Supplement to "Inference for matched tuples and fully blocked factorial designs"

(*Quantitative Economics*, Vol. 15, No. 2, May 2024, 279–330)

YUEHAO BAI
Department of Economics, University of Southern California

JIZHOU LIU
Booth School of Business, University of Chicago

MAX TABORD-MEEHAN
Department of Economics, University of Chicago

APPENDIX D: ADDITIONAL TABLES AND FIGURES

D.1 *Power plots*

In Section 4.3, we presented truncated power plots for the first and third configurations in order to make the horizontal axes the same as that of the second power plot. In Figure D.1, we present plots showing the entire "S" shape of the power curves for **MT** and **MT2** under all three configurations.

D.2 *Comparing superpopulation and finite population inference*

In this section, we compare the coverage properties of confidence intervals constructed using our proposed variance estimator versus two other well-known estimators, under both the super and finite population approaches to inference. First, we revisit the setting introduced in Section 4.2, but now we consider only the matched tuples design **(MT)**, and construct confidence intervals for the parameter $\Delta_{\nu^1_{-1}}$ using one of three variance estimators:

1. the variance estimator $\hat{\mathbb{V}}_{\nu,n}$ introduced in Section 3.1,
2. a standard heteroskedasticity-robust variance estimator obtained from the regression in (4), and
3. the block-cluster variance estimator considered in Theorem 3.4.

For the superpopulation simulations, we generate the data as in Section 4.2. For the finite population simulations, we simply use each DGP to generate the covariates and outcomes *once*, and then fix these in repeated samples.

Yuehao Bai: yuehao.bai@usc.edu
Jizhou Liu: jliu32@chicagobooth.edu
Max Tabord-Meehan: maxtm@uchicago.edu

Available at http://qeconomics.org. https://doi.org/10.3982/QE2354

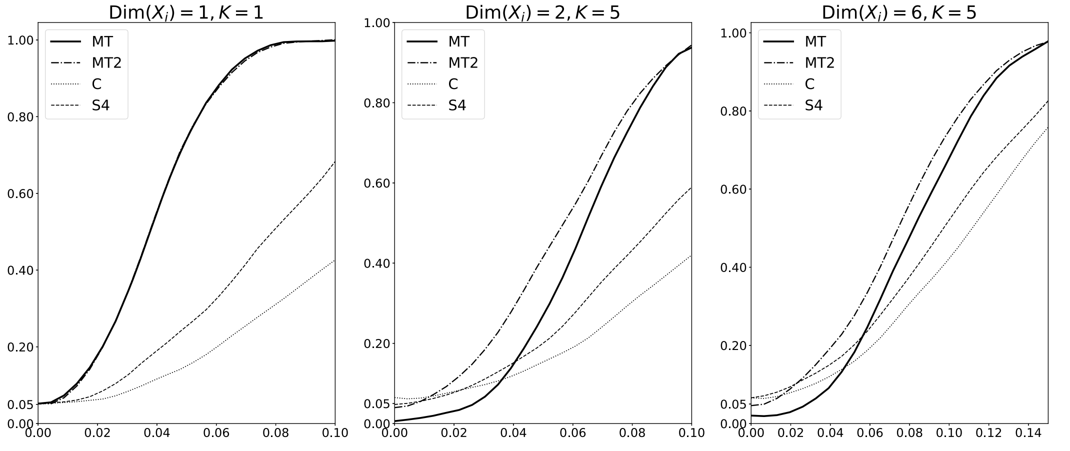

FIGURE D.1. Reject probability under various τs for the alternative hypothesis.

Table D.1 presents coverage probabilities and average confidence interval lengths (in parentheses) with varying sample sizes, based on 2000 Monte Carlo replications. As expected given our theoretical results, $\hat{\mathbb{V}}_{\nu,n}$ delivers exact coverage in large samples under the superpopulation framework in all cases, whereas the robust variance estimator and BCVE are both generally conservative. In the finite population framework, we find that both $\hat{\mathbb{V}}_{\nu,n}$ and BCVE deliver exact coverage for some model specifications in large populations, but all three methods are generally conservative. $\hat{\mathbb{V}}_{\nu,n}$ displays some undercoverage in small populations relative to BCVE, but as the population size increases, $\hat{\mathbb{V}}_{\nu,n}$ generally produces narrower confidence intervals.

Next, we repeat the above exercise using a calibrated simulation design analogous to that used in Section 4.3, but utilizing the wave 6 data from Fafchamps, McKenzie, Quinn, and Woodruff (2014). To construct our data generating process, we run an OLS regression of Y_i on a constant and the seven covariates X_i employed for matching, obtaining $\hat{\beta}$ and residuals $\hat{\epsilon}$. Subsequently, for $d \in \{0, 1, 2\}$ we compute $Y_i(d)$ based on the following model:

$$Y_i(d) = X_i'\hat{\beta} + (X_i - \bar{X}_i)'\hat{\beta} \cdot \gamma \cdot d + \epsilon_i,$$

with X_i drawn from the empirical distribution of the data and $\epsilon_i \sim N(0, \text{var}(\hat{\epsilon}))$. Note that when $\gamma = 0$ we obtain a model with a constant treatment effect of zero, but that as γ increases so does the amount of treatment effect heterogeneity. For the superpopulation simulations, the data is regenerated for each of the Monte Carlo replications. For the finite population simulations, the data is generated only *once* and then fixed in repeated samples. In each experimental assignment, we match the units into triplets and assign one unit to each of $d \in \{0, 1, 2\}$.

Table D.2 presents coverage probabilities and average confidence interval lengths (in parentheses) for the parameter $\Delta_\nu = E[Y_i(1) - Y_i(0)]$, based on 2000 Monte Carlo replications. Our first observation is that given the results for $\gamma = 0$, it is clear that the covariates X_i explain little of the variation in experimental outcomes in our simulation

TABLE D.1. Coverage rate and average CI length (parentheses) under the super and finite population approaches to inference.

Model	Method	Super Population					Finite Population				
		$4n=40$	$4n=80$	$4n=160$	$4n=480$	$4n=1000$	$4n=40$	$4n=80$	$4n=160$	$4n=480$	$4n=1000$
1	$\hat{\mathbb{V}}_{\nu,n}$	0.9340	0.9445	0.9435	0.9460	0.9470	0.9620	0.9550	0.9335	0.9445	0.9535
		(1.810)	(1.253)	(0.881)	(0.508)	(0.351)	(2.002)	(1.547)	(0.923)	(0.480)	(0.354)
	Robust	0.9855	0.9910	0.9930	0.9890	0.9920	0.9905	0.9895	0.9860	0.9950	0.9970
		(2.375)	(1.727)	(1.226)	(0.714)	(0.495)	(2.373)	(1.891)	(1.208)	(0.702)	(0.506)
	BCVE	0.9350	0.9470	0.9400	0.9455	0.9455	0.9185	0.9390	0.9405	0.9470	0.9525
		(1.821)	(1.262)	(0.885)	(0.509)	(0.351)	(1.822)	(1.475)	(0.938)	(0.483)	(0.354)
2	$\hat{\mathbb{V}}_{\nu,n}$	0.9295	0.9395	0.9400	0.9525	0.9505	0.9495	0.9375	0.9405	0.9370	0.9520
		(1.897)	(1.299)	(0.896)	(0.509)	(0.352)	(1.829)	(1.309)	(0.848)	(0.505)	(0.354)
	Robust	0.9850	0.9905	0.9955	0.9965	0.9955	0.9870	0.9820	0.9970	0.9945	0.9980
		(2.489)	(1.809)	(1.290)	(0.751)	(0.522)	(2.337)	(1.560)	(1.354)	(0.749)	(0.540)
	BCVE	0.9185	0.9395	0.9415	0.9545	0.9515	0.9340	0.9395	0.9425	0.9415	0.9530
		(1.858)	(1.282)	(0.893)	(0.508)	(0.352)	(1.789)	(1.311)	(0.852)	(0.518)	(0.356)
3	$\hat{\mathbb{V}}_{\nu,n}$	0.9445	0.9545	0.9600	0.9435	0.9450	0.9970	0.9790	0.9975	0.9890	0.9945
		(2.499)	(1.702)	(1.193)	(0.679)	(0.469)	(2.439)	(1.710)	(1.144)	(0.686)	(0.468)
	Robust	0.9800	0.9915	0.9920	0.9905	0.9910	1.0000	0.9985	1.0000	0.9995	1.0000
		(3.080)	(2.222)	(1.593)	(0.922)	(0.640)	(3.112)	(2.228)	(1.485)	(0.916)	(0.654)
	BCVE	0.9915	0.9940	0.9980	0.9960	0.9965	0.9995	0.9995	1.0000	1.0000	1.0000
		(3.748)	(2.578)	(1.811)	(1.032)	(0.714)	(3.766)	(2.628)	(1.729)	(1.015)	(0.709)

(*Continues*)

TABLE D.1. *Continued.*

Model	Method	Super Population					Finite Population				
		$4n=40$	$4n=80$	$4n=160$	$4n=480$	$4n=1000$	$4n=40$	$4n=80$	$4n=160$	$4n=480$	$4n=1000$
4	$\hat{\mathbb{V}}_{\nu,n}$	0.9355	0.9480	0.9375	0.9445	0.9470	0.9310	0.9345	0.9540	0.9535	0.9640
		(1.889)	(1.319)	(0.927)	(0.534)	(0.371)	(1.674)	(1.292)	(1.015)	(0.562)	(0.373)
	Robust	0.9470	0.9680	0.9580	0.9635	0.9655	0.9435	0.9560	0.9695	0.9685	0.9770
		(1.931)	(1.406)	(1.005)	(0.584)	(0.406)	(1.751)	(1.410)	(1.085)	(0.599)	(0.407)
	BCVE	0.9550	0.9740	0.9700	0.9710	0.9750	0.9730	0.9760	0.9750	0.9760	0.9815
		(2.208)	(1.543)	(1.077)	(0.617)	(0.428)	(2.190)	(1.572)	(1.149)	(0.655)	(0.432)
5	$\hat{\mathbb{V}}_{\nu,n}$	0.9315	0.9435	0.9495	0.9465	0.9530	0.9620	0.9615	0.9735	0.9625	0.9680
		(2.012)	(1.386)	(0.962)	(0.550)	(0.381)	(2.244)	(1.153)	(0.975)	(0.554)	(0.377)
	Robust	0.9530	0.9660	0.9790	0.9770	0.9850	0.9805	0.9870	0.9950	0.9870	0.9875
		(2.152)	(1.570)	(1.117)	(0.650)	(0.452)	(2.472)	(1.415)	(1.162)	(0.655)	(0.448)
	BCVE	0.9615	0.9730	0.9790	0.9785	0.9845	0.9610	0.9915	0.9930	0.9880	0.9870
		(2.419)	(1.667)	(1.155)	(0.662)	(0.458)	(2.506)	(1.530)	(1.151)	(0.656)	(0.453)
6	$\hat{\mathbb{V}}_{\nu,n}$	0.9065	0.9290	0.9305	0.9425	0.9505	0.9105	0.9675	0.9655	0.9715	0.9665
		(4.730)	(3.361)	(2.388)	(1.388)	(0.961)	(4.846)	(3.244)	(2.233)	(1.425)	(1.025)
	Robust	0.9425	0.9600	0.9615	0.9660	0.9670	0.9625	0.9835	0.9855	0.9835	0.9765
		(5.001)	(3.624)	(2.606)	(1.521)	(1.055)	(5.392)	(3.449)	(2.437)	(1.549)	(1.090)
	BCVE	0.9560	0.9675	0.9660	0.9725	0.9735	0.9670	0.9875	0.9865	0.9865	0.9860
		(5.623)	(3.930)	(2.767)	(1.595)	(1.101)	(5.886)	(3.812)	(2.537)	(1.611)	(1.166)

TABLE D.2. Coverage rate and average CI length (parentheses) under the super and finite population approaches to inference.

Model	Method	Super Population					Finite Population				
		$3n=60$	$3n=120$	$3n=360$	$3n=750$	$3n=1200$	$3n=60$	$3n=120$	$3n=360$	$3n=750$	$3n=1200$
$\gamma=0$	$\hat{\mathbb{V}}_{\nu,n}$	0.949	0.943	0.946	0.946	0.952	0.950	0.940	0.955	0.946	0.953
		(225.457)	(160.525)	(92.715)	(64.226)	(50.706)	(225.896)	(159.946)	(92.607)	(64.235)	(50.771)
	Robust	0.950	0.943	0.950	0.947	0.952	0.947	0.943	0.955	0.951	0.955
		(223.224)	(160.560)	(93.791)	(65.160)	(51.503)	(224.081)	(160.511)	(93.731)	(65.128)	(51.553)
	BCVE	0.948	0.938	0.943	0.940	0.946	0.953	0.944	0.954	0.943	0.950
		(229.461)	(162.261)	(92.762)	(64.198)	(50.674)	(230.041)	(161.019)	(92.765)	(64.089)	(50.685)
$\gamma=1$	$\hat{\mathbb{V}}_{\nu,n}$	0.940	0.946	0.953	0.960	0.959	0.946	0.941	0.947	0.948	0.953
		(229.287)	(164.518)	(94.925)	(65.239)	(51.591)	(233.870)	(165.423)	(94.580)	(65.390)	(51.554)
	Robust	0.936	0.955	0.961	0.970	0.963	0.945	0.950	0.954	0.958	0.960
		(230.262)	(166.659)	(97.449)	(67.499)	(53.449)	(232.131)	(167.113)	(97.281)	(67.482)	(53.420)
	BCVE	0.936	0.945	0.957	0.961	0.959	0.949	0.946	0.950	0.950	0.956
		(232.063)	(165.622)	(95.388)	(65.468)	(51.662)	(237.561)	(166.805)	(94.836)	(65.553)	(51.658)
$\gamma=3$	$\hat{\mathbb{V}}_{\nu,n}$	0.947	0.949	0.963	0.966	0.957	0.948	0.952	0.953	0.947	0.952
		(251.942)	(180.451)	(101.057)	(70.280)	(55.300)	(253.653)	(177.162)	(102.184)	(70.042)	(55.324)
	Robust	0.961	0.962	0.978	0.977	0.975	0.951	0.961	0.962	0.968	0.968
		(255.377)	(188.130)	(108.362)	(76.242)	(60.466)	(257.964)	(185.413)	(109.376)	(75.993)	(60.422)
	BCVE	0.947	0.955	0.969	0.971	0.963	0.958	0.957	0.954	0.959	0.961
		(256.837)	(185.391)	(103.913)	(72.470)	(57.259)	(260.735)	(181.843)	(105.186)	(72.325)	(57.091)
$\gamma=5$	$\hat{\mathbb{V}}_{\nu,n}$	0.945	0.947	0.966	0.964	0.957	0.940	0.959	0.978	0.968	0.966
		(285.897)	(199.748)	(111.957)	(78.191)	(60.960)	(284.327)	(200.163)	(113.900)	(77.267)	(60.890)
	Robust	0.959	0.965	0.986	0.981	0.977	0.955	0.970	0.986	0.983	0.982
		(295.771)	(215.171)	(125.135)	(88.824)	(70.149)	(293.489)	(215.318)	(127.164)	(88.177)	(70.040)
	BCVE	0.949	0.958	0.975	0.976	0.970	0.949	0.962	0.981	0.975	0.975
		(296.164)	(209.731)	(119.286)	(83.916)	(65.873)	(293.557)	(209.593)	(121.447)	(83.287)	(65.842)

design since all three variance estimators obtain exact coverage. However, as we artificially increase the amount of treatment effect heterogeneity by increasing the parameter γ, we find that, in line with our theoretical results, both the robust variance estimator and BCVE become slightly conservative. Moreover, in the finite population framework, $\hat{\mathbb{V}}_{\nu,n}$ starts to become conservative as well.

D.3 *Calibrated simulation design details*

In this section, we provide details for the calibrated simulation study considered in Section 4.3. Following Branson, Dasgupta, and Rubin (2016), we consider data obtained from the New York Department of Education, who were considering implementing a 2^5 factorial experiment to study five new intervention programs: a quality review, a periodic assessment, inquiry teams, a schoolwide-performance bonus program and an online resource program; details about each of these programs can be found in Dasgupta, Pillai, and Rubin (2015). The data set contains covariate information for 1376 schools. As in Branson, Dasgupta, and Rubin (2016), we consider experimental designs constructed using nine covariates, which were deemed likely to be correlated with schools' performance scores: total number of students, proportion of male students, enrollment rate, poverty rate, and five additional variables recording the proportion of students of various races.

Since the NYDE has yet to run such an experiment, and given the limitations of the available data set, we select one covariate ("number of teachers") from the original data set to use as the potential outcome under control, and then construct the potential outcomes under the various treatment combinations using the model described in Section 4.3. Specifically, we first demean and standardize all 9 covariates (denoted $\tilde{X}_i$), and then estimate a parameter vector β by ordinary least squares in the following linear model specification for $Y_i(-1, -1, \dots, -1)$:

$$Y_i(-1, -1, \dots, -1) = \gamma_{(-1,-1,\dots,-1)} \tilde{X}_i' \beta + \epsilon_i, \tag{S1}$$

where $\gamma_{(-1,-1,\dots,-1)} = -1$ as defined in Section 4.3. Table D.3 presents the regression results. For each treatment combination d, we then compute $Y_i(d)$ using the model from

TABLE D.3. Model (S1) OLS regression results.

	Coef	Std Err	z	$P > \lvert z\rvert$	[0.025	0.975]
constant	2.824e−06	0.007	0.000	1.000	−0.014	0.014
Total	−0.9808	0.016	−60.609	0.000	−1.012	−0.949
NativeAmerican	0.0374	0.054	0.699	0.485	−0.068	0.143
Black	2.9378	3.175	0.925	0.355	−3.285	9.160
Latino	2.6158	2.836	0.922	0.356	−2.942	8.174
Asian	1.6866	1.822	0.926	0.355	−1.884	5.258
White	1.9064	2.150	0.887	0.375	−2.308	6.121
Male	−0.0379	0.007	−5.355	0.000	−0.052	−0.024
Stability	0.0045	0.007	0.636	0.525	−0.009	0.018
Poverty Rate	−0.1818	0.011	−16.350	0.000	−0.204	−0.160

TABLE D.4. Point estimates and standard errors for testing the treatment effects of cash and in-kind grants using different methods (wave 7).

		All			High Initial	Low Initial
		Firms	Males	Females	Profit Women	Profit Women
		(1)	(2)	(3)	(4)	(5)
OLS without group fixed effects	Cash treatment	18.02	56.17	−8.43	−15.32	−3.84
		(29.66)	(67.95)	(18.25)	(38.99)	(17.14)
	In-kind treatment	31.59	62.02	4.63	42.10	−13.40
		(21.63)	(40.60)	(20.97)	(48.82)	(16.08)
	Cash = in-kind (p-val)	0.680	0.938	0.484	0.171	0.554
Matched tuples	Cash treatment	18.02	56.17	−8.43	−15.32	−3.84
		(26.07)	(60.09)	(17.25)	(42.10)	(16.60)
	In-kind treatment	31.59	62.02	4.63	42.10	−13.40
		(19.47)	(39.02)	(18.57)	(45.30)	(14.32)
	Cash = in-kind (p-val)	0.641	0.931	0.456	0.147	0.556

Section 4.3 given by

$$Y_i(d) = \tau \cdot \left(d^{(1)} + \frac{1}{K-1} \sum_{2 \le k \le K} d^{(k)} \right) + \gamma_d \tilde{X}_i' \beta + \epsilon_i ,$$

where $\tilde{X}_i$ is drawn from the empirical distribution of the data and $\epsilon_i \sim N(0, 0.1)$, where we note that 0.1 is approximately equal to the sample variance of the residuals of the regression in (S1).

D.4 *More results for the empirical application*

In Table D.4, we repeat our analysis for the data on long-term effects obtained through the final round (wave 7) of surveys from the original paper. For the analysis of long-term effects, we follow the same procedure as in the original paper, except we additionally drop the four groups with sizes ranging from 5 to 8. Note that the estimated effects are different for the fixed-effect regression. This is because, as in the analysis in the original paper, we do *not* drop entire quadruplets from our data set whenever one member of the quadruplet was missing due to nonresponse in the final survey round.

REFERENCES

Branson, Zach, Tirthankar Dasgupta, and Donald B. Rubin (2016), "Improving covariate balance in 2K factorial designs via rerandomization with an application to a New York City Department of Education High School Study." *The Annals of Applied Statistics*, 10 (4), 1958–1976. https://doi.org/10.1214/16-AOAS959. [6]

Dasgupta, Tirthankar, Natesh S. Pillai, and Donald B. Rubin (2015), "Causal inference from 2^k factorial designs by using potential outcomes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77 (4), 727–753. [6]

Fafchamps, Marcel, David McKenzie, Simon Quinn, and Christopher Woodruff (2014), "Microenterprise growth and the flypaper effect: Evidence from a randomized experiment in Ghana." *Journal of Development Economics*, 106, 211–226. https://www.sciencedirect.com/science/article/pii/S0304387813001375. [2]

Co-editor James Hamilton handled this manuscript.

Manuscript received 14 March, 2023; final version accepted 10 February, 2024; available online 13 February, 2024.