Inference for Matched Tuples and Fully Blocked Factorial Designs

Yuehao Bai, University of Michigan

COLLABORATORS:

Jizhou Liu, University of Chicago

Max Tabord-Meehan, University of Chicago

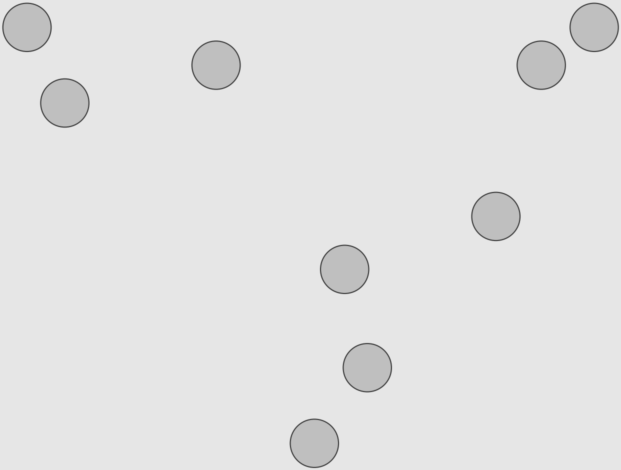
This paper studies randomized controlled trials (RCTs) with multiple treatments.

- "Multiple treatments" in AEA RCT Registry returns $>$700 experiments.

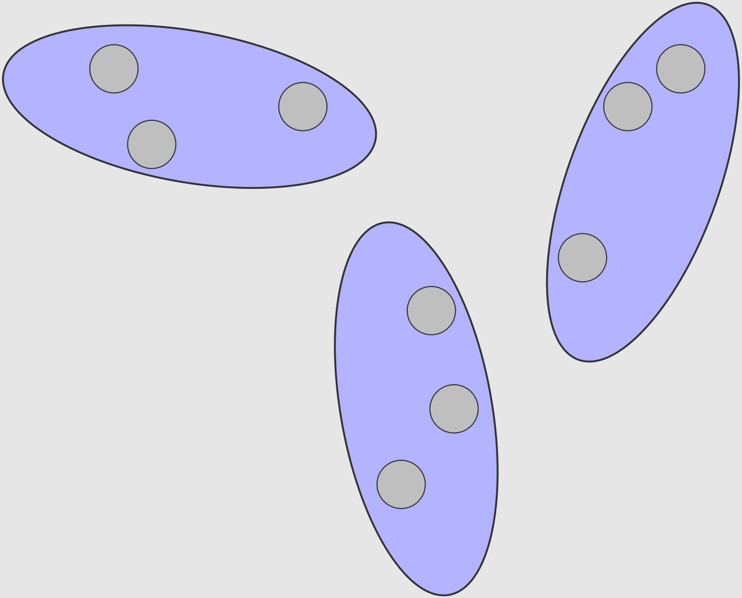
In particular, we study "matched tuples" design.

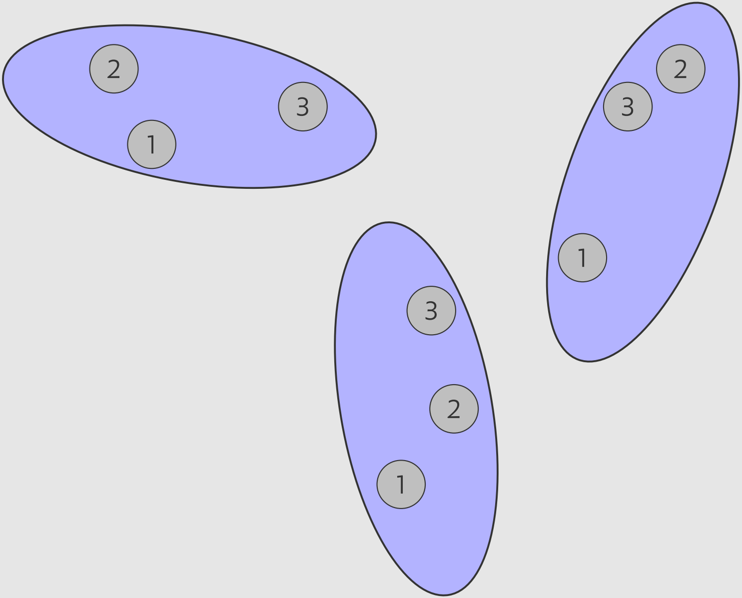
- Partition the units into blocks based on observed covariates.
- Within each block, each treatment is assigned exactly once.
- Examples:
  - Bold et al. (2018)
  - Brown and Andrabi (2020)
  - de Mel, McKenzie, & Woodruff (2013)
  - Fafchamps, McKenzie, Quinn & Woodruff (2014)

INTRODUCTION (CONT.)



INTRODUCTION (CONT.)




2
3
1
2
3
1
3
2
1

Despite popularity, limited formal results on inference beyond matched pairs.

We provide asymptotically exact inference methods for matched tuples.

- i.e., limiting size $=$ level.
- Parameters: pairwise average treatment effects (ATEs) and beyond.

In practice, inference on ATEs often relies on linear regression + block fixed effects.

- + heteroskedasticity-robust variance estimator.
- We show it is invalid.

For all parameters, we show blocking achieves smaller variance than large strata.

Results are also applied to factorial designs.

- e.g., 2×2 factorial design has two factors and four treatment status.
  - Main effect: effect of factor 1 averaging over values of factor 2.
  - Interaction effect between factors 1 and 2.
- Propose "fully blocked" factorial designs.
- In simulation, compares favorably to large strata and re-randomization.

Matched-pair designs:

Imai, King & Nall (2009), Athey and Imbens (2017), Bai, Romano & Shaikh (2021), Bai (2022), de Chaisemartin and Ramirez-Cuellar (2020), Cytrynbaum (2021)

Factorial designs:

Branson, Dasgupta & Rubin (2016), Dasgupta, Pillai & Rubin (2015), Li, Ding & Rubin (2020), Muradliharan, Romero & Wuthrich (2019), Pashley and Bind (2019), Wu and Hamada (2011)…

Application of factorial designs in economics:

Alatas et al. (2012), Besedes et al. (2012), Dellavigna et al. (2016), Karlan et al. (2014)…

i.i.d. units $1 \leq i \leq |\mathcal{D}|n$ drawn from superpopulation. For the ith unit:

$$\begin{aligned} X_i &= \text{observed covariates} \in \mathbf{R}^p \\ D_i &= \text{treatment status} \in \mathcal{D} = \{1, \ldots, |\mathcal{D}|\} \\ Y_i(d) &= \text{potential outcome if treament status were } d \\ Y_i &= \text{observed outcome} = \sum_{d \in \mathcal{D}} Y_i(d) I\{D_i = d\} . \end{aligned}$$

Let Q denote the distr. of $(Y_i(1), Y_i(2), \ldots, Y_i(|\mathcal{D}|), X_i)$.

For each random vector A, define $A^{(n)} = (A_1, \ldots, A_n)'$.

Example from Fafchamps, McKenzie, Quinn & Woodruff (2014):

- They study effects of capital aid program on profits of small businesses.
- X_i: sector, capital stock, pre-treatment profits.
- $D_i \in \{1, 2, 3\}$: control, in-kind grants, cash transfer.
- Y_i: profits.

Parameter of interest

Define $\Gamma(Q) = (\Gamma_1(Q), \ldots, \Gamma_{|\mathcal{D}|}(Q))$ with $\Gamma_d(Q) = E_Q[Y_i(d)]$.

Let ν be $m \times |\mathcal{D}|$ matrix.

Parameter of interest is $\Delta_\nu(Q) = \nu\Gamma(Q) \in \mathbf{R}^m$.

e.g., $\mathcal{D} = \{1, 2\}$, $\nu = (-1, 1)$, $\Delta_\nu(Q) = E_Q[Y_i(2) - Y_i(1)]$.

Also includes main and interaction effects in factorial designs (see later).

Assumptions on distribution Q:

(a) $E_Q[Y_i^2(d)] < \infty$.

(b) $E_Q[Y_i(d)|X_i = x]$ and $E_Q[Y_i^2(d)|X_i = x]$ are Lipschitz for $d \in \mathcal{D}$.

Blocking

First stratify sample into n blocks of size $|\mathcal{D}|$ based on $X^{(n)}$.

For example, if $\dim(X_i) = 1$, order and block adjacent $|\mathcal{D}|$ units.

Denote jth block by λ_j.

Assumption on blocks:

$$\frac{1}{n} \sum_{1 \le j \le n} \sum_{i,k \in \lambda_j} |X_i - X_k|^2 \xrightarrow{P} 0 \, .$$

If $\dim(X_i) = 1$, then satisfied when $E[X_i^2] < \infty$.

Cases with more covariates are discussed in paper.

Assumption on treatment assignment:

(a) $(Y^{(n)}(d) : d \in \mathcal{D}) \perp\!\!\!\perp D^{(n)} | X^{(n)}$.

(b) Conditional on $X^{(n)}$,

$$\{(D_i : i \in \lambda_j) : 1 \leq j \leq n\}$$

are i.i.d. and each uniformly distributed over all permutations of

$$\{1, 2, \ldots, |\mathcal{D}|\} .$$

Setup and Notation

Main Results

Factorial Designs

Simulations

Empirical Application

Estimator

For $d \in \mathcal{D}$, estimate $\Gamma_d(Q) = E_Q[Y_i(d)]$ by

$$\hat{\Gamma}_n(d) = \frac{1}{n} \sum_{1 \le i \le |\mathcal{D}|n} I\{D_i = d\} Y_i \, .$$

Under previous assumptions,

$$(\sqrt{n}(\hat{\Gamma}_n(d) - \Gamma_d(Q)) : d \in \mathcal{D}) \xrightarrow{d} N(0, \mathbb{V}) \, ,$$

where $\mathbb{V} = \mathbb{V}_1 + \mathbb{V}_2$ for

$$\mathbb{V}_1 = \operatorname{diag}(E[\operatorname{Var}[Y_i(d)|X_i]] : d \in \mathcal{D})$$

$$\mathbb{V}_2 = \left(\frac{1}{|\mathcal{D}|} \operatorname{Cov}[E[Y_i(d)|X_i], E[Y_i(d')|X_i]] \right)_{d,d' \in \mathcal{D}} \, .$$

Efficiency property

Compare with large strata design: stratify ith unit into S strata based on X_i.

Example: $\{$capital stock $> x$, capital stock $\leq x\}$.

Represented by discrete-valued $h : \text{support}(X) \to \{1, \ldots, S\}$.

Limiting variance is as above, but with X_i replaced by $h(X_i)$.

Theorem For $1 \times |\mathcal{D}|$ vector ν, variance of $\nu\hat{\Gamma}_n$ under matched tuples design is smaller than that under stratified design.

Variance estimator

Can show $\hat{\Gamma}_n(d) \xrightarrow{P} \Gamma_d(Q) = E_Q[Y_i(d)]$.

Can also show

$$\hat{\sigma}_n^2(d) = \frac{1}{n} \sum_{1 \leq i \leq |\mathcal{D}|n} (Y_i - \hat{\Gamma}_n(d))^2 I\{D_i = d\} \xrightarrow{P} \mathrm{Var}[Y_i(d)] .$$

For $\mathbb{V}_1$, note

$$\begin{aligned} E[\mathrm{Var}[Y_i(d)|X_i]] &= \mathrm{Var}[Y_i(d)] - \mathrm{Var}[E[Y_i(d)|X_i]] \\ &= \mathrm{Var}[Y_i(d)] - E[E[Y_i(d)|X_i]^2] + E[Y_i(d)]^2 \end{aligned}$$

Variance estimator (cont.)

Without additional assumption,

$$E[E[Y_i(d)|X_i]^2]$$

is hard to estimate—it requires two units with treatment d in same block.

Assumption on adjacent blocks:

$$\frac{1}{n}\sum_{1\le j\le n/2}\sum_{i\in\lambda_{2j-1},k\in\lambda_{2j}}|X_i-X_k|^2\xrightarrow{P}0\,.$$

Then, take products of units with treatment d in adjacent blocks.

Variance estimator (cont.)

Define

$$\hat{\rho}_n(d,d) = \frac{2}{n} \sum_{1 \le j \le n/2} \Big(\sum_{i \in \lambda_{2j-1}} Y_i I\{D_i = d\} \Big) \Big(\sum_{i \in \lambda_{2j}} Y_i I\{D_i = d\} \Big) .$$

Under previous assumptions, can show

$$\hat{\rho}_n(d,d) \xrightarrow{P} E[E[Y_i(d)|X_i]^2] .$$

Variance estimator (cont.)

For

$$E[E[Y_i(d)|X_i]E[Y_i(d')|X_i]] ,$$

sufficient to look within same block:

$$\hat{\rho}_n(d,d') = \frac{2}{n} \sum_{1 \le j \le n/2} \Big(\sum_{i \in \lambda_{2j-1}} Y_i I\{D_i = d\} \Big) \Big(\sum_{i \in \lambda_{2j}} Y_i I\{D_i = d'\} \Big) .$$

Variance estimator (cont.)

In summary, $\hat{\mathbb{V}}_n = \hat{\mathbb{V}}_{1,n} + \hat{\mathbb{V}}_{2,n} \xrightarrow{P} \mathbb{V}$, where

$$\hat{\mathbb{V}}_{1,n}(d) = \hat{\sigma}_n^2(d) - (\hat{\rho}_n(d,d) - \hat{\Gamma}_n^2(d))$$
$$\hat{\mathbb{V}}_{2,n}(d,d') = \frac{1}{|\mathcal{D}|}(\hat{\rho}_n(d,d') - \hat{\Gamma}_n(d)\hat{\Gamma}_n(d')) \, .$$

Variance estimator (cont.)

Example: $\mathcal{D} = \{1, 2, 3\}$. Parameter $\nu\Gamma(Q) = E_Q[Y_i(2) - Y_i(1)]$, with

$$\nu = (-1, 1, 0)\ .$$

Limiting variance is

$$E[\text{Var}[Y_i(1)|X_i]] + E[\text{Var}[Y_i(2)|X_i]] \\ + \frac{1}{3}E[(E[Y_i(2)|X_i] - E[Y_i(2)] - (E[Y_i(1)|X_i] - E[Y_i(1)]))^2]$$

Q: Focusing on units with $D_i \in \{1, 2\}$, can we treat problem as matched pairs?

In Bai, Romano & Shaikh '22, limiting variance has $\frac{1}{2}$ instead of $\frac{1}{3}$.

As a result, their variance estimator is conservative here!

The following tests are also conservative:

- Two-sample t-test: compare two samples with $D_i = 2$ and $D_i = 1$.
  - Two samples are not independent.
- "Matched triplets" t-test: treat difference in each triplet as an observation.
  - We don't draw triplets but draw units and block them into triplets.

Variance estimator (cont.)

Common practice in applications: estimate $E[Y_i(d)] - E[Y_i(1)]$ via OLS

$$Y_i = \sum_{d \in \mathcal{D} \setminus \{1\}} \beta(d) I\{D_i = d\} + \sum_{1 \le j \le n} \delta_j I\{i \in \lambda_j\} + \epsilon_i \ .$$

Use heteroskedasticity-robust variance estimator for testing

$$H_0 : E[Y_i(d)] - E[Y_i(1)] = \Delta_0 \text{ versus } H_1 : E[Y_i(d)] - E[Y_i(1)] \neq \Delta_0 \ .$$

See Bruhn and McKenzie (2009), Glennester and Takavarasha (2013).

Theorem Limiting rejection probability of this test may $>$ nominal level.

Intuition: number of regressors $\propto$ sample size!

Recap

We propose asymptotically exact methods of inference for linear constrasts.

Test in Bai, Romano & Shaikh (2019) is conservative.

As a result, two-sample t-test and "matched pairs" t-test are conservative.

Test based on block fixed effects is invalid.

Setup and Notation

K factors, each with two levels: -1 (off), $+1$ (on).

$\mathcal{D} = \{-1, +1\}^K$.

Main effect of factor 1: $[(+1,+1) - (-1,+1)] + [(+1,-1) - (-1,-1)]$.

Interaction effect: $[(+1,+1) - (-1,+1)] - [(+1,-1) - (-1,-1)]$.

Factor Combination	Factor 1	Factor 2	Factor 1/2 Interaction
1	-1	-1	+1
2	-1	+1	-1
3	+1	-1	-1
4	+1	+1	+1

Table: Example of a 2^2 factorial design.

Example: Karlan et al. (2014)

Factor 1: cash grant (no, yes)

Factor 2: insurance grant (no, yes)

Four levels of treatment

(no cash, no insurance)

(no cash, insurance)

(cash, no insurance)

(cash, insurance)

In lab experiment, also arises as "crossover" designs.

e.g., factors 1 and 2 are of interest, factor 3 is order of their appearances.

Proposal: fully blocked factorial designs with $|\mathcal{D}| = 2^K$ units in each block.

Previous results all apply with different choices of ν.

e.g., main effect of factor 1, $\nu = (-1, -1, +1, +1)$.

Alternative design: Large strata.

As before, limiting variance under large strata $\geq$ fully blocked designs.

Alternative design: pair units for factor 1, coin flips on all others.

Motivation: may be interested in main effect of factor 1 only.

Theorem Such design has limiting variance $\geq$ fully blocked design.

Unless $E[Y_i(d)|X_i]$ is the same for all d with factor 1 $= 1$ (and -1).

Takeaway: even if only interested in one factor, block on all factors!

Alternative design: re-randomization

- Branson et al. (2016), Dasgupta et al. (2015), Li et al. (2020)
- Draw treatment assignment from a set until it satisfies criterion.
  - e.g., Mahalanobis distances in covariates across groups $\leq$ threshold.
- We have no formal results on comparison vs. ours, but see simulations.

Setup and Notation

Main Results

Factorial Designs

Simulations

Empirical Application

Data generating process

Consider factorial design with $K = 2$. Sample size $4n = 1000$.

$Y_i(d) = \mu_d + \mu_d(X_i) + \sigma_d(X_i)\epsilon_{d,i}$.

$\mu_d(X_i)$ has zero mean, linear/sin/quadratic/combination.

Focus on the model with $X_i \sim N(0, 1)$ and

$$\mu_d(X_i) = \sin(\gamma_d X_i) + X_i^2 - 1\,.$$

Parameters

Main effects of factors 1 and 2.

Interaction effect between factors 1 and 2.

Partial effect of factor 1 fixing factor 2 $= 1$ or -1 (shown in paper).

Treatment assignment mechanism

B-B: Bernoulli for both factors.

C: completely randomized.

MP-B: matched pair for factor 1, bernoulli for factor 2.

MT: fully blocked design.

Large-2: two large strata.

Large-4: four large strata.

RE: re-randomization design with criterion in Branson et al. (2016).

One covariate, two factors: MSE

Parameter	**B-B**	**C**	**MP-B**	**MT**	**Large-2**	**Large-4**	**RE**
Δ_{ν_1}	19.771	19.301	2.866	1.000	16.282	11.322	16.995
Δ_{ν_2}	11.288	10.677	10.473	1.000	10.443	6.611	10.887
$\Delta_{\nu_{1,2}}$	20.901	21.778	19.943	1.000	19.382	12.819	19.556

Table: Ratio of MSE under all designs against those under matched tuples in one model. ν_1: main effect of factor 1. ν_2: main effect of factor 2. $\nu_{1,2}$: interaction effect.

One covariate, two factors: rej. prob.

Takeaway: despite conservativeness, MT may exhibit high power.

This is the only model where MT is conservative when $\dim(X_i) = 1$.

	Under H_0					Under H_1				
Parameter	**B-B**	**C**	**MT**	**Large-2**	**Large-4**	**B-B**	**C**	**MT**	**Large-2**	**Large-4**
Δ_{ν_1}	0.044	0.058	0.034	0.052	0.056	0.102	0.107	0.694	0.107	0.140
Δ_{ν_2}	0.046	0.053	0.033	0.046	0.051	0.062	0.068	0.224	0.070	0.069
$\Delta_{\nu_{1,2}}$	0.064	0.052	0.021	0.050	0.045	0.060	0.053	0.048	0.058	0.059

Table: Rejection probabilities under the null and alternative hypotheses in one model. ν_1: main effect of factor 1. ν_2: main effect of factor 2. $\nu_{1,2}$: interaction effect.

Experiments with more K and more covariates

Same model is fixed throughout:

$$Y_i(d) = \begin{cases} \tau d^{(1)} + (\tilde{X}_i - 1/2)'\beta + \epsilon_i, & \text{if } K = 1 \\ \tau \cdot \left(d^{(1)} + \frac{\sum_{k \geq 2} d^{(k)}}{K-1} \right) + \gamma_d (\tilde{X}_i - 1/2)'\beta + \epsilon_i, & \text{if } K \geq 2\,, \end{cases}$$

where $d^{(k)}$ denotes the value of the kth factor.

$\dim(\tilde{X}_i) \equiv 10$. Observed number of covariates $\dim(X_i)$ varies.

$\gamma_d = 1$ if $d^{(2)} = 1$ and $\gamma_d = -1$ otherwise.

$\beta = (1, 0.9, \ldots, 0.1)$.

K varies from 1 to 6, sample size $2^6 n = 64n = 1280$.

Parameter of interest is main effect of factor 1.

Experiments with more K and more covariates (cont.)

In terms of MSE, matched tuples is always the lowest.

Yet blocking on more observed covariates $\not\Rightarrow$ lower MSE.

$\dim(X_i)$	Method	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
1		1.000	0.957	0.952	0.985	0.986	0.974
2		0.714	0.725	0.732	0.700	0.723	0.741
4	**MT**	0.374	0.369	0.410	0.464	0.523	0.601
8		0.205	0.289	0.384	0.505	0.646	0.793
10		0.251	0.353	0.460	0.589	0.735	0.910

Table: Ratio of MSE against matched tuples with single factor and covariate.

Experiments with more K and more covariates (cont.)

Also consider "replicate" design MT2: double # units per block.

MT displays some conservativeness while replicate design MT2 doesn't!

		Under H_0					
Method	$\dim(X_i)$	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
MT	1	0.049	0.051	0.050	0.056	0.051	0.049
	2	0.052	0.051	0.056	0.048	0.045	0.038
	4	0.045	0.042	0.037	0.030	0.025	0.023
	8	0.026	0.019	0.014	0.016	0.016	0.020
	10	0.026	0.022	0.015	0.017	0.019	0.026
MT2	1	0.050	0.049	0.054	0.048	0.054	0.048
	2	0.053	0.051	0.048	0.051	0.048	0.053
	4	0.048	0.045	0.052	0.055	0.050	0.050
	8	0.046	0.048	0.049	0.050	0.052	0.050
	10	0.051	0.049	0.049	0.047	0.057	0.048

Table: Rejection probability when testing $H_0 : \Delta_{\nu_1} = 0$ under more factors and covariates.

Experiments with more K and more covariates (cont.)

Q: Why does replicate design control size better?

A: Because quality of variance estimator is higher.

Instructive to consider matched-pair design. Fix four units $\{1, 2, 3, 4\}$.

Component to estimate

$$E[E[Y(1) + Y(0)|X]^2]$$

is (average across all)

$$(Y_1 + Y_2)(Y_3 + Y_4) .$$

Design that put four units in one strata = mixing of three matched-pair designs.

Effectively, average over three variance estimator instead of one.

Large strata designs overreject with large K.

With $K = 6$, each block has only $1280/64/4 = 5$ units assigned to each d.

		Under H_0					
Method	$\dim(X_i)$	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
	1	0.044	0.056	0.064	0.046	0.051	0.063
	2	0.062	0.057	0.053	0.037	0.054	0.073
C	4	0.061	0.053	0.049	0.058	0.047	0.047
	8	0.051	0.049	0.043	0.048	0.061	0.052
	10	0.055	0.050	0.048	0.050	0.051	0.067
	1	0.045	0.050	0.050	0.048	0.071	0.091
	2	0.043	0.055	0.046	0.058	0.065	0.093
Large-4	4	0.059	0.059	0.061	0.062	0.054	0.087
	8	0.055	0.054	0.065	0.064	0.068	0.073
	10	0.054	0.058	0.055	0.061	0.063	0.079

Table: Rejection probability when testing $H_0 : \Delta_{\nu_1} = 0$ under more factors and covariates.

Fafchamps, McKenzie, Quinn & Woodruff (2014)

$D \in \{$control, cash transfer, in-kind grants$\}$.

$Y =$ profit of small business in Ghana.

Slight complication with double control per block, but can modify our procedure.

		All			High initial	Low initial
		Firms	Males	Females	Profit women	Profit women
		(1)	(2)	(3)	(4)	(5)
	Cash treatment	19.64	24.84	16.30	33.09	7.01
OLS		(15.42)	(27.29)	(18.13)	(42.56)	(11.58)
(standard t-test)	In-kind treatment	20.26	4.48	30.42	65.36	11.10
		(15.67)	(18.42)	(22.83)	(53.28)	(15.31)
	Cash=in-kind (p-val)	0.975	0.493	0.600	0.610	0.817
	Cash treatment	19.64	24.84	16.30	33.09	7.01
Difference-in-means		(14.24)	(26.05)	(15.21)	(39.27)	(11.15)
(adjusted t-test)	In-kind treatment	20.26	4.48	30.42	65.36	11.10
		(15.24)	(17.79)	(21.97)	(48.27)	(14.99)
	Cash=in-kind (p-val)	0.974	0.468	0.567	0.576	0.815

Table: Point estimates and standard errors for testing the treatment effects of cash and in-kind grants using different methods (wave 6).

With multiple treatments, use small blocks instead of large strata.

+ our adjusted variance estimator.

For inference, do not use regression with strata fixed effects.

With multiple covariates, consider designs with “replicates.”

$\mu_{1,1} = 2\mu_{1,-1} = 4\mu_{-1,1} = 2\tau$ for $\tau \in \{0, 0.05\}$, $\mu_{-1,-1} = 0$.

$\gamma_{1,1} = 2, \gamma_{-1,1} = 1, \gamma_{1,-1} = 1/2, \gamma_{-1,-1} = -1$.

$(\epsilon_{d,i})_{d \in \mathcal{D}} \perp\!\!\!\perp X_i$ and $\sim N(0, 1)$.

$\sigma_d(X_i) \equiv 1$.