0^{1,1} : &\quad Y_{i0}|T_i = 1, R_i = 0 \overset{d}{=} Y_{i0}|T_i = 0, R_i = 0, && (CA - TA) \\
H_0^{1,2} : &\quad Y_{i0}|T_i = 1, R_i = 1 \overset{d}{=} Y_{i0}|T_i = 0, R_i = 1, && (CR - TR) \\
H_0^{1} \quad : &\quad H_0^{1,1} \ \& \ H_0^{1,2}. && (Joint)
\end{aligned} \tag{SA5.1}$$

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

$$\begin{aligned}
KS_{n,r}^{1} &= \max_{i:R_i=r} \left| \sqrt{n} \left(F_{n,Y_{i0}}(y_{i0}|T_i = 1, R_i = r) - F_{n,Y_{i0}}(y_{i0}|T_i = 0, R_i = r) \right) \right|. \\
CM_{n,r}^{1} &= \frac{\sum_{i:R_i=r} \left(\sqrt{n} (F_{n,Y_{i0}}(y_{i0}|T_i = 1, R_i = r) - F_{n,Y_{i0}}(y_{i0}|T_i = 0, R_i = r)) \right)^2}{\sum_{i=1}^{n} 1\{R_i = r\}}
\end{aligned} \tag{SA5.2}$$

For the joint hypothesis H_0^1, which is the sharp testable restriction in Proposition 1(b.ii) in

the paper, we consider either $KS^1_{n,m} = \max\{KS^1_{n,0}, KS^1_{n,1}\}$ or $KS^1_{n,p} = p_{n,0}KS^1_{n,0} + p_{n,1}KS^1_{n,1}$, where $p_{n,r} = \sum_{i=1}^n 1\{R_i = r\}/n$ for $r = 0, 1$. $CM^1_{n,m}$ and $CM^1_{n,p}$ are similarly defined.

Table SA4 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, $KS^1_{n,0}$ and $KS^1_{n,1}$, 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^1_0 ($B = 199$). 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 $KS^1_{n,m}$ and $CM^1_{n,m}$ ($Joint$ (m)) exhibit better finite-sample power properties than $KS^1_{n,p}$ and $CM^1_{n,p}$ ($Joint$ (p)). We also note that the randomization procedure yields rejection probabilities for the two-sample KS statistics, $KS^1_{n,0}$ and $KS^1_{n,1}$, that are very similar to those obtained from the asymptotic critical values. 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,

$$\begin{aligned}
H_0^{2,1} &: \quad Y_{i0}|T_i = 0, R_i = 0 \overset{d}{=} Y_{i0}|T_i = 0, R_i = 1, && (CA - CR) \\
H_0^{2,2} &: \quad Y_{i0}|T_i = 0, R_i = 1 \overset{d}{=} Y_{i0}|T_i = 1, R_i = 0, && (CR - TA) \\
H_0^{2,3} &: \quad Y_{i0}|T_i = 1, R_i = 0 \overset{d}{=} Y_{i0}|T_i = 1, R_i = 1, && (TA - TR) \\
H_0^{2} &: \quad H_0^{2,1} \ \& \ H_0^{2,2} \ \& \ H_0^{2,3}. && (Joint)
\end{aligned} \tag{SA5.3}$$

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

$$\begin{aligned} KS_{n,j}^2 &= \max_{i:(T_i,R_i)2f(\tau_j,r_j),(\tau_{j+1},r_{j+1})g} \left|\sqrt{n}\left(F_{n,Y_{i0}jT_i=\tau_{j-1},R_i=r_{j-1}} - F_{n,Y_{i0}jT_i=\tau_j,R_i=r_j}\right)\right|, \\ CM_{n,j}^2 &= \frac{\sum_{i:(T_i,R_i)2f(\tau_j,r_j),(\tau_{j+1},r_{j+1})} \left(\sqrt{n}\left(F_{n,Y_{i0}jT_i=\tau_{j-1},R_i=r_{j-1}} - F_{n,Y_{i0}jT_i=\tau_j,R_i=r_j}\right)\right)^2}{\sum_{i=1}^n 1\left\{(T_i,R_i) \in \{(\tau_j,r_j),(\tau_{j+1},r_{j+1})\}\right\}}, \end{aligned} \tag{SA5.4}$$

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 SA5, 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.

SA5.2 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 $P(R_i = 0|T_i = 0) = 0.15$.[58] For the joint hypotheses, we

[58] 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.

report the simulation results for the KS statistic that takes the maximum over the individual statistics.

Panel A in Figure SA1 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 4 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 $0.9 \times P(R_i = 0|T_i = 0)$, where $p_{01} = 0$ corresponds to Design II and $p_{01} > 0$ to variants of Design I. *Design III to I* depicts the case in which we vary the correlation parameter between the unobservables in the outcome equation and the unobservables in the response equation, ρ, from zero to one. Hence, $\rho = 0$ corresponds to Design III while $\rho > 0$ corresponds to different versions of Design I. Finally, the results under *Design II to IV* are obtained by fixing $p_{01} = p_{10}$ and varying them from zero to $0.9 \times P(R_i = 0|Ti = 0)$. Design II corresponds to the case in which $p_{01} = p_{10} = 0$ and $p_{01} = p_{10} > 0$ corresponds to different versions of Design IV.

Overall, the simulation results illustrate that the *joint* tests that we propose in Section A 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 SA1 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 SA4: Simulation Results on the KS & CM Randomization Test of IV-R

Design	Att. Rate		KS (*Asym.*)		KS (R)				CM(R)			
	C	T	CR-TR	CA-TA	CR-TR	CA-TA	Joint (m)	Joint (p)	CR-TR	CA-TA	Joint (m)	Joint (p)
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$												
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 + $(U_{i0}, U_{i1}) \perp (R_i(0), R_i(1))$ (*Example 1*)*												
III	0.050	0.025	0.049	0.051	0.054	0.052	0.052	0.056	0.048	0.051	0.051	0.049
	0.100	0.050	0.047	0.042	0.050	0.046	0.046	0.047	0.053	0.047	0.047	0.043
	0.150	0.100	0.047	0.038	0.052	0.045	0.045	0.047	0.049	0.049	0.049	0.048
	0.200	0.150	0.054	0.031	0.053	0.036	0.036	0.047	0.055	0.036	0.036	0.044
	0.300	0.200	0.050	0.043	0.050	0.043	0.043	0.050	0.051	0.042	0.042	0.050
Equal Attrition Rates + Violation of Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$ (*Example 2*)												
IV	0.050	0.050	0.059	0.332	0.065	0.329	0.329	0.093	0.067	0.375	0.375	0.302
	0.100	0.100	0.102	0.569	0.102	0.577	0.577	0.230	0.116	0.663	0.663	0.593
	0.150	0.150	0.178	0.740	0.190	0.758	0.758	0.465	0.211	0.816	0.816	0.805
	0.200	0.200	0.313	0.854	0.319	0.859	0.859	0.709	0.368	0.917	0.916	0.910
	0.300	0.300	0.683	0.970	0.680	0.972	0.974	0.974	0.760	0.985	0.991	0.996

Notes: The above table presents the rejection probabilities of the KS and CM tests for the simple and joint null hypotheses in (SA5.1). 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 $KS^1_{n,m}$ $(CM^1_{n,m})$ and $KS^1_{n,p}$ $(CM^1_{n,p})$, respectively. Additional details of the design are provided in Table 4 in the paper.
† $(^{*})$ indicates IV-R only (IV-P).

Table SA5: Simulation Results on the KS & CM Randomization Test of IV-P

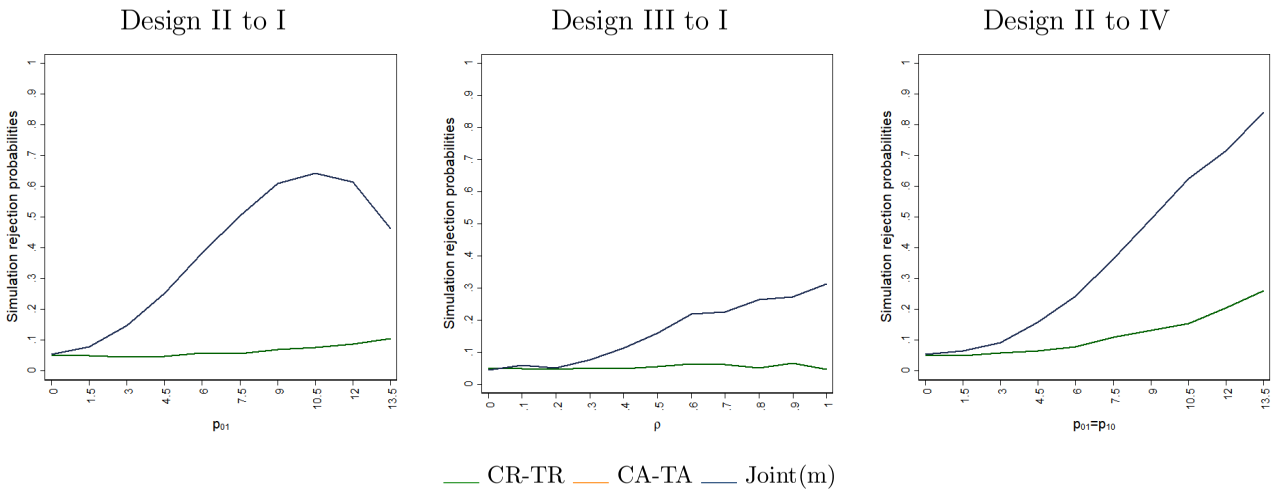
Design	Att. Rate		KS (*Asym.*)			KS (R)				CM(R)			
	C	T	CA-CR	CR-TA	TA-TR	CA-CR	CR-TA	TA-TR	*Joint* (m)	CA-CR	CR-TA	TA-TR	*Joint* (m)
Differential Attrition Rates + Monotonicity + $(U_{i0}, U_{i1}) \not\perp (R_i(0), R_i(1))$													
I	0.050	0.025	0.051	0.451	0.456	0.064	0.482	0.485	0.476	0.053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

Notes: The above table presents the rejection probabilities of the KS and CM tests for the simple and joint null hypotheses in (SA5.3). 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 $KS^2_{n,m}$ ($CM^2_{n,m}$). Additional details of the design are provided in Table 4 in the paper.

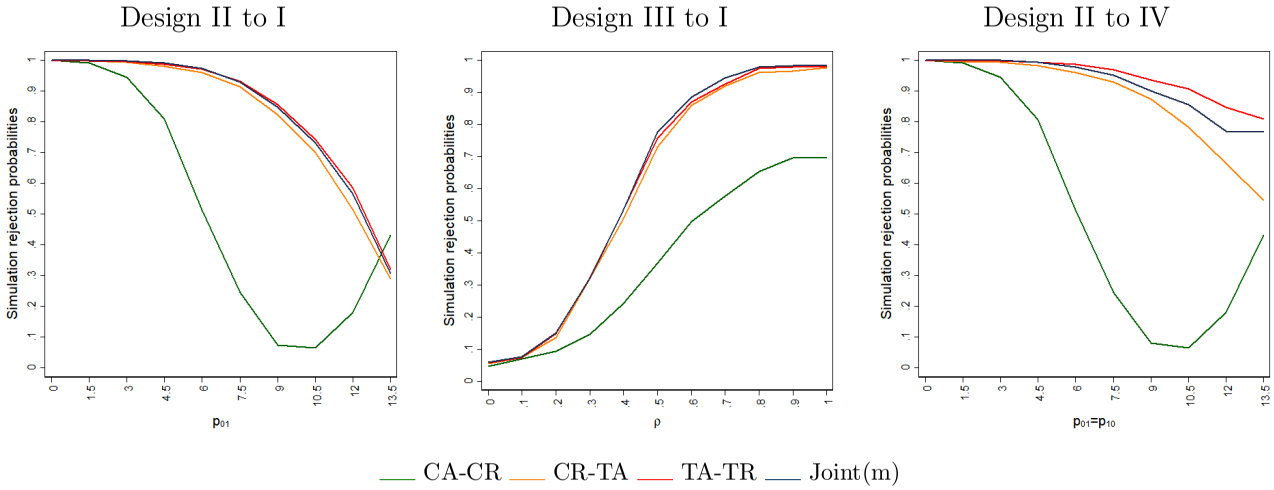
† (*) indicates IV-R only (IV-P).

Figure SA1: Additional Simulation Analysis for the KS Statistic of Internal Validity

Panel A. Internal Validity for Respondents



Panel B. Internal Validity for the Study Population



SA6 List of Papers Included in the Review of Field Experiments

Abeberese, Ama Baafra, Todd J. Kumler, and Leigh L. Linden. 2014. "Improving Reading Skills by Encouraging Children to Read in School: A Randomized Evaluation of the Sa Aklat Sisikat Reading Program in the Philippines." *Journal of Human Resources*, 49 (3): 611–33.

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. Avitabile, Ciro. 2012. "Does Information Improve the Health Behavior of Adults Targeted by a Conditional Transfer Program?" *Journal of Human Resources*, 47 (3): 785–825.

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.

Berry, James. 2015. "Child Control in Education Decisions: An Evaluation of Targeted Incentives to Learn in India." *Journal of Human Resources* 50 (4): 1051–80.

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 Experi-

ment 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.

Just, David R., and Joseph Price. 2013. "Using Incentives to Encourage Healthy Eating

in Children." *Journal of Human Resources* 48 (4): 855–72.

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.

Reichert, Arndt R. 2015. "Obesity, Weight Loss, and Employment Prospects: Evidence from a Randomized Trial." *Journal of Human Resources* 50 (3): 759–810.

Royer, H., Stehr, M., & Sydnor, J. (2015). Incentives, Commitments, and Habit Forma-

tion 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.