Testing Attrition Bias in Field Experiments*

Dalia Ghanem   Sarojini Hirshleifer   Karen Ortiz-Becerra

Online Appendix

*Dalia Ghanem is an associate professor of agricultural and resource economics at University of California, Davis (dghanem@ucdavis.edu). Sarojini Hirshleifer is an assistant professor of economics at University of California, Riverside and CEGA (sarojini.hirshleifer@ucr.edu). Karen Ortiz-Becerra is an assistant professor of economics at the University of San Diego (kortizbecerra@sandiego.edu).

SA1 Proofs

Proof. (Proposition 1)

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

Following Hsu, Liu and Shi (2019), we show that the testable restriction is sharp by showing that if (Y_{i0},Y_{i1},T_i,R_i) satisfy $Y_{i0}|T_i=0,R_i=r \stackrel{d}{=} Y_{i0}|T_i=1,R_i=r$ for $r=0,1$, then there exists (U_{i0},U_{i1}) such that $Y_{it}(d)=\mu_t(d,U_{it})$ for some $\mu_t(d,.)$ for $d=0,1$ and $t=0,1$, and $(U_{i0},U_{i1}) \perp T_i|R_i$ that generate the observed distributions. By the arbitrariness of U_{it} and 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

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

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,

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

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 $Y_{i0}|T_i,R_i \overset{d}{=} Y_{i0}(0)$, then there exists (U_{i0},U_{i1}) such that $Y_{it}(d) = \mu_t(d,U_{it})$ for some $\mu_t(d,.)$ for $d = 0,1$ and $t = 0,1$, and $(U_{i0},U_{i1}) \perp (T_i,R_i)$. Similar to (a.ii), we let $U_{it} = (Y_{it}(0),Y_{it}(1))'$ and $\mu_t(d,U_{it}) = dY_{it}(1) + (1-d)Y_{it}(0)$. Then $Y_{i0} = Y_{i0}(0)$ by similar arguments as in the above. Furthermore, $F_{Y_{i0}|T_i,R_i} = F_{Y_{i0}}$ by construction and it follows immediately that

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

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

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

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

$$F_{U_i|T_i,R_i(0),R_i(1)} = F_{U_i|T_i,\xi(0,V_i),\xi(1,V_i)} = F_{U_i|\xi(0,V_i),\xi(1,V_i)} = F_{U_i|R_i(0),R_i(1)}. \tag{SA1.1}$$

As a result,

$$F_{U_i|T_i=1,R_i=1} = \frac{p_{01}F_{U_i|(R_i(0),R_i(1))=(0,1)} + p_{11}F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i=1|T_i=1)}, \tag{SA1.2}$$

$$F_{U_i|T_i=0,R_i=1} = \frac{p_{10}F_{U_i|(R_i(0),R_i(1))=(1,0)} + p_{11}F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i=1|T_i=0)}. \tag{SA1.3}$$

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

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

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

Alternatively, if we assume (ii), $R_i(0) \leq R_i(1)$ implies $p_{10} = 0$. As a result, $P(R_i = 0|T_i = 1) = P(R_i = 0|T_i = 0)$ iff $p_{01} = 0$. It follows that the terms in (SA1.2) and (SA1.3) both equal $F_{U_i|(R_i(0),R_i(1))=(1,1)}$. Similarly, it follows that $F_{U_i|T_i=1,R_i=0} = F_{U_i|T_i=0,R_i=0} = F_{U_i|(R_i(0),R_i(1))=(0,0)}$, which implies the result.

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

$$\begin{aligned} F_{U_i|T_i=1,R_i=1} &= \frac{p_{01}F_{U_i|(R_i(0),R_i(1))=(0,1)} + p_{11}F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i=1|T_i=1)} \\ &= \frac{p_{10}F_{U_i|(R_i(0),R_i(1))=(1,0)} + p_{11}F_{U_i|(R_i(0),R_i(1))=(1,1)}}{P(R_i=1|T_i=0)} = F_{U_i|T_i=0,R_i=1}. \end{aligned} \tag{SA1.4}$$

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. □

1 Supplementary Example for Section IV.A.

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

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

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

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

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

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

$$Y_{it} = U_{it}^1 + T_i(\beta_1 U_{it}^1 + \beta_2 U_{it}^2) \tag{SA1.5}$$

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.

SA2 Randomization Tests of Internal Validity

We present randomization procedures to test the IVal-R and IVal-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.[69] 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 in the paper. The randomization procedures we propose, however, can be applied to test joint distributional hypotheses that include covariates as in Section IV.B..

We first outline a general randomization procedure that we adapt to the different settings we consider.[70] 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 $g \in \mathscr{G}_0$ (Lehmann and Romano, 2005, Chapter 15.2). Specifically, the transformations $g \in \mathscr{G}_0$ satisfy $\mathbf{Z} \stackrel{d}{=} g(\mathbf{Z})$ under H_0 only.

[69] 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.

[70] See Lehmann and Romano (2005); Canay, Romano and Shaikh (2017) for a more detailed review.

Procedure 1. *(Randomization)*

1. *For g_b, which is i.i.d. Uniform($\mathscr{G}_0$), compute $\hat{T}_n(g_b) = T(g_b(\mathbf{Z}))$,*

2. *Repeat Step 1 for $b = 1, \dots, B$ times,*

3. *Compute the p-value, $\hat{p}_{n,B} = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} 1\{\hat{T}_n(g_b) \geq T_n\}\right)$.*

A test that rejects when $\hat{p}_{n,B} \leq \alpha$ is level α for any B (Lehmann and Romano, 2005, Chapter 15.2). In our application, the invariant transformations in $\mathscr{G}_0$ 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.

1 Completely Randomized Trials

The testable restriction of the IVal-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

$$\text{(SA2.1)} \qquad H_0^1 : F_{Y_{i0}|T_i=0,R_i=r} = F_{Y_{i0}|T_i=1,R_i=r} \text{ for } r = 0,1.$$

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

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

where for a random variable X_i, F_{n,X_i} denotes the empirical cdf, i.e. the sample analogue of F_{X_i}, and $\|.\|$ denotes some non-random or random norm. Different choices of the norm give rise to different statistics. For instance, the KS and CM statistics are the most widely known and used. The former is obtained by using the L^∞ norm over the sample points, i.e. $\|f\|_{n,\infty} = \max_i |f(y_i)|$, whereas the latter is obtained by using an L^2 norm, i.e. $\|f\|_{n,2} = \sum_{i=1}^{n} f(y_i)^2/n$. In order to test the *joint* hypothesis in (SA2.1), 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),[71]

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

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

$$KS^1_{n,m} = \max\{KS^1_{n,0}, KS^1_{n,1}\}, \text{where for } r = 0,1$$
$$\text{(SA2.2)} \qquad KS^1_{n,r} = \max_{i:R_i=r} \left|\sqrt{n}\left(F_{n,Y_{i0}}(y_{i0}|T_i = 1, R_i = r) - F_{n,Y_{i0}}(y_{i0}|T_i = 0, R_i = r)\right)\right|.$$

Let $\mathscr{G}^1_0$ denote the set of all permutations of individual observations within respondent and attritor subgroups, for $g \in \mathscr{G}^1_0$, $g(\mathbf{Z}) = \{(Y_{i0}, T_{g(i)}, R_{g(i)}) : R_{g(i)} = R_i, 1 \leq i \leq n\}$. Under H^1_0 and the cross-sectional i.i.d. assumption, $\mathbf{Z} \stackrel{d}{=} g(\mathbf{Z})$ for $g \in \mathscr{G}^1_0$. 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 $\mathscr{G}^1_0$.

We now consider testing the restriction of the IVal-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 $(\tau, r) \in \mathscr{T} \times \mathscr{R} = \{(0,0),(0,1),(1,0),(1,1)\}$ and (τ_j, r_j) denote the j^{th} element of $\mathscr{T} \times \mathscr{R}$. Then, the joint hypothesis is given wlog by

$$\text{(SA2.3)} \qquad H^2_0 : F_{Y_{i0}|T_i=\tau_j, R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1}, R_i=r_{j+1}} \text{ for } j = 1, \ldots, |\mathscr{T} \times \mathscr{R}| - 1.$$

[71] 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 SA7.3.

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

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

for some fixed or data-dependent non-negative weights w_j for $j = 1,\ldots,|\mathscr{T}\times\mathscr{R}|-1$. In the simulation and empirical sections, we use the following KS statistic to test H^2_0

(SA2.4) $$KS^2_n = \max_{j=1,2,3} KS^2_{n,j}, \text{ where}$$
$$KS^2_{n,j} = \max_i \left|\sqrt{n}\left(F_{n,Y_{i0}}(y_{i0}|T_i=\tau_j,R_i=r_j) - F_{n,Y_{i0}}(y_{i0}|T_i=\tau_{j+1},R_i=r_{j+1})\right)\right|.$$

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

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

2 Stratified Randomized Trials

As pointed out in Section III.B.3. of the paper, 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 SA2.1.

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

(SA2.5) $$H^{1,\mathscr{S}}_0 : F_{Y_{i0}|T_i=0,S_i=s,R_i=r} = F_{Y_{i0}|T_i=1,S_i=s,R_i=r} \text{ for } r = 0,1,\ s \in \mathscr{S}.$$

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

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

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

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

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

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

$$
\text{(SA2.6)} \qquad H_0^{2,\mathscr{S}} : F_{Y_{i0}|T_i=\tau_j,S_i=s,R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1},S_i=s,R_i=r_{j+1}} \text{ for } j = 1, \ldots, |\mathscr{T} \times \mathscr{R}| - 1,\ s \in \mathscr{S}.
$$

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

$s \in \mathscr{S}$,

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

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

$$T^{2,\mathscr{S}}_{n,m} = \max_{s\in\mathscr{S}} T^{2,\mathscr{S}}_{n,m,s},$$
$$T^{2,\mathscr{S}}_{n,p} = \sum_{s\in\mathscr{S}} w_s T^{2,\mathscr{S}}_{n,p,s},$$

given fixed or random non-negative weights w_s for $s \in \mathscr{S}$.

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

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 IVal-R assumption, the stratum-specific hypothesis is $H^{1,s}_0 : F_{Y_{i0}|T_i=0,S_i=s,R_i=r} = F_{Y_{i0}|T_i=1,S_i=s,R_i=r}$ for $r = 0,1$. Hence, for each $s \in \mathscr{S}$, one can use $\mathscr{G}^{1,\mathscr{S}}_0$ in the above procedure to obtain p-values for $T^{1,\mathscr{S}}_{n,m,s}$ and $T^{1,\mathscr{S}}_{n,p,s}$, 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 IVal-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.[72] 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 IVal-R and IVal-P assumptions using regressions for completely, cluster, and stratified randomized experiments in Appendix A in the paper.

SA3 Selection of Articles from the Field Experiment Literature

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*.[73] By searching for *RCT*, *randomized controlled trial*, or *field experiment* in each journal's website, we identified 160 articles that estimate the impacts of a field experiment intervention and were published between 2009 and 2015.[74]

Of these 160 experiments, we exclude five articles with a study design for which attrition is irrelevant due to the use of repeated cross-sections or the fact that attrition is the only outcome reported in the abstract. Further, since the testable restrictions proposed in Section III are conditions on the baseline outcome, we also excluded 62 articles that did not have available baseline data for any of the abstract outcomes. Half of these papers did not collect baseline outcomes (29) or had a

[72] 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.

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

[74] Our initial search using these keywords yielded a total of 235 articles, but 75 of them were neither field experiments nor studies that report the impacts of an intervention on a specific outcome for the first time. Of these 75 papers, 33 were observational studies exploiting quasi-experimental variation, and 27 were lab experiments or lab in the field (which usually take place over a very short period of time). The remaining 15 articles had a primary goal different from reporting an intervention's impact. In particular, some papers used existing field experiments to calibrate structural models or illustrate the application of a new econometric technique, and others used the random allocation of survey formats to test for the best approach to elicit information on variables such as consumption and poverty.

response rate at baseline below fifty percent (4). The other experiments targeted a population for which the baseline outcome takes the same value for everyone by design (29).[75]

Thus, we review 93 papers with a study design for which attrition is relevant and baseline data on at least one main outcome variable reported in the abstract.[76] Of these articles, 61% were published in the *Journal of Development Economics*, the *American Economic Journal: Applied Economics*, and the *Quarterly Journal of Economics* (see Table SA2).

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.

2 Selection of Articles for the Empirical Applications

In order to conduct the empirical applications in Section V, we identified 47 articles that had publicly available analysis files from the 93 articles in our review (see Section II). To select the four articles included in the empirical applications, we reviewed the data files of the twelve articles with the highest reported survey attrition rates. 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 eight 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, one did not provide sufficient information to identify the attritors, and one had a unique outcome of interest that was nearly degenerate at baseline. In two cases, an exceptionally high number of missing values at baseline was the limiting factor since the attrition rate at follow-up

[75]Some examples in this last category include training interventions that target unemployed individuals and measure impacts on employment, as well as studies that estimate the effect of an intervention on the take-up of a newly introduced product.

[76]These 93 articles correspond to 96 field experiments since some papers report results for more than one intervention.

conditional on baseline response was lower than the attrition rate reported in the paper.

SA4 Attrition Tests in the Field Experiment Literature

In this section, we describe the different empirical strategies used to test for attrition bias in the articles we review and classify them into differential attrition rate tests, selective attrition tests, and determinants of attrition tests. We classify the strategies for the differential attrition rate 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. For the selective attrition tests, we specify the null hypotheses since they are closely related to the tests that we propose. 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 $Z_{i0} = (X'_{i0}, Y'_{i0})'$.

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

1 Differential Attrition Rate Test

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

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

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

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

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 subtypes 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 SA4 for details). Thus, for each estimation strategy, we report the null hypotheses that are used in each case.

A 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. $T_i = \gamma + X_{i0}'\theta + Y_{i0}'\alpha + U_i$ 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 re-

spondents.

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

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

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

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

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

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

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

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

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

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

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

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

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

B 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 $j = 1, 2, \dots, (l+k)$; may include strata fixed effects.

(a) *Multiple hypotheses for individual baseline variables:*[77]

[77] Although this null hypothesis is testing for the equality of means for treatment and control respondents, we classify

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:*

   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:*

this strategy as one that includes both respondents and attritors given that the regression test is based on both samples.

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

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

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.

SA5 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,

$$P(R_i = 0|T_i = 0) = P_i((0,0,0)) + P_i((0,0,1)) + P_i((0,1,0)) + P_i((0,1,1))$$
$$P(R_i = 0|T_i = 1) = P_i((0,0,0)) + P_i((0,0,1)) + P_i((1,0,0)) + P_i((1,0,1))$$
$$P(R_i = 0|T_i = 2) = P_i((0,0,0)) + P_i((1,0,0)) + P_i((0,1,0)) + P_i((1,1,0)) \tag{SA5.1}$$

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,

$$P_i((0,1,0)) + P_i((0,1,1)) = P_i((1,0,0)) + P_i((1,0,1))$$

$$P_i((0,0,1)) + P_i((0,1,1)) = P_i((1,0,0)) + P_i((1,1,0)) \tag{SA5.2}$$

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

SA6 Identification and Testing for the Multiple Treatment Case

In this section, we present the generalization of Propositions 1 and 2 (Section SA6.1) as well as the distributional test statistics (Section SA6.2) 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, $D_{i1} \in \mathscr{D}$, where wlog $\mathscr{D} = \{0, 1, \ldots, |\mathscr{D}| - 1\}$, $|\mathscr{D}| < \infty$. $D_i \equiv (D_{i0}, D_{i1}) \in \{(0,0), (0,1), \ldots, (0, |\mathscr{D}| - 1)\}$. Let T_i denote the indicator for membership in the treatment group defined by D_i, i.e. $T_i \in \mathscr{T} = \{0, 1, \ldots, |\mathscr{D}| - 1\}$, where $T_i = D_{i1}$ and hence $|\mathscr{T}| = |\mathscr{D}|$ by construction.

1 Identification and Sharp Testable Restrictions

A 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) (Identification) $Y_{i1} | T_i = \tau, R_i = 1 \stackrel{d}{=} Y_{i1}(\tau) | R_i = 1$ *for* $\tau \in \mathscr{T}$.

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

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

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

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

Proof. (Proposition 4) (a) Under the assumptions imposed it follows that $F_{U_{i0},U_{i1}|T_i,R_i} = F_{U_{i0},U_{i1}|R_i}$, which implies that for $d \in \mathscr{D}$, $F_{Y_{it}(d)|T_i,R_i} = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}|T_i,R_i}(u) = \int 1\{\mu_t(d,u) \leq .\} dF_{U_{it}|R_i}(u) = F_{Y_{it}(d)|R_i}$. (i) follows by letting $t = 1$ and $d = \tau$, while conditioning the left-hand side of the last equation on $T_i = \tau$ and $R_i = 1$ and the right-hand side on $R_i = 1$. The testable implication in (ii) follows by letting $t = d = 0$ and conditioning the left-hand side on $T_i = \tau$ and $R_i = r$ and the right-hand side on $T_i = \tau'$ and $R_i = r$, where $\tau \neq \tau'$.

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

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

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

$$\begin{aligned} F_{U_i|T_i=\tau,R_i=1} &= F_{\{Y_{i0}(d)\}_{d=1}^{|\mathscr{D}|-1},\mathbf{Y}_{i1}(.)|Y_{i0},T_i=\tau,R_i=1} F_{Y_{i0}|T_i=\tau,R_i=1} \\ &= F_{\{Y_{i0}(d)\}_{d=1}^{|\mathscr{D}|-1},\{Y_{i1}(d)\}_{d=0}^{\tau-1},Y_{i1},\{Y_{i1}(d)\}_{d=\tau+1}^{|\mathscr{D}|-1}|Y_{i0},T_i=\tau,R_i=1} F_{Y_{i0}|T_i=\tau,R_i=1}. \end{aligned} \tag{SA6.1}$$

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

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

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,

$$\text{(SA6.2)} \qquad F_{U_i|T_i=\tau,R_i=0} = F_{\{Y_{i0}(d)\}_{d=1}^{|\mathscr{D}|-1},\mathbf{Y}_{it}(.)|Y_{i0},T_i=\tau,R_i=0}F_{Y_{i0}|T_i=\tau,R_i=0}$$

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 $\tau \in \mathscr{T}$ using the same $F_{\{Y_{i0}(d)\}_{d=1}^{|\mathscr{D}|-1},\mathbf{Y}_{it}(.)|Y_{i0},R_i=0}$. 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 $d \in \mathscr{D}$, $F_{Y_{it}(d)|T_i,R_i}(.) = \int 1\{\mu_t(d,u) \leq .\}dF_{U_{it}|T_i,R_i}(u) = \int 1\{\mu_t(d,u) \leq .\}dF_{U_{it}}(u) = F_{Y_{it}(d)}$. (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 $\tau \in \mathscr{T}$, $r = 0,1$.

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 $d \in \mathscr{D}$ and $t = 0,1$ and $(U_{i0},U_{i1}) \perp (T_i,R_i)$. Similar to (a.ii), we let $U_{it}' = \mathbf{Y}_{it}(.) = (Y_{it}(0),Y_{it}(1),\ldots,Y_{it}(|\mathscr{D}|-1))$ and $\mu_t(d,U_{it}) = \sum_{j=0}^{\mathscr{D}-1} 1\{j = d\}Y_{it}(j)$ for $d \in \mathscr{D}$, $t = 0,1$. 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 $\tau \in \mathscr{T}$

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

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

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

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

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

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

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. □

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.

A Completely randomized trials

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

$$\text{(SA6.3)} \qquad H_0^{1,\mathscr{T}} : F_{Y_{i0}|T_i=\tau,R_i=r} = F_{Y_{i0}|T_i=\tau',R_i=r} \text{ for } \tau,\tau' \in \mathscr{T},\ \tau \neq \tau',\ r = 0,1.$$

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

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

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

Let $(\tau,r) \in \mathscr{T} \times \mathscr{R}$, where $\mathscr{R} = \{0,1\}$. Let (τ_j, r_j) denote the j^{th} element of $\mathscr{T} \times \mathscr{R}$, then the null hypothesis implied by Proposition 4(b.ii) is given by the following:

$$\text{(SA6.4)} \qquad H_0^{2,\mathscr{T}} :\ F_{Y_{i0}|T_i=\tau_j,R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1},R_i=r_{j+1}} \text{ for } j = 1,\ldots,|\mathscr{T}\times\mathscr{R}|-1.$$

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

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

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

B Stratified randomized trials

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

$$\text{(SA6.5)} \qquad H_0^{1,\mathscr{S},\mathscr{T}} : F_{Y_{i0}|T_i=\tau,S_i=s,R_i=r} = F_{Y_{i0}|T_i=\tau',S_i=s,R_i=r} \text{ for } \tau,\tau' \in \mathscr{T},\ \tau \neq \tau',\ s \in \mathscr{S},\ r = 0,1.$$

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

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

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

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

$$\text{(SA6.6)} \qquad H_0^{2,\mathscr{S},\mathscr{T}} :\ F_{Y_{i0}|T_i=\tau_j,S_i=s,R_i=r_j} = F_{Y_{i0}|T_i=\tau_{j+1},S_i=s,R_i=r_{j+1}} \text{ for } j = 1,\ldots,|\mathscr{T}\times\mathscr{R}|-1,\ s \in \mathscr{S}.$$

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

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

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

SA7 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 IVal-R and IVal-P assumptions.

1 Simulation Design and Test Statistics

The data-generating process (DGP) is described in Panel A of Table SA1. 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 IVal-R and IVal-P as summarized in Panel B of Table SA1.[78] Designs I and II present cases where the differential rate test would have desirable properties as a test of IVal-R.[79] 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 IVal-R. Design II rules out treatment-only responders and, as a result, we have IVal-R, but not IVal-P.

Designs III and IV illustrate *Examples 1* and *2* in Section III.C., respectively. Design III demonstrates a setting in which we have differential attrition rates and IVal-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.

[78] 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 IVal-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 IVal-R only (Proposition 3(iii)).

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

Table SA1 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:	$\begin{cases} U_{it} = (\alpha_i, \eta_{it})', \ t = 0,1, \\ \alpha_i \mid R_i(0), R_i(1) \overset{i.i.d.}{\sim} \begin{cases} N(\delta_{00}, 1) \ if \ (R_i(0), R_i(1)) = (0,0), \\ N(\delta_{01}, 1) \ if \ (R_i(0), R_i(1)) = (0,1), \\ N(\delta_{10}, 1) \ if \ (R_i(0), R_i(1)) = (1,0), \\ N(\delta_{11}, 1) \ if \ (R_i(0), R_i(1)) = (1,1). \end{cases} \\ \eta_{i1} = 0.5\eta_{i0} + \varepsilon_{i0}, \ (\eta_{i0}, \varepsilon_{i0})' \overset{i.i.d.}{\sim} N(0, 0.5 I_2) \end{cases}$

Panel B. Variants of the Design

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

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 IVal-R and IVal-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 IVal-R and IVal-P tests (denoted with an M subscript) are given by:

(SA7.1) $$Y_{i0} = \gamma_{11}T_iR_i + \gamma_{01}(1-T_i)R_i + \gamma_{10}T_i(1-R_i) + \gamma_{00}(1-T_i)(1-R_i) + \varepsilon_i$$

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

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

(SA7.2) $$H^{1}_{0,M}: \ \gamma_{10} = \gamma_{00} \ \& \ \gamma_{11} = \gamma_{01}, \qquad (IV\text{-}R)$$

(SA7.3) $$H^{2}_{0,M}: \ \gamma_{11} = \gamma_{01} = \gamma_{10} = \gamma_{00}, \qquad (IV\text{-}P)$$

$H^{1,1}_{0,M}$ ($H^{1,2}_{0,M}$) 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 IVal-R assumption. $H^{1}_{0,M}$ ($H^{2}_{0,M}$) are the hypotheses of the mean IVal-R (IVal-P) tests in Section III.B.2., which we implement using Wald statistics and asymptotic χ^2 critical values. To implement the distributional IVal-R and IVal-P tests, we use Kolmogorov-Smirnov-type (KS) statistics of their respective hypotheses,

(SA7.4) $$H^1_0: \ Y_{i0}|T_i, R_i = r \overset{d}{=} Y_{i0}|R_i = r, \text{ for } r = 0, 1,$$

(SA7.5) $$H^2_0: \ Y_{i0}|T_i, R_i \overset{d}{=} Y_{i0}.$$

We formally define the KS statistics for the above hypotheses in Section SA2.1, where we also describe the randomization procedures we use to obtain their p-values.

2 Simulation Results

Table SA9 reports simulation rejection probabilities for the differential attrition rate test as well as the mean and distributional tests of the IVal-R and IVal-P assumptions for Designs I-IV. First, we consider the performance of the differential attrition rate test. Columns 1 through 3 of Table

SA9 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 IVal-R. In Design I, where internal validity is violated, the test rejects above 5%, while in Design II, where IVal-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 IVal-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 IVal-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 IVal-R tests, which are given in Columns 4 through 7 of Table SA9. As expected, where IVal-R holds (Designs II and III), the tests control size. Similarly, where IVal-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).[80]

The test statistics of the IVal-P assumption (Columns 8 and 9 in Table SA9) 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 IVal-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. IVal-R holds, but IVal-P does not). Thus, although the IVal-P test does reject, the IVal-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

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

IVal-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 IVal-R and IVal-P assumptions we propose behave according to our theoretical predictions. In what follows, we examine the finite-sample performance of a wider variety of the distributional tests of the IVal-R and IVal-P assumptions.

3 Extended Simulations for the Distributional Tests

A Comparing different statistics of the distributional hypotheses

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

$$\begin{aligned} H_0^{1,1} &: \; Y_{i0}|T_i=1,R_i=0 \stackrel{d}{=} Y_{i0}|T_i=0,R_i=0, && (CA-TA) \\ H_0^{1,2} &: \; Y_{i0}|T_i=1,R_i=1 \stackrel{d}{=} Y_{i0}|T_i=0,R_i=1, && (CR-TR) \\ H_0^{1} &: \; H_0^{1,1} \;\&\; H_0^{1,2}. && (Joint) \end{aligned} \tag{SA7.6}$$

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

$$\begin{aligned} KS^1_{n,r} &= \max_{i:R_i=r} \left|\sqrt{n}\left(F_{n,Y_{i0}}(y_{i0}|T_i=1,R_i=r) - F_{n,Y_{i0}}(y_{i0}|T_i=0,R_i=r)\right)\right|. \\ CM^1_{n,r} &= \frac{\sum_{i:R_i=r}\left(\sqrt{n}(F_{n,Y_{i0}}(y_{i0}|T_i=1,R_i=r) - F_{n,Y_{i0}}(y_{i0}|T_i=0,R_i=r)\right)^2}{\sum_{i=1}^n 1\{R_i=r\}} \end{aligned} \tag{SA7.7}$$

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 SA10 presents the simulation rejection probabilities of the aforementioned statistics of the IVal-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 IVal-R holds. They also have non-trivial power in finite samples in Designs I and IV, when IVal-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 IVal-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 IVal-P assumption. Proposition 1(b.ii) in the paper implies the three simple null hypotheses as well as their joint hypothesis below,

$$
\begin{aligned}
H_0^{2,1} &: \ Y_{i0}|T_i = 0, R_i = 0 \overset{d}{=} Y_{i0}|T_i = 0, R_i = 1, && (CA-CR)\\
H_0^{2,2} &: \ Y_{i0}|T_i = 0, R_i = 1 \overset{d}{=} Y_{i0}|T_i = 1, R_i = 0, && (CR-TA)\\
H_0^{2,3} &: \ Y_{i0}|T_i = 1, R_i = 0 \overset{d}{=} Y_{i0}|T_i = 1, R_i = 1, && (TA-TR)\\
H_0^{2} &: \ H_0^{2,1} \ \&\ H_0^{2,2} \ \&\ H_0^{2,3}. && (Joint)
\end{aligned}
\tag{SA7.8}
$$

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

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

$$\text{(SA7.9)} \qquad CM_{n,j}^2 = \frac{\sum_{i:(T_i,R_i)\in\{(\tau_j,r_j),(\tau_{j+1},r_{j+1})\}} \left(\sqrt{n}\left(F_{n,Y_{i0}|T_i=\tau_{j-1},R_i=r_{j-1}} - F_{n,Y_{i0}|T_i=\tau_j,R_i=r_j}\right)\right)^2}{\sum_{i=1}^n 1\left\{(T_i,R_i)\in\{(\tau_j,r_j),(\tau_{j+1},r_{j+1})\}\right\}},$$

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 SA11, we report the simulation rejection probabilities for distributional tests of the IVal-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, IVal-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 IVal-P holds, our simulation results illustrate that the test statistics we consider control size.

B 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 $P(R_i=0|T_i=0)=0.15$.[81] For the joint hypotheses, we report the simulation results for the KS statistic that takes the maximum over the individual statistics.

Panel A in Figure SA1 displays the simulation rejection probabilities of the tests of the IVal-R assumption while Panel B displays the simulation rejection probabilities of the tests of the IVal-

[81] 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 II in the paper for details.

P assumption. We present these rejection probabilities for alternative parameter values of the designs we consider in Section SA7 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 SA2 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 IVal-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.

SA8 Tables and Figures

Table SA2 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.

Table SA3 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 SA4 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	17.3	10.3	1	46	10	22
Papers that test on multiple baseline variables:						
Multiple hypotheses for individual variables (76%)	16.9	9.7	2	46	10	21
Joint hypothesis for all variables (24%)	20.3	11.3	4	44	13	23

Notes: Of the 47 experiments that conduct a selective attrition test, 45 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 45 experiments that test on multiple baseline variables.

Table SA5 Empirical Applications: Outcomes from The Four Field Experiments

ID	Paper	Outcome		Follow-Up		Baseline Sample	
		Description	Target Population	Round	Months Since Baseline	N	# Clusters
1	Duflo et al. (2012)	Student took written exam (=1)	Children 7-10 yrs old	1st	8	2264	113
2		Student's math test score		1st		2227	
3		Student's language test score		1st		2128	
4		Student's total test score		1st		2230	
5		Student took written exam (=1)		2nd	13	2268	
6		Student's math test score		2nd		2242	
7		Student's language test score		2nd		2139	
8		Student's total test score		2nd		2245	
9	Dupas & Robinson (2013)	Contrib. to ROSCA last yr ($), full sample	Self-emp. w/o bank account	Unique	15	375	–
10		Contrib. to ROSCA last yr ($), market vendors				286	–
11		Contrib. to ROSCA last yr ($), bike-taxi drivers				89	–
12	Ambler et al. (2015)	Remittances to target hh ($USD)	Migrants w kin in sec./tert. school	Unique	8	974	126
13	Karlan & Valdivia (2011)	Business results index*	Adult female entrepreneurs	Unique	24	4304	226
14		Total number of workers				4415	
15		Paid workers (=1)				4404	
16		Empowerment index (hh decisions)				4030	
17		Partake in savings decisions (=1)				4467	
18		Partake in fertility decisions (=1)				4141	
19		Partake in decisions on bills' tracking (=1)				4393	
20		Empowerment index (business decisions)**				4138	
21		Empowerment index (all decisions)				3731	
22		Tax formality (=1)				4424	
23		Keep records of sales (=1)				4357	
24		Number of sales locations				4485	
25		Keep records of withdrawal (=1)				1296	
26		Number of income sources				3188	

Notes: The table reports details of the 26 outcomes included in the empirical application in Section V. *Months since baseline* refers to the maximum number of months between baseline and the last follow-up for those analyses that pool data from different rounds or cohorts. * The *business results index* summarizes seven outcomes related to sales and the number of workers. We include only two of these outcomes since the effective attrition rate for the other five outcomes is zero. ** The *index on empowerment in business decisions* summarizes three outcomes related to the participation of the client in these decisions. We do not include these variables separately since they are binary variables with low variance at baseline due to the sample proportions of the event being less than 10%.

Table SA6 Empirical Applications: Mean Baseline Outcome by Treatment-Response Subgroups

ID	Paper	Outcome	Follow-Up	Sample Size at Baseline	Attrition Rate (%)	Mean Baseline Outcome by Group			
						TR	CR	TA	CA
1	Duflo et al. (2012)	Student took written exam (=1)	1st	2264	17.7	0.174	0.197	0.147	0.143
2		Student's math test score	1st	2227	16.4	8.016	8.077	7.559	8.233
3		Student's language test score	1st	2128	16.2	3.713	3.840	3.932	4.231
4		Student's total test score	1st	2230	16.5	11.579	11.791	11.430	12.042
5		Student took written exam (=1)	2nd	2268	22.1	0.170	0.196	0.174	0.143
6		Student's math test score	2nd	2242	21.4	8.016	8.066	7.798	8.336
7		Student's language test score	2nd	2139	21.6	3.794	3.873	3.521	4.137
8		Student's total test score	2nd	2245	21.3	11.635	11.747	11.289	12.257
9	Dupas & Robinson (2013)	Contrib. to ROSCA last yr ($), full sample		375	33.3	4274	3337	3755	3382
10		Contrib. to ROSCA last yr ($), market vendors	Unique	286	31.8	4827	3910	4384	4965
11		Contrib. to ROSCA last yr ($), bike taxi drivers		89	38.2	2777	685	607	1151
12	Ambler et al. (2015)	Remittances to target hh ($USD)	Unique	974	25.6	2429	3005	2342	2296
13	Karlan & Valdivia (2011)	Business results index*		4304	36.1	0.011	0.050	-0.095	-0.050
14		Total number of workers		4415	32.8	1.988	1.980	1.779	1.820
15		Paid workers (=1)		4404	32.7	0.270	0.233	0.210	0.223
16		Empowerment index (hh decisions)		4030	28.2	0.034	0.031	0.032	0.074
17		Partake in savings decisions (=1)		4467	23.9	0.850	0.836	0.833	0.866
18		Partake in fertility decisions (=1)		4141	26.3	0.685	0.715	0.721	0.740
19		Partake in decisions on bills' tracking (=1)		4393	23.6	0.606	0.600	0.609	0.616
20		Empowerment index (business decisions)**		4138	34.8	0.009	0.020	-0.094	-0.018
21		Empowerment index (all decisions)		3731	37.1	0.041	0.045	0.022	0.043
22		Tax formality (=1)	Unique	4424	32.4	0.143	0.161	0.099	0.114
23		Keep records of sales (=1)		4357	33.2	0.284	0.302	0.297	0.285
24		Number of sales locations		4485	23.5	1.061	1.091	1.066	1.075
25		Keep records of withdrawal (=1)		1296	23.8	0.093	0.096	0.095	0.109
26		Number of income sources		3188	25.4	2.318	2.336	0.328	0.305

Notes: The table reports the mean baseline outcome by groups for the 26 outcomes included in the empirical application in Section V. *TR* refers to treatment respondents, *CR* refers to control respondents, *TA* refers to treatment attritors, and *CA* refers to control attritors. * The *business results index* summarizes seven outcomes related to sales and the number of workers. We include only two of these outcomes since the effective attrition rate for the other five outcomes is zero. ** The *index on empowerment in business decisions* summarizes three outcomes related to the participation of the client in these decisions. We do not include these variables separately since they are binary variables with low variance at baseline due to the sample proportions of the event being less than 10%.

Table SA7 Mean Baseline Outcome and Covariates by Group: School Enrollment

Follow-up Sample	School Enrollment				Age				Poverty Index				Head's Educ			
	TR	CR	TA	CA	TR	CR	TA	CA	TR	CR	TA	CA	TR	CR	TA	CA
Pooled	0.88	0.87	0.62	0.60	10.16	10.19	12.48	12.44	618.44	620.11	627.95	627.10	2.77	2.69	2.45	2.37
1st	0.88	0.87	0.55	0.55	10.19	10.22	13.02	12.81	618.18	619.86	632.61	630.25	2.76	2.69	2.44	2.30
2nd	0.90	0.90	0.59	0.60	9.93	9.98	12.79	12.58	617.34	620.20	629.79	625.22	2.79	2.70	2.46	2.41
3rd	0.86	0.86	0.70	0.66	10.34	10.35	11.66	11.90	619.76	620.29	622.00	627.02	2.77	2.68	2.44	2.38

Notes: This table presents the mean baseline value of the variables included in the attrition tests for the outcome of school enrollment in the *Progresa* example discussed in section IV.B.. The sample size is 24,094 children. TR and CR refer to treatment and control respondents, while TA and CA refer to treatment and control attritors. *Pooled* refers to all the three follow-ups.

Table SA8 Mean Baseline Outcome and Covariates by Group: Adult Employment

Panel A: Employment, Age, and Gender

Follow-up Sample	Employment				Age				Male (=1)			
	TR	CR	TA	CA	TR	CR	TA	CA	TR	CR	TA	CA
Pooled	0.46	0.47	0.47	0.48	38.04	38.34	35.77	35.07	0.49	0.48	0.51	0.50
1st	0.46	0.47	0.47	0.47	37.89	38.10	35.53	35.46	0.49	0.48	0.50	0.49
2nd	0.46	0.46	0.47	0.50	38.23	38.57	35.34	34.81	0.48	0.48	0.51	0.51
3rd	0.46	0.47	0.47	0.48	38.01	38.40	36.30	35.15	0.49	0.48	0.50	0.49

Panel B: Marital Status and Household Size by Age Group

Follow-up Sample	Married (=1)				# Children <= 5				# Children 5 – 18				# Adults			
	TR	CR	TA	CA	TR	CR	TA	CA	TR	CR	TA	CA	TR	CR	TA	CA
Pooled	0.79	0.80	0.62	0.64	1.23	1.23	1.25	1.25	2.31	2.34	2.11	2.18	2.88	2.88	3.20	3.17
1st	0.78	0.78	0.63	0.65	1.23	1.23	1.26	1.25	2.30	2.34	2.00	2.03	2.91	2.90	3.16	3.11
2nd	0.80	0.80	0.61	0.64	1.22	1.23	1.26	1.24	2.30	2.33	2.18	2.26	2.86	2.87	3.25	3.16
3rd	0.80	0.80	0.63	0.62	1.23	1.23	1.23	1.26	2.32	2.34	2.08	2.17	2.87	2.86	3.17	3.20

Notes: This table presents the mean baseline value of the variables included in the attrition tests for the outcome of adult employment in the *Progresa* example discussed in section IV.B.. The sample size is 31,175 adults. *TR* and *CR* refer to treatment and control respondents, while *TA* and *CA* refer to treatment and control attritors. *Pooled* refers to all the three follow-ups.

Table SA9 Simulation Results on Differential Attrition Rates and Tests of Internal Validity ($ATE = 0.25$)

Design	Attrition Rates		Differential Attrition Rate Test	Tests of the IVal-R Assumption				Tests of the IVal-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 IVal-R (IVal-P) assumption refer to the regression tests (Appendix A in the paper) of the null hypothesis in (SA7.2) ((SA7.3)). The KS statistics of the IVal-R (IVal- P) assumption are given in Equations (SA2.2) ((SA2.4)), and their p-values are obtained using the proposed randomization procedures in Section SA2.1 ($B = 199$). 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 SA1.

† (*) indicates IVal-R only (IVal-P).

Table SA10 Simulation Results on the KS & CM Randomization Test of IVal-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 (SA7.6). 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 SA1 in the paper.
† (*) indicates IVal-R only (IVal-P).

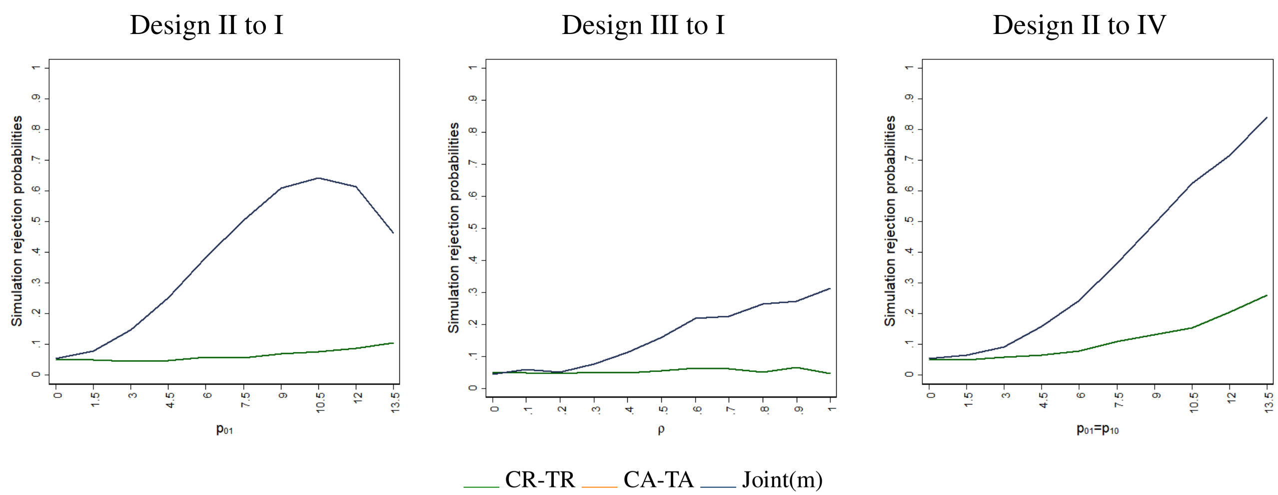
Table SA11 Simulation Results on the KS & CM Randomization Test of IVal-P

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

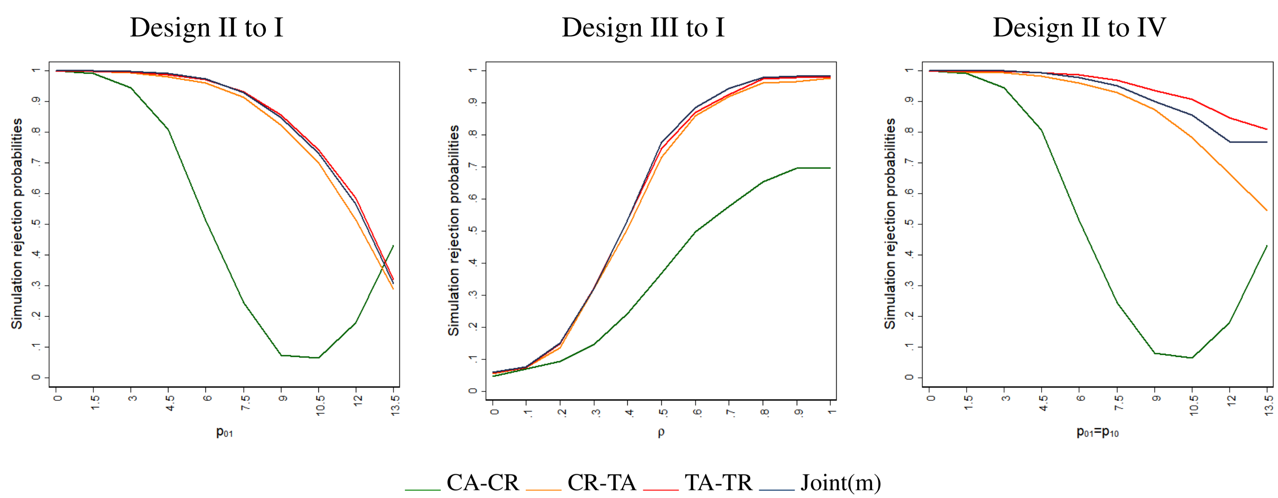
Notes: The above table presents the rejection probabilities of the KS and CM tests for the simple and joint null hypotheses in (SA7.8). We use the nominal level $\alpha = 0.05$, 2,000 simulation replications and $n = 2,000$. C denotes the control group, T denotes the treatment group. $KS(Asym.)$ refers 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 SA1 in the paper.
† (*) indicates IVal-R only (IVal-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



SA9 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 Experiment in Nigeria. *Economic Journal*, 124(574), F327–F355.

Crépon, B., Devoto, F., Duflo, E., & Parienté, W. (2015). Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco. *American Economic Journal: Applied Economics*, 7(1), 123-150.

Cunha, J. M. (2014). Testing Paternalism: Cash versus In-Kind Transfers. *American Economic Journal: Applied Economics*, 6(2), 195-230.

De Grip, A., & Sauermann, J. (2012). The Effects of Training on Own and Co-worker Productivity: Evidence from a Field Experiment. *Economic Journal*, 122(560), 376-399.

de Mel, S., McKenzie, D., & Woodruff, C. (2014). Business training and female enterprise start-up, growth, and dynamics: Experimental evidence from Sri Lanka. *Journal of Development Economics*, 106, 199-210.

De Mel, S., McKenzie, D., & Woodruff, C. (2012). Enterprise Recovery Following Natural Disasters. *Economic Journal*, 122(559), 64-91.

de Mel, S., McKenzie, D., & Woodruff, C. (2013). The Demand for, and Consequences of, Formalization among Informal Firms in Sri Lanka. *American Economic Journal: Applied Economics*, 5(2), 122-150.

Dinkelman, T., & Martínez A., C. (2014). Investing in Schooling In Chile: The Role of Information about Financial Aid for Higher Education. *Review of Economics and Statistics*, 96(2), 244-257.

Doi, Y., McKenzie, D., & Zia, B. (2014). Who you train matters: Identifying combined effects of financial education on migrant households. *Journal of Development Economics*, 109, 39–55.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5), 1739-1774.

Duflo, E., Greenstone, M., Pande, R., & Ryan, N. (2013). Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India. *Quarterly Journal of Economics*, 128(4), 1499-1545.

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives Work: Getting Teachers to Come to School. *American Economic Review*, 102(4), 1241-1278.

Dupas, P., & Robinson, J. (2013). Savings Constraints and Microenterprise Development: Evidence from a Field Experiment in Kenya. *American Economic Journal: Applied Economics*, 5(1), 163-192.

Edmonds, E. V, & Shrestha, M. (2014). You get what you pay for: Schooling incentives and child labor. *Journal of Development Economics*, 111, 196-211.

Fafchamps, M., McKenzie, D., Quinn, S., & Woodruff, C. (2014). Microenterprise growth and the flypaper effect: Evidence from a randomized experiment in Ghana. *Journal of Development Economics*, 106(Supplement C), 211-226.

Fafchamps, M., & Vicente, P. C. (2013). Political violence and social networks: Experimental evidence from a Nigerian election. *Journal of Development Economics*, 101(Supplement C), 27-48.

Ferraro, P. J., & Price, M. K. (2013). Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment. *Review of Economics and Statistics*, 95(1), 64-73.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Baicker, K. (2012). The Oregon Health Insurance Experiment: Evidence from the First Year. *Quarterly Journal of Economics*, 127(3), 1057-1106.

Fryer, J. R. G. (2011). Financial Incentives and Student Achievement: Evidence from Randomized Trials. *Quarterly Journal of Economics*, 126(4), 1755-1798.

Fryer, J. R. G. (2014). Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments. *Quarterly Journal of Economics*, 129(3), 1355-1407.

Gertler, P. J., Martinez, S. W., & Rubio-Codina, M. (2012). Investing Cash Transfers to Raise Long-Term Living Standards. *American Economic Journal: Applied Economics*, 4(1), 164-192.

Giné, X., Goldberg, J., & Yang, D. (2012). Credit Market Consequences of Improved Personal Identification: Field Experimental Evidence from Malawi. *American Economic Review*, 102(6), 2923-2954.

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

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

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

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

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

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

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

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

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, intrahousehold 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 Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company. *American Economic Journal: Applied Economics*, 7(3), 51-84.

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

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

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

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

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

Valdivia, M. (2015). Business training plus for female entrepreneurship? Short and mediumterm 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.