# A SIMPLE, GRAPHICAL APPROACH TO COMPARING MULTIPLE TREATMENTS: ONLINE SUPPLEMENT 

## S1. THE CASE OF A SINGLE TREATMENT

Here, we provide some additional insight into the choice of $\gamma$ by considering the special case where $k=$ 1 (i.e., the case where a single treatment is compared to a control). In doing so, we will assume that the limiting distribution of $\sqrt{n}\left(\hat{\beta}_{n}-\beta\right)$ is bivariate normal, with mean vector $(0,0)^{\prime}$ and covariance matrix

$$
\left(\begin{array}{cc}
\Sigma_{0} & 0 \\
0 & \Sigma_{1}
\end{array}\right)
$$

We will also assume that $\hat{\Sigma}_{n, s}$ is a consistent estimate of $\Sigma_{s}$, so that se $\left(\hat{\beta}_{n, s}\right)=\sqrt{\hat{\Sigma}_{n, s} / n}$, for $s \in\{0,1\}$.
With $k=1$, our problem only involves testing the hypothesis that $\beta_{0}=\beta_{1}$ (or, equivalently, $\delta_{1}=0$ ). If this hypothesis is true, then

$$
\begin{equation*}
\lim _{n \rightarrow \infty} P\left(\frac{\sqrt{n}\left(\hat{\beta}_{n, 0}-\hat{\beta}_{n, 1}\right)}{\sqrt{\hat{\Sigma}_{n, 0}+\hat{\Sigma}_{n, 1}}}>\Phi^{-1}(1-\alpha / 2)\right)=\alpha / 2 \tag{S.1}
\end{equation*}
$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution function (so that, e.g., $\Phi^{-1}(0.975)=1.960$ ). Rearranging the above, we have

$$
\begin{equation*}
\lim _{n \rightarrow \infty} P\left(L_{n, 0}\left(B_{n} \Phi^{-1}(1-\alpha / 2)\right)>U_{n, 1}\left(B_{n} \Phi^{-1}(1-\alpha / 2)\right)\right)=\alpha / 2 \tag{S.2}
\end{equation*}
$$

where

$$
B_{n}=\frac{\sqrt{\hat{\Sigma}_{n, 0}+\hat{\Sigma}_{n, 1}}}{\sqrt{\hat{\Sigma}_{n, 0}}+\sqrt{\hat{\Sigma}_{n, 1}}}
$$

Next, swapping the indices in (S.2), we have

$$
\begin{equation*}
\lim _{n \rightarrow \infty} P\left(L_{n, 1}\left(B_{n} \Phi^{-1}(1-\alpha / 2)\right)>U_{n, 0}\left(B_{n} \Phi^{-1}(1-\alpha / 2)\right)\right)=\alpha / 2 \tag{S.3}
\end{equation*}
$$

Thus, since the events within (S.2) and (S.3) are disjoint, we have

$$
\lim _{n \rightarrow \infty} P\left(\max _{s \in\{0,1\}} L_{n, s}\left(B_{n} \Phi^{-1}(1-\alpha / 2)\right)>\min _{s \in\{0,1\}} U_{n, s}\left(B_{n} \Phi^{-1}(1-\alpha / 2)\right)\right)=\alpha
$$

Accordingly, rather than choosing for $\gamma$ via re-sampling, we could form uncertainty intervals for $\beta_{0}$ and $\beta_{1}$ as

$$
\begin{equation*}
C_{n, 0}\left(B_{n} \Phi^{-1}(1-\alpha / 2)\right) \tag{S.4}
\end{equation*}
$$

and

$$
\begin{equation*}
C_{n, 1}\left(B_{n} \Phi^{-1}(1-\alpha / 2)\right) \tag{S.5}
\end{equation*}
$$

respectively, and use these uncertainty intervals to make inferences about the ordering of $\beta_{0}$ and $\beta_{1}$ as usual. In doing so, the probability that we spuriously infer that either $\beta_{0}>\beta_{1}$ or $\beta_{1}>\beta_{0}$ is exactly equal to $\alpha$ asymptotically.

These results demonstrate that basing inferences about the ordering of $\beta_{0}$ and $\beta_{1}$ on the non-overlap of their (asymptotically valid) $(1-\alpha)$-level confidence intervals, $C_{n, 0}\left(\Phi^{-1}(1-\alpha / 2)\right)$ and $C_{n, 0}\left(\Phi^{-1}(1-\alpha / 2)\right)$, respectively, is overly conservative (since $B_{n}<1$ ). On the other hand, we can use our uncertainty intervals to form an asymptotically valid $(1-\alpha)$-level confidence interval for $\delta_{1} \equiv \beta_{1}-\beta_{0}$. As shown in Section 3.2
of Bennett and Thompson (2016), hereafter BT, such a confidence interval is simply the difference in the estimates of $\beta_{1}$ and $\beta_{0}$, plus or minus the average lengths of their uncertainty intervals, i.e.,

$$
\left[\left(\hat{\beta}_{n, 1}-\hat{\beta}_{n, 0}\right) \pm B_{n} \Phi^{-1}(1-\alpha / 2) \times\left(\operatorname{se}\left(\hat{\beta}_{n, 1}\right)+\operatorname{se}\left(\hat{\beta}_{n, 0}\right)\right)\right]
$$

Indeed, since $\hat{\delta}_{n, 1} \equiv \hat{\beta}_{n, 1}-\hat{\beta}_{n, 0}$, we have se $\left(\hat{\delta}_{n, 1}\right)=\sqrt{\left(\hat{\Sigma}_{n, 0}+\hat{\Sigma}_{n, 1}\right) / n}$, and the above is simply

$$
\left[\hat{\delta}_{n, 1} \pm \Phi^{-1}(1-\alpha / 2) \times \operatorname{se}\left(\hat{\delta}_{n, 1}\right)\right]
$$

Finally, it can be seen from (S.1) that inferences based on the non-overlap of the uncertainty intervals in (S.4) and (S.5) will be identical to those based on a hypothesis testing procedure that rejects $\beta_{s}=\beta_{t}$ in favour of $\beta_{s}>\beta_{t}$ if

$$
\frac{\sqrt{n}\left(\hat{\beta}_{n, s}-\hat{\beta}_{n, t}\right)}{\operatorname{se}\left(\hat{\beta}_{n, s}-\hat{\beta}_{n, t}\right)}>\Phi^{-1}(1-\alpha / 2) .
$$

## S2. SIMULATION EVIDENCE

S2.1. A Simple Illustration. To illustrate the perils of ignoring the issue of multiple testing in our setting, we generate, for $k \in\{2, \ldots, 10\}$, one million samples from the model (2.1) of the main paper as follows. We assign 100 observations to each of the $k$ treatment groups and to a control group, so that $n=$ $100(k+1)$. We set $\beta_{0}=1$ and $\delta_{1}=\cdots=\delta_{k}=0$, and exclude the $\mathbf{X}_{i}^{\prime} \eta$ term. For each $i, V_{i}$ is an independent standard normal draw.

Within each sample, we use individual $t$-tests at the $5 \%$ nominal level to test (A) each of the $k$ hypotheses in (2.2) of the main paper, and (B) each of the $\binom{k}{2}$ hypotheses in (2.3) of the main paper. As a point of comparison, we also use a single $F$-test at the $5 \%$ nominal level to jointly test the hypothesis that $\delta_{1}=\cdots=$ $\delta_{k}=0$. Clearly, if this hypothesis is true, then all of the hypotheses in (2.2) and (2.3) of the main paper are true as well. However, rejection of this hypothesis provides no guidance as to which of the hypotheses in (2.2) or (2.3) of the main paper ought to be rejected individually; it could be that just one treatment effect is non-zero, or even that all $k$ treatment effects are non-zero (in which case some or all of the treatment effects may be different from each other). ${ }^{19}$

The rejection frequencies for the different types of tests are shown in Figure S1. Specifically, the dashdotted line shows the frequency of rejecting at least one of the $k$ hypotheses in (2.2) of the main paper, the dashed line shows the frequency of rejecting at least one of the $\binom{k}{2}$ hypotheses in (2.3) of the main paper, and the solid line shows the frequency of rejecting at least one of the total of $\binom{k+1}{2}$ hypotheses, i.e., the empirical FWER. The dotted line shows the rejection frequency for the single $F$-test, which is, unsurprisingly, 0.050 for all $k$. Note, however, that even with $k=2$, the empirical FWER across the individual tests is 0.122 ; the frequency of rejecting (A) $\delta_{1}=0$ and/or $\delta_{2}=0$ is 0.091 , and (B) $\delta_{1}=\delta_{2}$ is, unsurprisingly, 0.050 . With $k$ $=10$, the empirical FWER across the individual tests is 0.675 .

S2.2. The Overlap Procedure. We now examine the finite-sample performance of the overlap procedures described above by way of several Monte Carlo experiments. As in BT, we consider a basic (unrefined) max- $T$ procedure as a benchmark for the basic (unrefined) overlap procedure. ${ }^{20}$ Specifically, this procedure rejects the hypothesis $\beta_{s}=\beta_{t}$ in favour of $\beta_{s} \neq \beta_{t}$ whenever the absolute value of

$$
\begin{equation*}
T_{n,(s, t)}=\frac{\hat{\beta}_{n, s}-\hat{\beta}_{n, t}}{\operatorname{se}\left(\hat{\beta}_{n, s}-\hat{\beta}_{n, t}\right)} \tag{S.6}
\end{equation*}
$$

[^0]

Figure S1. Rejection rates for individual $t$-tests and joint $F$-test.
exceeds the $1-\alpha$ quantile of $\max _{(s, t) \in K^{2}}\left|T_{n,(s, t)}^{*}\right|$, where

$$
T_{n,(s, t)}^{*}=\frac{\left(\hat{\beta}_{n, s}^{*}-\hat{\beta}_{n, s}\right)-\left(\hat{\beta}_{n, t}^{*}-\hat{\beta}_{n, t}\right)}{\operatorname{se}\left(\hat{\beta}_{n, s}^{*}-\hat{\beta}_{n, t}^{*}\right)}
$$

Note that, if there are no other explanatory variables (i.e., if the $\mathbf{X}_{i}^{\prime} \eta$ term is excluded from model (2.1) in the main paper), as in the design of our simulations in Section S3.1 above and in what follows, then the estimate of $\operatorname{Cov}\left(\hat{\beta}_{n, s}, \hat{\beta}_{n, t}\right)$, which se $\left(\hat{\beta}_{n, s}-\hat{\beta}_{n, t}\right)$ generally depends on, will be zero for all $s \neq t$. That is, $\operatorname{se}\left(\hat{\beta}_{n, s}-\hat{\beta}_{n, t}\right)=\sqrt{\operatorname{se}\left(\hat{\beta}_{n, s}\right)^{2}+\operatorname{se}\left(\hat{\beta}_{n, t}\right)^{2}}$ in such cases. Nonetheless, $T_{n,(s, t)}$ and $T_{n,\left(s, t^{\prime}\right)}$ will be correlated, since $\hat{\beta}_{n, s}-\hat{\beta}_{n, t}$ and $\hat{\beta}_{n, s}-\hat{\beta}_{n, t^{\prime}}$ are correlated.

It is interesting to point out here that inferences based on the modified overlap procedure, which ignores the comparisons between treatment effects (see Section 2.5 of the main paper), will be identical to those based on a max- $T$ procedure, which rejects $\delta_{s}=0$ in favour of $\delta_{s}>0$ if $T_{n, s}>\lambda$, and of $\delta_{s}<0$ if $T_{n, s}<$ $-\lambda$, where

$$
T_{n, s}=\frac{\hat{\delta}_{n, s}}{\operatorname{se}\left(\hat{\delta}_{n, s}\right)}
$$

The reason for this is that the choice of $\lambda$ suggested in Section 2.5 of the main paper is just the $1-\alpha$ quantile of $\max _{s \in\{1, \ldots, k\}}\left|T_{n, s}^{*}\right|$, where

$$
T_{n, s}^{*}=\frac{\hat{\delta}_{n, s}^{*}-\hat{\delta}_{n, s}}{\operatorname{se}\left(\hat{\delta}_{n, s}^{*}\right)}
$$

The design of our simulations is the same as the one described in Section S3.1 above, but with several variations. First, we consider $k \in\{5,10\}$ and assign $n_{0} \in\{50,100,200\}$ observations to each of the $k$

Table S1. Empirical FWERs for max- $T$ and overlap procedures

| $k$ | $n_{0}$ | Homoskedasticity |  |  | Heteroskedasticity |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | max- $T$ | Overlap | Mod. overlap | max-T | Overlap | Mod. overlap |
| 5 | 50 | 0.052 | 0.050 | 0.052 | 0.053 | 0.051 | 0.051 |
|  | 100 | 0.051 | 0.049 | 0.049 | 0.051 | 0.050 | 0.051 |
|  | 200 | 0.052 | 0.050 | 0.051 | 0.052 | 0.050 | 0.048 |
| 10 | 50 | 0.053 | 0.051 | 0.050 | 0.052 | 0.050 | 0.051 |
|  | 100 | 0.053 | 0.051 | 0.051 | 0.052 | 0.051 | 0.050 |
|  | 200 | 0.051 | 0.049 | 0.051 | 0.051 | 0.049 | 0.050 |

treatment groups and to a control group (so that $n=n_{0}(k+1)$ ). Second, we set $\beta_{s}=\theta(s+1)$, with $\theta \in$ $\{0,0.1,0.2, \ldots, 1\}$, which allows us to examine both control of the FWER (when $\theta=0$ ) and power (when $\theta>0$ ). Finally, we consider two different specifications for the error term distribution: a homoskedastic case in which all of the errors are drawn from the standard normal distribution, and a heteroskedastic case in which the errors for observations assigned to the control group are standard normal, while the observations assigned to treatment group $s \in\{1, \ldots, k\}$ are normal with mean zero and variance $s+1$. In both cases, we estimate $\beta$ using OLS and obtain heteroskedasticity-consistent standard errors (specifically, the HC0 variant of MacKinnon and White, 1985). The bootstrap counterparts of these objects are obtained using 499 replications of the wild bootstrap (Wu, 1986; Liu, 1988; Mammen, 1993) with the Rademacher distribution (Davidson and Flachaire, 2008). Throughout all of our simulations, we set the nominal FWER $\alpha$ equal to 0.05 and generate 100,000 samples.

Table S1 shows that control of the FWER for the max- $T$ procedure and both the unmodified and modified overlap procedures is adequate at all of the sample sizes considered in both the homoskedastic and heteroskedastic cases for both $k=5$ and $k=10$.

In order to compare the power of the max- $T$ procedure and the (unmodified) overlap procedure, we follow BT and Romano and Wolf (2005b) in examining average power, which is the proportion of false hypotheses (of the form $\beta_{s}=\beta_{t}$ when $\theta>0$ ) that are rejected. Figures S2(a) and S2(b) display the empirical average power for these two procedures as a function of $\theta$ in the homoskedastic and heteroskedastic cases, respectively (to save space, we only present results for $k=5$ ). Within these figures, black lines correspond to the overlap procedure, and red lines correspond to the max- $T$ procedure, while lines that are solid, dashed, and dash-dotted correspond to $n_{0}=50, n_{0}=100$, and $n_{0}=200$, respectively. Evidently, both procedures have nearly identical average power for all of the sample sizes considered in the homoskedastic case. In the heteroskedastic case, the max- $T$ procedure has slightly higher average power in the larger sample sizes.

Finally, we turn to the gain in power that results from using the modified overlap procedure. Specifically, we examine the probability that the largest (i.e., the 'best') parameter is correctly identified (here, we always have $\theta>0$, so the largest parameter is $\beta_{k+1}$ ). Figures $\mathrm{S} 3(\mathrm{a})$ and $\mathrm{S} 3(\mathrm{~b})$ display the empirical probability that the 'best' is identified by the two overlap procedures as a function of $\theta$ in the homoskedastic and heteroskedastic cases, respectively (as above, we only present results for $k=5$ ). Within these figures, the black line corresponds to the unmodified overlap procedure, while the blue line corresponds to the modified overlap procedure. To reduce clutter, we only plot the results for $n_{0}=200$, but the results for the other sample sizes are qualitatively similar. Namely, the modified overlap procedure does a much better job in identifying the 'best', particularly in the heteroskedastic case (where the largest parameter estimate has the largest variance). As noted in the previous section, this is due to the fact that $\gamma$ is chosen to be much smaller with the modified overlap procedure, resulting in narrower uncertainty intervals. In all cases shown here, we find that the value of $\gamma$ that is chosen by the modified overlap procedure is slightly less than half as large (on average, over the 10,000 samples) as the value of $\gamma$ that is chosen by the unmodified overlap procedure. More


Figure S2. Empirical average power for max- $T$ and overlap procedures $(k=5)$.


Figure S3. Empirical probability of identifying the best for the unmodified and modified overlap procedures $(k=5)$.
generally, this ratio will decrease as $k$ is increased (since an increasing number of 'irrelevant' comparisons can be eliminated), but will increase as $n$ is increased (since asymptotically, both procedures will correctly identify the largest parameter with probability one).

## S3. COMPUTATIONAL DETAILS FOR PERFORMANCE PAY EXAMPLE

Here, we outline the computational steps involved in choosing $\gamma$ for the example in Section 2.4 of the main paper. Note that the first four steps are concerned with obtaining the parameter estimates and their corresponding standard errors, as well as the bootstrap counterparts of these objects, and would thus be required for any conventional bootstrap-based hypothesis testing procedure; only the fifth step is specific to the overlap procedure (however, this step is not specific to this particular example).
(1) Use OLS to estimate model (2.8) of the main paper. Store the estimates $\hat{\beta}_{n}=\left(\hat{\beta}_{n, 0}, \hat{\beta}_{n, 1}, \hat{\beta}_{n, 2}\right)^{\prime}$ and their corresponding standard errors (clustered by school in this example), as well as the fitted values $\left\{\text { Score }_{i}\right\}_{i=1}^{n}$ and the residuals $\left\{\hat{V}_{i}\right\}_{i=1}^{n}$.
(2) For each $i \in\{1, \ldots, n\}$, generate

$$
\operatorname{Score}_{i}^{*}=\hat{\operatorname{Score}}_{i}+\hat{V}_{i} \sum_{g=1}^{G} W_{g}^{*} I\left(\operatorname{School}_{i}=g\right)
$$

where School is a categorical variable taking values in $\{1, \ldots, G\}$ (here, $G$ is the number of schools), and $\left\{W_{g}^{*}\right\}_{g=1}^{G}$ are drawn from a bootstrap weight distribution (in this example, we use the Rademacher distribution, which places equal probability on -1 and 1; see Davidson and Flachaire, 2008). ${ }^{21}$
(3) Use OLS to estimate

$$
\operatorname{Score}_{i}^{*}=\beta_{0} \operatorname{Control}_{i}+\beta_{1} \operatorname{Group}_{i}+\beta_{2} \text { Individual }_{i}+\mathbf{X}_{i}^{\prime} \eta+\text { errors. }
$$

Store the estimates $\hat{\beta}_{n}^{*}=\left(\hat{\beta}_{n, 0}^{*}, \hat{\beta}_{n, 1}^{*}, \hat{\beta}_{n, 2}^{*}\right)^{\prime}$ and their corresponding standard errors.
(4) Repeat Steps 2-3 $B$ times, where $B$ is a large number (we set $B=9,999$ in this example).
(5) Find the smallest value of $\gamma$ satisfying (2.5) of the main paper. This can be accomplished using the following algorithm:
( $\ddagger$ ick a candidate value for $\gamma$.
(W)or each $b \in\{1, \ldots, B\}$, calculate the maximum lower endpoint

$$
\bar{L}^{* b}=\max _{s \in K}\left\{\left(\hat{\beta}_{n, s}^{* b}-\hat{\beta}_{n, s}\right)-\gamma \times \operatorname{se}\left(\hat{\beta}_{n, s}^{* b}\right)\right\}
$$

and the minimum upper endpoint

$$
\underline{U}^{* b}=\min _{s \in K}\left\{\left(\hat{\beta}_{n, s}^{* b}-\hat{\beta}_{n, s}\right)+\gamma \times \operatorname{se}\left(\hat{\beta}_{n, s}^{* b}\right)\right\},
$$

where $\hat{\beta}_{n}^{* b}$ is the $b^{\text {th }}$ replicate of $\hat{\beta}_{n}^{*}$.
(Walculate $\hat{\alpha}=\#\left(\bar{L}^{* b}>\underline{U}^{* b}\right) / B$.
(dterate over Steps $5 \mathrm{~b}-5 \mathrm{c}$, using a root finder to solve $(\alpha-\hat{\alpha})-I(\alpha<\hat{\alpha})=0 .{ }^{22}$
Figure S 1 displays $\hat{\alpha}$ as a function of $\gamma$ for this example (the horizontal dotted lines here correspond to $\alpha$ $=0.05$ ). Notice that $\hat{\alpha}$ is weakly decreasing in $\gamma .{ }^{23}$ Using the algorithm in Step 5 above, we obtain a value

[^1]

Figure S4. Frequency of observing at least one pair of non-overlapping uncertainty intervals across bootstrap replications ( $\hat{\alpha}$ ) as a function of $\gamma$ in the performance pay example.
of 0.497216 for $\gamma$ (at this value of $\gamma$, we have $\hat{\alpha}=499 / 9999=0.04990499<\alpha$; with $\gamma=0.497215$, we have $\hat{\alpha}=5000 / 9999=0.050005>\alpha$ ). With $B$ suitably large, $\hat{\alpha}$ can be made as close to $\alpha$ as one desires.

## S4. Test Statistics for Charitable Giving Example

Table S2 displays $T$-statistics of the form (S.6) for each of the $\binom{36+1}{2}=666$ relevant pairwise parameter comparisons for the example in Section 3.1 of the main paper. The $17 T$-statistics that are greater than 1.960 in absolute value are in bold.
based on the non-overlap of confidence intervals in this example would be conservative (in cases where $k$ is larger, it is likely that the opposite would be true).
Table S2. $T$-statistics in charitable giving example.

|  | 1:1/25/1 | 1:1/50/1 | 1:1/100/1 | 1:1NM/1 | 2:1/25/1 | 2:1/50/1 | 2:1/100/1 | 2:1/NM/1 | 3:1/25/1 | 3:1/50/1 | 3:1/100/1 | 3:1/nM/1 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Control | $-1.302$ | 0.090 | -0.127 | 0.404 | 1.561 | $-1.023$ | -1.806 | 0.822 | 1.374 | 0.041 | 1.413 | -0.045 |
| 1:1/25/1 |  | 0.705 | 0.697 | 1.101 | 1.997 | 0.041 | -0.303 | 1.385 | 1.859 | 0.871 | 1.923 | 0.652 |
| 1:1/50/1 |  |  | -0.151 | 0.185 | 1.155 | -0.643 | -0.903 | 0.519 | 0.954 | -0.053 | 0.926 | -0.099 |
| 1:1/100/1 |  |  |  | 0.398 | 1.428 | -0.606 | -0.959 | 0.755 | 1.234 | 0.125 | 1.234 | 0.047 |
| 1:1/NM/1 |  |  |  |  | 1.095 | -0.992 | -1.353 | 0.391 | 0.875 | -0.290 | 0.845 | -0.312 |
| 2:125/1 |  |  |  |  |  | -1.896 | -2.179 | -0.701 | -0.259 | -1.362 | -0.345 | -1.304 |
| 2:1150/1 |  |  |  |  |  |  | -0.309 | 1.284 | 1.747 | 0.761 | 1.793 | 0.583 |
| 2:1/100/1 |  |  |  |  |  |  |  | 1.598 | 2.059 | 1.150 | 2.153 | 0.873 |
| 2:1/NM/1 |  |  |  |  |  |  |  |  | 0.463 | -0.668 | 0.407 | -0.654 |
| 3:125/1 |  |  |  |  |  |  |  |  |  | -1.160 | $-0.080$ | -1.108 |
| 3:1150/1 |  |  |  |  |  |  |  |  |  |  | 1.155 | -0.062 |
| 3:1/100/1 |  |  |  |  |  |  |  |  |  |  |  | -1.091 |
|  | 1:1/25/1.25 | 1:1/501/.25 | 1:1/1001.25 | 1:11NM/1.25 | 2:1/25/1.25 | 2:1/501.1.25 | 2:1/100/1.25 | 2:1/NM/1.25 | 3:1/25/1.25 | 3:1/501/.25 | 3:11/00/1.25 | 3:1/NM/1.25 |
| Control | -0.234 | 0.605 | 1.647 | 0.975 | 0.017 | 1.131 | 0.310 | 1.411 | 0.452 | ${ }^{0.272}$ | -1.342 | 0.557 |
| 1:1/25/1 | 0.413 | 1.251 | 2.133 | 1.361 | 0.787 | 1.559 | 1.067 | 1.930 | 1.152 | 1.015 | -0.135 | 1.170 |
| 1:1/50/1 | -0.233 | 0.325 | 1.105 | 0.778 | -0.062 | 0.859 | 0.111 | 0.914 | 0.211 | 0.093 | -0.776 | 0.316 |
| 1:1/100/1 | -0.118 | 0.557 | 1.435 | 0.948 | 0.104 | 1.071 | 0.321 | 1.226 | 0.431 | 0.294 | -0.781 | 0.531 |
| 1:1/NM/1 | -0.445 | 0.163 | 1.046 | 0.691 | -0.288 | 0.773 | -0.094 | 0.832 | 0.027 | -0.111 | -1.167 | 0.159 |
| 2:125/1 | $-1.368$ | -0.950 | -0.171 | -0.182 | $-1.330$ | -0.211 | -1.198 | -0.368 | $-1.082$ | -1.200 | -2.034 | -0.915 |
| 2:150/1 | 0.364 | 1.140 | 1.992 | 1.308 | 0.696 | 1.488 | 0.948 | 1.793 | 1.036 | 0.906 | -0.161 | 1.074 |
| 2:1/100/1 | 0.607 | 1.496 | 2.358 | 1.494 | 1.040 | 1.722 | 1.339 | 2.162 | 1.409 | 1.279 | 0.143 | 1.393 |
| 2:1/NM/1 | -0.756 | -0.237 | 0.595 | 0.392 | -0.652 | 0.428 | -0.489 | 0.390 | -0.370 | -0.499 | $-1.438$ | -0.223 |
| 3:125/1 | -1.182 | -0.720 | 0.105 | 0.330 | -1.129 | 0.022 | -0.984 | -0.101 | -0.861 | -0.986 | $-1.896$ | -0.689 |
| 3:1150/1 | -0.222 | 0.457 | 1.361 | 0.887 | -0.015 | 1.003 | 0.205 | 1.146 | 0.324 | 0.180 | -0.949 | 0.435 |
| 3:1/100/1 | -1.168 | -0.679 | 0.195 | 0.093 | -1.119 | 0.093 | -0.963 | -0.022 | -0.829 | -0.966 | -1.965 | -0.646 |
| 3:1/NM/1 | -0.149 | 0.459 | 1.280 | 0.879 | 0.047 | 0.980 | 0.238 | 1.081 | 0.342 | 0.216 | -0.729 | 0.443 |
| 1:1/25/1.25 |  | 0.580 | 1.343 | 0.956 | 0.201 | 1.057 | 0.381 | 1.158 | 0.474 | 0.359 | -0.491 | 0.561 |
| 1:11501.25 |  |  | 0.878 | 0.580 | -0.447 | 0.642 | -0.262 | 0.664 | -0.139 | -0.276 | -1.312 | 0.004 |
| 1:1/1001.25 |  |  |  | -0.052 | -1.318 | -0.069 | -1.169 | -0.219 | -1.031 | -1.170 | -2.165 | -0.837 |
| 1:1/1/2M/1.25 |  |  |  |  | -0.877 | -0.009 | -0.762 | -0.109 | -0.678 | -0.769 | -1.402 | -0.563 |
| 2:1/25/1.25 |  |  |  |  |  | 0.986 | 0.207 | 1.109 | 0.320 | 0.184 | ${ }^{-0.866}$ | 0.428 |
| 2:11501. 25 |  |  |  |  |  |  | $-0.857$ | -0.111 | -0.757 | -0.863 | -1.605 | -0.621 |
| 2:1/100/1.25 |  |  |  |  |  |  |  | 0.952 | 0.124 | -0.020 | -1.136 | 0.252 |
| 2:1/NM/1.25 |  |  |  |  |  |  |  |  | -0.816 | -0.955 | -1.968 | ${ }^{0.631}$ |
| 3:1/251.25 |  |  |  |  |  |  |  |  |  | $-0.141$ | -1.216 | 0.136 |
| 3:11501/.25 |  |  |  |  |  |  |  |  |  |  | $-1.086$ | 0.266 |
| 3:1/11001.25 |  |  |  |  |  |  |  |  |  |  |  | 1.231 |
|  | 1:1/251. 5 | 1:1/501.5 | 1:1/100/1.5 | 1:1/NM/1.5 | 2:1/25/1.5 | 2:1/501.5 | 2:1/100/1.5 | 2:1/NM/1.5 | 3:1/251. 5 | 3:1/501. 5 | 3:1/1001.5 | 3:1/NM/1.5 |

Table S2. Continued

| Control | 0.434 | -0.129 | -0.183 | 0.628 | 1.520 | 0.810 | 0.940 | 0.020 | 1.633 | -0.477 | -0.433 | 0.469 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1:1/25/1 | 1.059 | 0.686 | 0.573 | 1.209 | 1.904 | 1.456 | 1.528 | 0.744 | 2.084 | 0.508 | 0.533 | 1.198 |
| 1:1/50/1 | 0.231 | -0.152 | -0.189 | 0.381 | 1.196 | 0.443 | 0.562 | -0.058 | 1.171 | -0.336 | -0.315 | 0.207 |
| 1:1/100/1 | 0.431 | -0.002 | -0.055 | 0.595 | 1.427 | 0.708 | 0.829 | 0.102 | 1.469 | -0.223 | -0.196 | 0.437 |
| 1:1/NM/1 | 0.064 | -0.397 | -0.425 | 0.238 | 1.131 | 0.296 | 0.435 | -0.273 | 1.114 | -0.635 | -0.607 | 0.017 |
| 2:1/25/1 | -0.991 | -1.425 | - 1.415 | -0.822 | 0.135 | -0.866 | -0.716 | - 1.293 | -0.034 | -1.637 | - 1.616 | -1.112 |
| 2:1/50/1 | 0.969 | 0.599 | 0.502 | 1.117 | 1.834 | 1.320 | 1.406 | 0.662 | 1.971 | 0.422 | 0.446 | 1.070 |
| 2:1/100/1 | 1.278 | 0.946 | 0.807 | 1.414 | 2.065 | 1.713 | 1.767 | 0.978 | 2.282 | 0.798 | 0.820 | 1.475 |
| 2:1/NM/1 | -0.308 | -0.754 | -0.764 | -0.138 | 0.777 | -0.128 | 0.017 | -0.627 | 0.695 | -0.977 | -0.952 | -0.390 |
| 3:1/25/1 | -0.771 | -1.230 | -1.221 | -0.595 | 0.379 | -0.626 | -0.470 | - 1.092 | 0.234 | -1.463 | - 1.438 | -0.890 |
| 3:1/50/1 | 0.333 | -0.127 | -0.172 | 0.505 | 1.367 | 0.609 | 0.738 | -0.011 | 1.399 | -0.367 | -0.338 | 0.326 |
| 3:1/100/1 | -0.733 | -1.229 | -1.218 | -0.546 | 0.458 | -0.579 | -0.413 | -1.076 | 0.324 | -1.486 | -1.459 | -0.862 |
| 3:1/NM/1 | 0.351 | -0.049 | -0.094 | 0.507 | 1.325 | 0.592 | 0.711 | 0.048 | 1.331 | -0.246 | -0.222 | 0.342 |
| 1:1/25/1.25 | 0.476 | 0.115 | 0.066 | 0.618 | 1.389 | 0.704 | 0.814 | 0.197 | 1.394 | -0.054 | -0.034 | 0.477 |
| 1:1/50/1.25 | -0.089 | -0.556 | -0.575 | 0.087 | 1.001 | 0.124 | 0.270 | -0.426 | 0.959 | -0.796 | -0.769 | -0.154 |
| 1:1/100/1.25 | -0.921 | -1.430 | - 1.411 | -0.731 | 0.301 | -0.783 | -0.610 | -1.270 | 0.141 | - 1.692 | - 1.665 | -1.069 |
| 1:1/NM/1.25 | -0.627 | -0.948 | -0.957 | -0.494 | 0.290 | -0.503 | -0.392 | -0.859 | 0.158 | - 1.095 | - 1.082 | -0.693 |
| 2:1/25/1.25 | 0.331 | -0.105 | -0.150 | 0.496 | 1.345 | 0.591 | 0.716 | 0.003 | 1.362 | -0.327 | -0.300 | 0.321 |
| 2:1/50/1.25 | -0.691 | -1.068 | -1.074 | $-0.543$ | 0.323 | -0.560 | -0.429 | -0.962 | 0.186 | -1.247 | -1.227 | -0.778 |
| 2:1/100/1.25 | 0.153 | -0.320 | -0.353 | 0.330 | 1.223 | 0.405 | 0.543 | -0.194 | 1.226 | -0.569 | -0.540 | 0.118 |
| 2:1/NM/1.25 | -0.718 | -1.221 | - 1.210 | -0.530 | 0.480 | -0.562 | -0.395 | -1.066 | 0.347 | -1.483 | -1.454 | -0.849 |
| 3:1/25/1.25 | 0.040 | -0.431 | -0.457 | 0.216 | 1.123 | 0.272 | 0.414 | -0.303 | 1.101 | -0.676 | -0.647 | -0.011 |
| 3:1/50/1.25 | 0.168 | -0.295 | -0.328 | 0.341 | 1.221 | 0.416 | 0.552 | -0.173 | 1.226 | -0.535 | -0.507 | 0.135 |
| 3:1/100/1.25 | 1.121 | 0.771 | 0.656 | 1.262 | 1.943 | 1.506 | 1.579 | 0.820 | 2.121 | 0.607 | 0.630 | 1.263 |
| 3:1/NM/1.25 | -0.089 | -0.529 | -0.550 | 0.080 | 0.975 | 0.114 | 0.253 | -0.409 | 0.921 | -0.752 | -0.726 | -0.150 |
| 1:1/25/1.5 |  | -0.430 | -0.455 | 0.165 | 1.042 | 0.208 | 0.344 | -0.316 | 1.000 | -0.645 | -0.621 | -0.051 |
| 1:1/50/1.5 |  |  | -0.052 | 0.595 | 1.417 | 0.705 | 0.827 | 0.103 | 1.464 | -0.219 | -0.192 | 0.436 |
| 1:1/100/1.5 |  |  |  | 0.610 | 1.419 | 0.716 | 0.832 | 0.147 | 1.449 | -0.147 | -0.123 | 0.462 |
| 1:1/NM/1.5 |  |  |  |  | 0.888 | 0.024 | 0.163 | -0.476 | 0.823 | -0.808 | -0.786 | -0.231 |
| 2:1/25/1.5 |  |  |  |  |  | -0.931 | -0.792 | -1.314 | -0.170 | -1.610 | -1.586 | -1.145 |
| 2:1/50/1.5 |  |  |  |  |  |  | 0.155 | -0.562 | 0.872 | -0.969 | -0.938 | -0.293 |
| 2:1/100/1.5 |  |  |  |  |  |  |  | -0.686 | 0.712 | -1.076 | - 1.048 | -0.437 |
| 2:1/NM/1.5 |  |  |  |  |  |  |  |  | 1.321 | -0.312 | -0.287 | 0.303 |
| 3:1/25/1.5 |  |  |  |  |  |  |  |  |  | -1.698 | - 1.672 | -1.135 |
| 3:1/50/1.5 |  |  |  |  |  |  |  |  |  |  | 0.028 | 0.695 |
| 3:1/100/1.5 |  |  |  |  |  |  |  |  |  |  |  | 0.665 |

## REFERENCES- APPENDIX

Bennett, C. J., B. S. Thompson (2016). Graphical procedures for multiple comparisons under general dependence. Journal of the American Statistical Association 111, 1278-88.

Davidson, R., E. Flachaire (2008). The wild bootstrap, tamed at last. Journal of Econometrics 146, 162-9.
Liu, R. Y. (1988). Bootstrap procedures under some Non-I.I.D models. Annals of Statistics 16, 1696-708.
MacKinnon, J. G., H. White (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. Journal of Econometrics 29, 305-25.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. Annals of Statistics 21, 255-85.

Romano, J. P., M. Wolf (2005a). Exact and approximate stepdown methods for multiple hypothesis testing. Journal of the American Statistical Association 100, 94-108.

Romano, J. P., M. Wolf (2005b). Stepwise multiple testing as formalized data snooping. Econometrica 73, 1237-82.

Webb, M. D. (2014). Reworking wild bootstrap based inference for clustered errors. Working Papers 1315, Queen's University.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. Annals of Statistics 14, 1261-95.

Young, A. (2018). Channelling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. Forthcoming in Quarterly Journal of Economics.


[^0]:    ${ }^{19}$ As noted in the main paper, Young (2019) jointly tests the hypothesis that all of the treatment effects are zero.
    ${ }^{20}$ Romano and Wolf (2005a) propose a stepwise refinement that is applicable to this max- $T$ procedure. As in BT, we compare the unrefined max- $T$ procedure to the unrefined overlap procedure in order to reduce the computational cost of our experiments.

[^1]:    ${ }^{21}$ See also Webb () if $G \leq 12$.
    ${ }^{22}$ The basic idea is to minimize the distance $|\alpha-\hat{\alpha}|$ while ensuring that $\alpha \geq \hat{\alpha}$ (notice that the indicator function acts as a penalty term here). If $\alpha<\hat{\alpha}$ (e.g., points to the left of $\gamma=0.497216$ in Figure S4b), then $(\alpha-\hat{\alpha})-I(\alpha<\hat{\alpha})<-1$; on the other hand, if $\alpha \geq \hat{\alpha}$ (e.g., points at or to the right of $\gamma=0.497216$ in Figure S4b), then $0 \leq(\alpha-\hat{\alpha})-I(\alpha<\hat{\alpha})<1$. Thus, $(\alpha-\hat{\alpha})-I(\alpha<\hat{\alpha})$ is guaranteed to be closer to zero (in absolute value) if $\alpha \geq \hat{\alpha}$ than if $\alpha<\hat{\alpha}$.
    ${ }^{23}$ It is interesting to note that, with $\gamma=1.960$ (which, under mild regularity conditions, would result in the uncertainty intervals being asymptotically-valid 0.95 -level confidence intervals), we have $\hat{\alpha}=0$. This suggests that making inferences

