A Simple, Graphical Approach to Comparing Multiple Treatments

Brennan S. Thompson† and Matthew D. Webb‡

First version: February 1, 2015
This version: March 15, 2017

Abstract

We propose a graphical approach to comparing multiple treatments that allows users to easily infer differences between any treatment effect and zero, and between any pair of treatment effects. Our approach makes use of a flexible, resampling-based procedure that asymptotically controls the familywise error rate (the probability of making one or more spurious inferences). We demonstrate the usefulness of our approach with three empirical examples.

Keywords: multiple hypothesis testing; treatment effects; bootstrap
1 Introduction

When an experiment involves more than one treatment (e.g., several different drugs designed to treat a particular disease), there is often interest in comparing each treatment to not only a control, but also to the other treatment(s). With \( k \) treatments under consideration, this can be seen to involve testing a total of \( \binom{k+1}{2} \) hypotheses. For example, with \( k = 2 \) treatments, there are 3 hypotheses of interest: (i) that the effect of the first treatment is equal to zero; (ii) that the effect of the second treatment is equal to zero; and (iii) that the effects of the first and second treatments are equal to each other. With \( k = 3 \) treatments, there are 6 hypotheses of interest, and so on.

Of course, when testing more than one hypothesis at a given nominal level, the probability of rejecting at least one true hypothesis, i.e., the familywise error rate (FWER), is typically well in excess of that given nominal level. In recognition of this issue, a wide variety of multiple testing procedures, ranging from the simple Bonferroni correction to resampling-based stepwise procedures (Romano & Wolf, 2005a,b), have been developed to control the FWER. While such procedures are often used when multiple treatments are examined in biostatistics (Dunnett, 1955; Dunnett & Tamhane, 1991), the econometrics literature has, with the exception of the recent working paper by List, Shaikh & Xu (2016), hereafter LSX, ignored the problem of multiple testing whenever multiple treatments are considered.

---

1See, e.g., Angrist, Lang & Oreopoulos (2009), Karlan & List (2007), and Muralidharan & Sundararaman (2011), all of which we revisit in this paper.

2In general, the FWER is bounded from above by \( m\alpha_0 \), where \( m \) is the number of hypotheses under consideration (here, \( m = \binom{k+1}{2} \)), and \( \alpha_0 \) is the nominal level that each hypothesis is tested at. More specifically, the FWER is equal to \( \alpha_m \equiv 1 - (1 - \alpha_0)^m < m\alpha_0 \) if the tests are mutually independent. However, if the tests are mutually dependent, as is the case here, the FWER may be greater or less than \( \alpha_m \).

3Multiple testing procedures have also been developed to control other generalized error rates, such as the false discovery rate (the expected proportion of true hypotheses rejected; Benjamini & Hochberg, 1995). If one desired, it would be straightforward to modify the procedures used in this paper to control the false discovery rate rather than the FWER.

4Recently, some researchers have used multiple-testing procedures when examining heterogeneous treatment effects, in which different types of individuals (say, men and women) may respond differently to the same treatment; see Anderson (2008), Fink et al. (2014), Lee & Shaikh (2014), Lehrer, Pohl & Song (2015), and Gu & Shen (2016). Young (2016), on the other hand, jointly tests the (single) hypothesis that all of the treatment effects – which may differ not only across different treatments, but also across different types of individuals – are zero.
Nonetheless, as LSX note, this issue is pervasive in many areas of economics as multiple treatments are considered “in nearly every experiment that is published today” (p. 3).

In this paper, we propose a graphical approach to comparing multiple treatments that uses the resampling-based procedure of Bennett & Thompson (2016), hereafter BT, to asymptotically control the FWER. The advantage of this graphical approach is that it allows users to easily determine both statistical and practical significance in the differences between each treatment effect and zero, and between each pair of treatment effects. That is, unlike standard multiple testing procedures, such as that utilized by LSX, it offers users more than a “Yes-No” decision on the \( \binom{k+1}{2} \) hypotheses of interest.

In the following section, we provide a formal description of our problem and present the results of a simple Monte Carlo simulation that illustrates the perils of ignoring the issue of multiple testing in this setting. Next, in Section 3, we review the graphical procedure of BT as well as a modification of this procedure designed to identify the “best” treatment under consideration. Section 3 also includes a simple illustration using data from a field experiment in which \( k = 2 \) types of performance pay for teachers are considered. Section 4 describes the results of a set of additional Monte Carlo simulations designed to examine the finite-sample performance of the procedures. In Section 5, we present the results of two additional empirical examples. The first of these examples is of interest as it involves a large number of treatments (\( k = 36 \)). In the second, we consider a case where treatment effects are estimated with instrumental variables. In both examples, controlling for multiple testing meaningfully changes the statistical inferences. Finally, Section 6 concludes.

2 Setup

In order to make the discussion of our problem more concrete, consider the following regression model:

\[
Y_i = \beta_0 + \sum_{s=1}^{k} \delta_s D_{s,i} + X_i' \eta + V_i,
\]  

(1)
where $D_{s,i}$ equals one if individual $i \in \{1, \ldots, n\}$ participates in treatment $s \in \{1, \ldots, k\}$ and zero otherwise; $X_i$ is a vector of other explanatory variables (e.g., age, gender, etc.); and $V_i$ is an idiosyncratic error term.\footnote{In cases where selection issues are a concern, one might, for example, treat participation in a treatment as endogenous and use assignment to that treatment as an instrument (see Section 5.2 for an example). The crucial assumption we make is that the parameters of interest can be $\sqrt{n}$-consistently estimated, whether using OLS, 2SLS, or some other method.} We assume that each individual receives only one of the $k$ treatments or is in a control group; if there are individuals receiving a combination of treatments, such individuals would be included in a distinct treatment group (see Section 5.2 for an example). In what follows, we define the treatment effect of the $s^{th}$ treatment as $\delta_s$.

The first part of our problem involves comparing each treatment to the control, i.e., testing the following $k$ hypotheses:

\begin{equation}
\delta_s = 0, \quad \text{for each } s \in \{1, \ldots, k\}.
\end{equation}

The second part of our problem involves comparing each treatment to the other treatment(s), i.e., testing the following $\binom{k}{2}$ hypotheses:

\begin{equation}
\delta_s = \delta_t, \quad \text{for each unique } (s, t) \in \{1, \ldots, k\}^2.
\end{equation}

Hence, as noted in the previous section, our problem involves testing a total of $\binom{k+1}{2}$ hypotheses.\footnote{In Appendix A, we consider the $k = 1$ case. Although there is only a single hypothesis of interest ($\delta_1 = 0$) in this case, it does provide some important insight on our proposed approach. In Appendix B, we simplify our problem by ignoring the hypotheses in (3), i.e., we focus on the so-called problem of “multiple comparisons with a control” (Hsu, 1996) that was first explored by Dunnett (1955). Interestingly, in the work on testing for heterogenous treatment effects cited above, each treatment effect is compared only to zero (and not any of the other treatment effects).}

To illustrate the perils of ignoring the issue of multiple testing in this setting, we generate, for $k \in \{2, \ldots, 10\}$, one million samples from the model (1) 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 $X_i'\eta$ 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), and (B) each of the $\binom{k}{2}$ hypotheses in (3). 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) and (3) are true as well. However, rejection of this hypothesis provides no guidance as to which of the hypotheses in (2) or (3) 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).\footnote{As noted above, Young (2016) jointly tests the hypothesis that all of the treatment effects are zero.}

The rejection frequencies for the different types of tests are shown in Figure 1. Specifically, the dash-dotted line shows the frequency of rejecting at least one of the $k$ hypotheses in (2), the dashed line shows the frequency of rejecting at least one of the $\binom{k}{2}$ hypotheses in (3), 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.

3 The Overlap Procedure

The procedure of BT is designed to facilitate all pairwise comparisons within a set of parameters, so we re-write model (1) as follows:

$$Y_i = \sum_{s=0}^{k} \beta_s D_{s,i} + X_i' \eta + V_i,$$

where $D_{0,i}$ equals one if individual $i$ belongs to the control group and zero otherwise (so that $\sum_{s=0}^{k} D_{s,i} = 1$ for all $i$); and $\beta_s \equiv \beta_0 + \delta_s$, for $s \in \{1, \ldots, k\}$.\footnote{Note the absence of a constant term in this model.} In what follows, we denote the parameter vector $(\beta_0, \ldots, \beta_k)'$ by $\beta$.

Notice that, since $\delta_s = 0$ is equivalent to $\beta_s = \beta_0$, and $\delta_s = \delta_t$ is equivalent to $\beta_s = \beta_t$, our problem boils down to testing the following $\binom{k+1}{2}$ hypotheses:

$$\beta_s = \beta_t, \quad \text{for each unique (s, t) } \in \mathcal{K}^2,$$

where $\mathcal{K} = \{0, \ldots, k\}$. However, since interest ultimately lies in the treatment effects, $\delta_1, \ldots, \delta_k$, we show how users can make inferences about the hypotheses in (2) and (3) more directly later in this section.

The procedure of BT, which can be seen as a resampling-based generalization of of Tukey's (1953) procedure, involves presenting each of the parameter estimates $\hat{\beta}_{n,s}$, $s \in K$, together
with a corresponding uncertainty interval,

\[ C_{n,s}(\gamma) = \left[ \hat{\beta}_{n,s} \pm \gamma \times \text{se} \left( \hat{\beta}_{n,s} \right) \right], \]

whose length is determined by the parameter \( \gamma > 0 \) (discussed below) and \( \text{se} \left( \hat{\beta}_{n,s} \right) \), the standard error of \( \hat{\beta}_{n,s} \). High-level assumptions on the large-sample behaviour of these objects are given at the end of this section. We denote the lower and upper endpoints of \( C_{n,s}(\gamma) \) by \( L_{n,s}(\gamma) \) and \( U_{n,s}(\gamma) \), respectively.

These uncertainty intervals are used to make inferences about the ordering of the parameters of interest as follows. We infer that \( \beta_s > \beta_t \) if the uncertainty interval for \( \hat{\beta}_s \) lies entirely above the uncertainty interval for \( \hat{\beta}_t \) (i.e., if \( L_{n,s} > U_{n,t} \)). If the uncertainty intervals for \( \hat{\beta}_s \) and \( \hat{\beta}_t \) overlap one another (i.e., if \( C_{n,s} \cap C_{n,t} \neq \emptyset \)), we can make no such inference.\(^9\)

For this reason, BT refer to their procedure as the overlap procedure.

The choice of \( \gamma \) is motivated as follows. If all \( k+1 \) parameters are equal, then the “ideal” choice of \( \gamma \) would ensure that the probability that at least one pair of uncertainty intervals is non-overlapping is as close to, but no higher than, the nominal FWER \( \alpha \). That is, the “ideal” choice of \( \gamma \) is the smallest value satisfying

\[ P \left( \max_{s \in K} L_{n,s}(\gamma) > \min_{s \in K} U_{n,s}(\gamma) \right) \leq \alpha, \]

when all \( k+1 \) parameters are equal (notice that the probability above is weakly decreasing in \( \gamma \); values of \( \gamma \) larger than the ideal value – but still satisfying the above condition – will result in a FWER that is weakly further below \( \alpha \)). Since this choice is infeasible, we choose \( \gamma \) using the bootstrap analogue of the above.\(^{10}\) Towards this end, for \( b \in \{1, \ldots, B\} \), let \( \hat{\beta}_{n,s}^{*b} \) be the \( b \)th replicate of \( \hat{\beta}_{n,s}^{*} \), the bootstrap counterpart of \( \hat{\beta}_{n,s} \). Then, a feasible choice of \( \gamma \) is

\(^9\)It should be pointed out that using confidence intervals to make inferences in this manner would be inappropriate. Specifically, when \( k = 1 \), such inferences would be overly conservative (cf. Appendix A); as \( k \) grows, the FWER would quickly become larger than one minus the nominal level of the confidence intervals.

\(^{10}\)In Appendix A, we show that, when \( k = 1 \) and the limiting distribution of \( \sqrt{n}(\hat{\beta}_n - \beta) \) is known, we can easily choose \( \gamma \) without resorting to the bootstrap.
the smallest value satisfying

\[
\frac{1}{B} \sum_{b=1}^{B} I \left( \max_{s \in K} L_{n,s}^b(\gamma) > \min_{s \in K} U_{n,s}^b(\gamma) \right) \leq \alpha, \tag{5}
\]

where \( I(\cdot) \) is an indicator function and \( L_{n,s}^b \) and \( U_{n,s}^b \) are, respectively, the lower and upper endpoints of

\[
C_{n,s}^b(\gamma) = \left[ (\hat{\beta}_{n,s}^b - \hat{\beta}_{n,0}) \pm \gamma \times se(\hat{\beta}_{n,s}^b) \right].
\]

We make only the following high-level assumptions:

- \( \sqrt{n}(\hat{\beta}_n - \beta) \) and \( \sqrt{n}(\hat{\beta}_n^* - \hat{\beta}_n) \) both have the same (continuous and strictly increasing) \((k + 1)\)-variate limiting distribution.
- For each \( s \in K \), \( \sqrt{n} \times se(\hat{\beta}_{n,s}) \) and \( \sqrt{n} \times se(\hat{\beta}_{n,s}^*) \) both converge in probability to the same (positive) constant.

Under these conditions, BT show that the overlap procedure described above (1) bounds the FWER from above by \( \alpha \) asymptotically, and (2) is consistent, in the sense that any differences between parameter pairs are inferred (in the correct direction) with probability one asymptotically. Simulation evidence presented in BT and in Section 4 below suggests that the overlap procedure provides satisfactory control of the FWER and has good (average) power properties in finite samples.

We conclude this section by showing how the overlap procedure can be used to allow users to make inferences about the treatment effects, \( \delta_1, \ldots, \delta_k \), more directly. First, we subtract \( \hat{\beta}_{n,0} \) from the endpoints of the uncertainty intervals for \( \beta_0, \ldots, \beta_k \) (leaving their lengths unchanged). That is, for each \( s \in K \), we compute the interval

\[
\tilde{C}_{n,s}(\gamma) = \left[ (\hat{\beta}_{n,s} - \hat{\beta}_{n,0}) \pm \gamma \times se(\hat{\beta}_{n,s}) \right].
\]
Note that, for $s = 0$, this interval is simply

$$
\tilde{C}_{n,0}(\gamma) = \left[ 0 \pm \gamma \times \text{se} \left( \hat{\beta}_{n,0} \right) \right],
$$

while, for $s \in \{1, \ldots, k\}$, this interval is

$$
\tilde{C}_{n,s}(\gamma) = \left[ \hat{\delta}_{n,s} \pm \gamma \times \text{se} \left( \hat{\beta}_{n,s} \right) \right],
$$

where $\hat{\delta}_{n,s} \equiv \hat{\beta}_{n,s} - \hat{\beta}_{n,0}$. It is important to emphasize that (6) is not an uncertainty interval for $\delta_s$ (cf. Appendix B).

Denoting the lower and upper endpoints of $\tilde{C}_{n,s}(\gamma)$ by $\tilde{L}_{n,s}(\gamma)$ and $\tilde{U}_{n,s}(\gamma)$, respectively, we can then infer that $\delta_s > 0$ if $\tilde{L}_{n,s} > \tilde{U}_{n,0}$, that $\delta_s < 0$ if $\tilde{U}_{n,s} < \tilde{L}_{n,0}$, and that $\delta_s > \delta_t$ if $\tilde{L}_{n,s} > \tilde{U}_{n,t}$.

### 3.1 Stepwise Refinement

BT also propose an iterative stepwise refinement for the overlap procedure that (weakly) increases its power without sacrificing asymptotic control of the FWER. The idea behind this refinement is to iterate the overlap procedure while eliminating any pairwise parameter comparisons that are “resolved” at a previous step. In this sense, it is analogous to Holm’s (1979) stepwise refinement of the Bonferroni correction, which eliminates from consideration any null hypotheses that are rejected at a previous step.

We begin by defining, for each $s \in K$,

$$
A_{n,s}(\gamma) = \{ t \in K : C_{n,s}(\gamma) \cap C_{n,t}(\gamma) \neq \emptyset \},
$$

so that $t \in A_{n,s}$ whenever the uncertainty intervals for $\beta_s$ and $\beta_t$ overlap one another (i.e., whenever the pairwise comparison between $\beta_s$ and $\beta_t$ is “unresolved”). Note that, if $A_{n,s} = \{s\}$ for all $s \in K$, all pairwise comparisons have been resolved.
Next, let $\gamma_{n,1}$ be the value of $\gamma$ obtained via the basic (i.e., unrefined) overlap procedure. Then, for $j \in \{2, \ldots, k\}$, the $j^{th}$ iteration of the procedure involves choosing $\gamma$ as the smallest value satisfying

$$\frac{1}{B} \sum_{b=1}^{B} I \left( \max_{s \in K} \left\{ \max_{t \in A_{n,s}(\gamma_{n,j-1})} L_{n,t}^{ub}(\gamma) - \min_{t \in A_{n,s}(\gamma_{n,j-1})} U_{n,t}^{ub}(\gamma) \right\} > 0 \right) \leq \alpha.$$ 

Notice that, here, we are only concerned with the non-overlap of (re-centered) uncertainty intervals which correspond to comparisons that were unresolved at the previous iteration.

Of course, if $A_{n,s}(\gamma_{n,j}) = A_{n,s}(\gamma_{n,j-1})$ for all $s \in K$, or if $A_{n,s}(\gamma_{n,j}) = \{s\}$ for all $s \in K$, then no further refinement is possible (i.e., no further pairwise comparisons can possibly be resolved), and the iterations are halted. More generally, $A_{n,s}(\gamma_{n,j}) \subseteq A_{n,s}(\gamma_{n,j-1})$ for all $s \in K$, meaning that $\gamma_{n,j} \leq \gamma_{n,j-1}$. Thus, the stepwise refinement can resolve at least as many pairwise comparisons as the basic procedure. Moreover, BT show that, so long as at least one pair of parameters is equal, the stepwise refinement results in an FWER exactly equal to $\alpha$ asymptotically; this is true of the basic procedure only if all of the parameters are equal.

### 3.2 Empirical Example

In order to provide a simple illustration of our proposed approach, we utilize data from Muralidharan & Sundararaman (2011), hereafter MS (Section 5 presents the results of two additional empirical examples). This paper describes the results of a field experiment designed to examine the effects of offering teachers performance pay conditional upon students’ academic performance.\footnote{The data used in this example is publicly-available at: \url{http://www.jstor.org/stable/suppl/10.1086/659655/suppl_file/2009434data.zip}} Specifically, MS analyze outcomes from three separate groups of schools: a control group, a group in which teachers were paid based on the scores of their own students, and a group in which teachers were paid based on the performance of all students at their school. In other words, there are $k = 2$ treatments.
Among many other things, MS compare the impact of the group incentive to the impact of the individual incentive on combined math and language scores over the two years that the experiment ran. To make these comparisons, they first estimate the following model:

$$\text{Score}_i = \beta_0 + \delta_1 \text{Group}_i + \delta_2 \text{Individual}_i + X_i' \eta + V_i,$$  

(7)

where Score is the combined math and language score in year 2; Group and Individual are indicator variables indicating membership in the group incentive treatment group and individual incentive treatment group, respectively; and $X$ contains the combined math and language score in year 0, as well as a set of indicator variables for subdistricts. There are $n = 29,760$ observations (approximately one-third of these observations correspond to the control group and to each of the two treatment groups), and the model is estimated using OLS. Standard errors are clustered by school. Results are shown in the first column of Table 1 (cf. the fourth column of Table 8 in MS).

Next, MS test the following three hypotheses:

MS1: $\delta_1 = 0$

MS2: $\delta_2 = 0$

MS3: $\delta_2 = \delta_1$

$T$-statistics corresponding to tests of these three hypotheses are 2.701, 4.879, and 1.897, respectively. Thus, MS conclude that both treatment effects are statistically different from zero, and from one another (even MS3 could be rejected in favour of a two-sided alternative at a nominal level of slightly less than 6% if the tests were conducted separately, i.e., without controlling the FWER).

In order to apply the overlap procedure, we first need to re-write the model above in the form of model (4), i.e.,

$$\text{Score}_i = \beta_0 \text{Control}_i + \beta_1 \text{Group}_i + \beta_2 \text{Individual}_i + X_i' \eta + V_i,$$  

(8)
Table 1: Performance pay example: Parameter estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model (7)</th>
<th>Model (8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>0.132</td>
<td>0.132</td>
</tr>
<tr>
<td></td>
<td>(0.168)</td>
<td>(0.168)</td>
</tr>
<tr>
<td>$\delta_1$</td>
<td>0.154</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.057)</td>
<td></td>
</tr>
<tr>
<td>$\delta_2$</td>
<td>0.283</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.058)</td>
<td></td>
</tr>
<tr>
<td>$\delta_2 - \delta_1$</td>
<td>0.129</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.068)</td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.286</td>
<td>0.286</td>
</tr>
<tr>
<td></td>
<td>(0.172)</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.415</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.168)</td>
<td></td>
</tr>
<tr>
<td>$\beta_2 - \beta_0$</td>
<td>0.283</td>
<td>0.283</td>
</tr>
<tr>
<td></td>
<td>(0.057)</td>
<td>(0.058)</td>
</tr>
<tr>
<td>$\beta_2 - \beta_1$</td>
<td>0.129</td>
<td>0.129</td>
</tr>
<tr>
<td></td>
<td>(0.068)</td>
<td>(0.068)</td>
</tr>
</tbody>
</table>

Note: Clustered standard errors are in brackets.
where Control is an indicator variable for membership in the control group. Notice that MS1 \( (\delta_1 = 0) \), MS2 \( (\delta_2 = 0) \), and MS3 \( (\delta_2 = \delta_1) \) are equivalent to \( \beta_1 = \beta_0 \), \( \beta_2 = \beta_0 \), and \( \beta_2 = \beta_1 \), respectively. Indeed, Table 1 shows that the estimates of \( \delta_1 \), \( \delta_2 \), and \( \delta_2 - \delta_1 \) arising from model (7) are identical to the estimates of \( \beta_1 - \beta_0 \), \( \beta_2 - \beta_0 \), and \( \beta_2 - \beta_1 \), respectively, arising from model (8).

Given a nominal FWER of \( \alpha = 0.05 \) and 9,999 replications of the wild cluster bootstrap (Cameron et al., 2008), we obtain a value of 0.497 for \( \gamma \) (this is the value obtained after the first iteration; no further refinement was possible).\(^{12}\) The resulting uncertainty intervals are shown in Figure 2a, and can be interpreted as follows:

- Since the uncertainty intervals for \( \beta_1 \) and \( \beta_0 \) overlap, we cannot infer anything about their ordering (or, equivalently, anything about the sign of \( \delta_1 \)).

- Since the uncertainty interval for \( \beta_2 \) lies entirely above the uncertainty interval for \( \beta_0 \), we infer that \( \beta_2 > \beta_0 \) (or, equivalently, that \( \delta_2 > 0 \)).

\(^{12}\)Supplementary Appendix I provides computational details for this example.
• Since the uncertainty intervals for $\beta_2$ and $\beta_1$ overlap, we can we cannot infer anything about their ordering (or, equivalently, anything about the ordering of $\delta_2$ and $\delta_1$).

Thus, while our results are consistent with rejecting MS2, they are not consistent with rejecting either MS1 or MS3.\footnote{We also applied the overlap procedure at a nominal FWER of $\alpha = 0.06$ (obtaining a value of 0.469 for $\gamma$), and found that the uncertainty intervals for $\beta_2$ and $\beta_1$ were still overlapping (recall that the absolute value of the $T$-statistic for the test of MS3 was 1.897, which corresponds to a non-multiplicity-adjusted $p$-value of just under 0.06). In fact, the smallest nominal FWER at which the uncertainty intervals for $\beta_2$ and $\beta_1$ are non-overlapping is $\alpha = 0.149$ (see BT, Section 3.3, for a discussion of multiplicity-adjusted $p$-values).}

Figure 2b displays the same uncertainty intervals centered around the treatment effects, $\delta_1 \equiv \beta_1 - \beta_0$ and $\delta_2 \equiv \beta_2 - \beta_0$. That is, we subtract $\hat{\beta}_{n,0} = 0.132$ from the endpoints of the uncertainty intervals for $\beta_1$ and $\beta_2$ (leaving their lengths unchanged). Moreover, we include a dotted horizontal line at $\tilde{U}_{n,0}$ (if the vertical axis extended far enough below zero, we would include another dotted horizontal line at $\tilde{L}_{n,0}$). Given that $\tilde{L}_{n,2}$ lies above this dotted horizontal line, for example, one can quickly infer that $\delta_2 > 0$.

3.3 A Modification for Multiple Comparisons with the Best

Thus far, we have been concerned with controlling the FWER across all pairwise parameter comparisons. This approach allows for a (potentially complete) ranking of all the treatments under consideration. For example, assuming that a larger value of the outcome variable is “better”, one could infer that treatment $s \in \{1, \ldots, k\}$ is the “best”, i.e., $\beta_s > \beta_t$ for all $t \in K \setminus \{s\}$, if $L_{n,s} > U_{n,t}$ for all $t \in K \setminus \{s\}$.\footnote{Note that $\beta_s > \beta_t$ for all $t \in K \setminus \{s\}$ is equivalent to $\delta_s > 0$ and $\delta_s > \delta_t$ for all $t \in \{1, \ldots, k\} \setminus \{s\}$. That is, a treatment is declared the “best” if its treatment effect is both positive and larger than all of the $k - 1$ other treatment effects.} Similarly, one may be able to identify a “second best” treatment, a “third best” treatment, and so on.

While such a complete ranking may occasionally be of value, interest often centers on identifying only the (first) best treatment. That is, we may only want to know whether or not the treatment effect which is estimated to be the largest is actually statistically distinguishable from the other treatment effect(s) and from zero. Such a problem is the focus...
of so-called “multiple comparisons with the best” procedures (Hsu, 1981, 1984; Horrace and Schmidt, 2000).

Here, we follow BT in developing a modification of the overlap procedure to focus on this problem. The basic idea behind this modification is that, by eliminating “irrelevant” pairwise comparisons (i.e., those in which neither of the parameters is estimated to be largest), it may be possible to substantially increase the power of the procedure.

We begin by introducing some further notation. Let $[1], [2], \ldots, [k+1]$ be the random indices such that $\hat{\beta}_{n,[1]} > \hat{\beta}_{n,[2]} > \cdots > \hat{\beta}_{n,[k+1]}$. This means that $\beta_{[1]}$ is the true value of the parameter which is estimated to be largest, and not necessarily the largest parameter value. Moreover,

$$L_{n,[1]}(\gamma) = \hat{\beta}_{n,[1]} - \gamma \times se\left(\hat{\beta}_{n,[1]}\right)$$

is the lower endpoint of the uncertainty interval for $\beta_{[1]}$. Interestingly, $L_{n,[1]}$ may not be the largest lower endpoint; if the standard error of $\hat{\beta}_{n,[1]}$ is relatively large, it could be the case that the lower endpoint associated with this point estimate extends below the lower endpoint associated with some smaller point estimate.

Similar to what is done in the unmodified overlap procedure, we infer that $\beta_{[1]}$ is the largest parameter value in the collection if $L_{n,[1]} > U_{n,[s]}$ for all $s \in \{2, \ldots, k+1\}$. Thus, a feasible choice of $\gamma$ here is the smallest value satisfying

$$\frac{1}{B} \sum_{b=1}^{B} I\left(L_{n,[1^*]}(\gamma) > \max_{s^* \in \{2, \ldots, k+1\}} U_{n,[s^*]}(\gamma)\right) \leq \alpha,$$

where $[1^*], [2^*], \ldots, [(k+1)^*]$ are random indices such that

$$\left(\hat{\beta}_{n,[1^*]} - \hat{\beta}_{n,[1^*]}\right) > \left(\hat{\beta}_{n,[2^*]} - \hat{\beta}_{n,[2^*]}\right) > \cdots > \left(\hat{\beta}_{n,[(k+1)^*]} - \hat{\beta}_{n,[(k+1)^*]}\right)$$

15In fact, BT introduce a generalization of the “multiple comparisons with the best” approach which allows for comparisons within the “r best” ($r$ being some integer smaller than the total number of parameters under consideration). Such an approach may be of use when the number of parameters under consideration is very large (perhaps in the hundreds or thousands), and one is willing to abandon pursuit of a complete ranking in return for the ability to resolve more comparisons within the top $r$.

16Effectively, the number of comparisons is reduced from from $\binom{k+1}{2}$ to $k$. 15
That is, $L_{n,[1]}^b$ is the lower endpoint of the uncertainty interval for the parameter which is estimated to be largest in the $b$th bootstrap sample after re-centering.

Simulation evidence presented in BT and in the next section suggests that the choice of $\gamma$ resulting from this modification may be substantially smaller than the choice resulting from the unmodified overlap procedure, resulting in greatly increased power.

Before moving on, we illustrate the modified overlap procedure using the performance pay example introduced in Section 3.2. That is, we seek to determine only whether or not the treatment effect for the individual incentive (the treatment effect estimated to be the largest) is statistically distinguishable from the treatment effect for the group incentive and from zero.

Here, we obtain a value of 0.316 for $\gamma$, which is less than two-thirds as large as the value we obtained using the unmodified procedure (0.497). Figure 3 displays the lower half of $\tilde{C}_{n,[1]} = \tilde{C}_{n,2}$ and the upper halves of $\tilde{C}_{n,[2]} = \tilde{C}_{n,1}$ and $\tilde{C}_{n,[3]} = \tilde{C}_{n,0}$. We explicitly include the upper half of $\tilde{C}_{n,0}$ here (rather than a dotted horizontal line corresponding to its upper endpoint, as in Figure 2b) since the modified overlap procedure cannot be used to make inferences about the sign of any treatment effect that is not estimated to be largest (i.e., we cannot compare the group incentive to the control here).

Our inferences here are much more in line with MS. Specifically, we infer that $\delta_2 > 0$ and that $\delta_2 > \delta_1$, decisions which are consistent with rejecting MS2 and MS3. However, since the modified overlap procedure focuses solely on comparisons with the “best”, it does not allow us to infer anything about the sign of $\delta_1$. That is, we cannot say anything about MS1. The increased power of the modified overlap procedure comes entirely at the cost of remaining silent on such comparisons.

Ultimately, one must decide which procedure to use based on which comparisons are actually of interest: if identifying “second best”, “third best”, etc., (or even having the ability to infer whether or not any of the treatment effects that are not estimated to be largest are different from zero) is of no concern, the modified overlap procedure can be recommended...
on the grounds of potentially much higher power. Of course, this choice should be made 
a priori so as to avoid the temptation to “cherry pick” results. With this caveat in mind, we use only the unmodified procedure in the empirical example in Section 5.1, and only the modified procedure in the empirical example in Section 5.2.

4 Simulation Evidence

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. Specifically, this procedure rejects the hypothesis $\beta_s = \beta_t$ in favour of $\beta_s \neq \beta_t$ whenever the absolute value of

$$T_{n,(s,t)} = \frac{\hat{\beta}_{n,s} - \hat{\beta}_{n,t}}{\text{se}(\hat{\beta}_{n,s} - \hat{\beta}_{n,t})}$$

17Romano & 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.
exceeds the $1 - \alpha$ quantile of $\max_{(s,t) \in K^2} |T^*_{n,(s,t)}|$, where

$$T^*_{n,(s,t)} = \frac{(\hat{\beta}^*_{n,s} - \hat{\beta}_{n,s}) - (\hat{\beta}^*_{n,t} - \hat{\beta}_{n,t})}{\text{se}(\hat{\beta}^*_{n,s} - \hat{\beta}^*_{n,t})}.$$  

Note that, if there are no other explanatory variables (i.e., if the $X_i \eta$ term is excluded from model (4)), as in the design of our simulations in Section 2 and below, then the estimate of $\text{Cov}(\hat{\beta}_{n,s}, \hat{\beta}_{n,t})$, which $\text{se}(\hat{\beta}_{n,s} - \hat{\beta}_{n,t})$ generally depends on, will be zero for all $s \neq t$. That is, $\text{se}(\hat{\beta}_{n,s} - \hat{\beta}_{n,t}) = \sqrt{\text{se}(\hat{\beta}_{n,s})^2 + \text{se}(\hat{\beta}_{n,t})^2}$ in such cases. Nonetheless, $T_{n,(s,t)}$ and $T_{n,(s,t')}$/ will be correlated, since $\hat{\beta}_{n,s} - \hat{\beta}_{n,t}$ and $\hat{\beta}_{n,s} - \hat{\beta}_{n,t'}$ are correlated.

The design of our simulations is the same as the one described in Section 2, but with several variations. First, we consider $k \in \{5, 10\}$ and assign $n_0 \in \{50, 100, 200\}$ observations to each of the $k$ 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 & 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 & Flachaire, 2008). Throughout all of our simulations, we set the nominal FWER $\alpha$ equal to 0.05 and generate 100,000 samples.

Table 2 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 $k = 10$. 

18
In order to compare the power of the max-$T$ procedure and the (unmodified) overlap procedure, we follow BT and Romano & 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 4a and 4b 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 5a and 5b 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

Table 2: Empirical FWERs for max-$T$ and overlap procedures

<table>
<thead>
<tr>
<th>$k$</th>
<th>$n_0$</th>
<th>Homoskedasticity</th>
<th>Heteroskedasticity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>max-$T$</td>
<td>Overlap</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>0.052</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.051</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.052</td>
<td>0.050</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>0.053</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.053</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.051</td>
<td>0.049</td>
</tr>
</tbody>
</table>
Figure 4: Empirical average power for max-$T$ and overlap procedures ($k = 5$)

(a) Homoskedastic errors  (b) Heteroskedastic errors

Figure 5: Empirical probability of identifying the best for the unmodified and modified overlap procedures ($k = 5$)

(a) Homoskedastic errors  (b) Heteroskedastic errors
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 generally, this ratio will decrease as $k$ is increased, but increase as $n$ is increased (asymptotically, both procedures will correctly identify the largest parameter with probability one).

5 Additional Empirical Examples

5.1 Matching Grants in Charitable Giving

Karlan & List (2007), hereafter KL, conducted a large-scale field experiment to examine the effect of matching grants on charitable giving.\footnote{Data for this paper is available for download from: http://www.aeaweb.org/aer/data/dec07/20060421_data.zip} Matching grants are schemes in which an individuals’ donation to a charity is amplified by a third party (the “matching donor”). For example, with a 2:1 matching ratio, the matching donor donates $2 for every $1 donated by the individual.

The experiment involved sending letters to 50,083 previous donors of a politically-oriented charity asking them to donate again. Approximately one-third of these donors were randomly assigned to a control group, and received letters which made no mention of a matching grant. The remaining (“treated”) donors received letters which varied along three dimensions: the matching ratio (either 1:1, 2:1, or 3:1), the maximum size of the matching grant (either $25,000, $50,000, $100,000, or none), and the donation amount used to illustrate how the
matching grant worked (either 1, 1.25, or 1.50 times the donor’s maximum previous donation). That is, there are \( k = 3 \times 4 \times 3 = 36 \) different treatments (the experiment was designed so that “treated” donors received one of these treatments with probability 1/36).

Although KL consider two outcomes, response (a binary variable) and amount given, we focus here solely on the latter.\(^{19}\) Moreover, our model differs from that of KL in two important ways. First, KL utilize a more restrictive (but also more parsimonious) model in which the different treatments interact, while our model—which conforms to the specification in (1)—includes a distinct treatment effect for each of the \( k = 36 \) treatments. Second, unlike KL, we include the following individual-level explanatory variables in our model: the number of months since the last donation, the highest previous donation, the number of previous donations, the number of years since the initial donation, an indicator for having previously donated in the same year, an indicator for being female, and an indicator for being a couple. Because data on some of these explanatory variables are missing for some individuals, we are left with \( n = 48,934 \) observations.

We estimate our model using OLS and obtain heteroskedasticity-consistent standard errors (specifically, the HC0 variant of MacKinnon & White, 1985). Given a nominal FWER of \( \alpha = 0.05 \) and 999 replications of the wild bootstrap, we obtain a value of 2.809 for \( \gamma \) using the unmodified overlap procedure (this is the value obtained after the first iteration; no further refinement was possible).

Figure 6 displays our uncertainty intervals centered around the treatment effects (the dotted horizontal lines correspond to the endpoints of \( \tilde{C}_{n,0} \)). From this figure, it is immediately obvious that we cannot infer anything about the ordering of the treatment effects or any of their signs.

As a point of comparison, we also compute \( T \)-statistics of the form (9) for each of the

\(^{19}\)In their recent multiple-testing based analysis of the same data, LSX use four different approaches: one in which just different outcomes are considered, one in which just different treatments are considered, one in which just different “types” of donors are considered, and one in which all the different outcomes, treatments, and “types” are simultaneously considered. Note that LSX group the 36 different treatments that we consider into just 3 treatments, which vary only on the basis of the matching ratio.
Figure 6: Charitable giving example

Note: Treatment labels are formed as follows: matching ratio / maximum matching grant in thousands of dollars (NM = no maximum) / size of illustrative contribution relative to the donor’s maximum previous donation.

Figure 7: Histogram of $T$-statistics in charitable giving example
(36+1) = 666 relevant pairwise parameter comparisons. A histogram of these $T$-statistics is shown in Figure 7. It is interesting to note that 17 of these $T$-statistics fall outside of the interval $[-1.960, 1.960]$ (under mild regularity conditions, the limiting distribution of $T_{n,(s,t)}$ when $\beta_s = \beta_t$ is standard normal here). In other words, had we separately tested the equality of each pair of parameters at the 5% nominal level (i.e., without any consideration of the FWER), we would have rejected 17 out of 666 hypotheses. On the other hand, the critical value for the max-$T$ procedure (i.e., the 0.95 quantile of $\max_{(s,t) \in K^2} |T^*_{n,(s,t)}|$; see Section 4) is 3.382. Since none of the $T$-statistics are larger in absolute value than this critical value (cf. Figure 7), the max-$T$ procedure would not reject any of the 666 hypotheses (which is consistent with the results of the overlap procedure). The primary advantage of the overlap procedure is that its graphical nature allows users to make these inferences much more simply (i.e., without having to examine a massive table of $T$-statistics or $p$-values).

5.2 Student Achievement Programs

Angrist, Lang & Oreopoulos (2009), hereafter ALO, conducted a field experiment at a large university in Canada in order to examine programs aimed at improving students’ academic performance. The experiment involved sorting students into a control group and $k = 3$ treatment groups. Students in the first treatment group were offered support services (supplemental instruction and peer advising), while students in the second treatment group were offered financial incentives (cash awards depending on their performance). Students in the third treatment group were offered both support services and financial incentives.

Although ALO present results for several different outcome variables, we focus solely on GPA, which was measured at the end of first-year and again at the end of second-year.

---

20 A complete listing of these $T$-statistics is given in Supplementary Appendix II.
21 Recall from Section 4 that these test statistics will, in general, be correlated. Thus, even if all of the treatment effects were equal to zero, we would expect that less than 5% of the $T$-statistics (obtained from a single sample) would fall outside of the interval $[-1.960, 1.960]$. However, the probability that at least one of the $T$-statistics would fall outside of this interval is well in excess of 0.05 (cf. Section 2).
22 This dataset is publicly-available from: https://www.aeaweb.org/aej-applied/data/2007-0062_data.zip
In order to examine the effects of the different treatments in this case, ALO estimate the following model:

\[ \text{GPA}_i = \beta_0 + \delta_1 \text{PF}_i + \delta_2 \text{PS}_i + \delta_3 \text{PFS}_i + \mathbf{X}_i' \eta + V_i, \] (10)

where PF, PS, PFS are indicator variables indicating participation in the financial incentives only treatment, the support services only treatment, and the combined treatment, respectively; and \( \mathbf{X} \) contains sets of indicator variables for mother tongue, high school group, number of courses in fall term, self reports on how often the student procrastinates, mother’s education, and father’s education. ALO treat participation in the three treatments as endogenous, and use assignment to these groups as instruments.\(^{23}\)

There are \( n = 1,542 \) observations, and the model is estimated using 2SLS. Standard errors are clustered by student. The first column of Table 8 in ALO provides detailed results.

Our focus here is solely on determining whether or not there is a single “best” treatment (i.e., we use the modified overlap procedure). In doing so, we first re-write the above model in the form of model (4), where \( \beta_0 \) is multiplied by an indicator variable for membership in the control group.\(^{24}\) For simplicity, however, we center our uncertainty intervals around the treatment effects (and zero).

Given a nominal FWER of \( \alpha = 0.05 \), we obtain a value of 0.504 for \( \gamma \) using 999 replications of the wild cluster bootstrap (which we modified for 2SLS following the approach of Davidson & MacKinnon, 2010). Figure 8 displays the lower half of \( \tilde{C}_{n,[1]} = \tilde{C}_{n,3} \) and the upper halves of \( \tilde{C}_{n,[2]} = \tilde{C}_{n,2} \), \( \tilde{C}_{n,[3]} = \tilde{C}_{n,1} \), and \( \tilde{C}_{n,[4]} = \tilde{C}_{n,0} \).

Since \( \tilde{L}_{n,[1]} > \tilde{U}_{n,[8]} \), for \( t \in \{2, 3, 4\} \), we can infer that the combined treatment is the “best”. Note, however, that the modified overlap procedure does not allow us to compare the other treatments to one another (or to the control). ALO, on the other hand, simply test that each of the treatment effects are zero, and conclude that only \( \delta_3 \) is positive (each

\(^{23}\)ALO also examine “intention-to-treat” effects, where the treatment effects are the coefficients on indicator variables for assignment to the treatments. These effects are estimated for both men and women (both separately and together), while the treatment effects we focus on here are estimated only for women.

\(^{24}\)In the first stage of obtaining 2SLS estimates, the indicator variable for membership in the control group and the indicator variables for each of the treatments are regressed on the instruments.
test is conducted at the 5% nominal level, without any consideration of the FWER).\textsuperscript{25}

6 Conclusion

In this paper, we have shown how multiple treatments can be compared using a simple, graphical procedure which (asymptotically) controls the FWER. Our proposed approach complements the growing literature within econometrics that focuses on testing for heterogenous treatment effects (i.e., situations where different types of individuals may respond differently to the \textit{same} treatment). A natural extension of our approach would be to incorporate such heterogenous treatment effects. We leave this to future work.

\textsuperscript{25}ALO do informally compare estimates of different treatment effects in other parts of the paper. For example, on p. 14 they state that “the [intention-to-treat] estimates for women suggest the combination of services and fellowships \ldots had a larger impact than fellowships alone.”
Appendix A: The $k = 1$ Case

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 $(0, 0)'$ and covariance matrix

$$
\begin{pmatrix}
\Sigma_0 & 0 \\
0 & \Sigma_1
\end{pmatrix}.
$$

We will also assume $\hat{\Sigma}_{n,s}$ is a consistent estimate of $\Sigma_s$, so that $\text{se} \left( \hat{\beta}_{n,s} \right) = \sqrt{\hat{\Sigma}_{n,s}/n}$, 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

$$
\lim_{n \to \infty} P \left( \frac{\sqrt{n} \left( \hat{\beta}_{n,0} - \hat{\beta}_{n,1} \right)}{\sqrt{\hat{\Sigma}_{n,0} + \hat{\Sigma}_{n,1}}} > \Phi^{-1} \left( 1 - \alpha/2 \right) \right) = \alpha/2,
$$

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

$$
\lim_{n \to \infty} P \left( L_{n,0}(B_n \Phi^{-1}(1 - \alpha/2)) > U_{n,1}(B_n \Phi^{-1}(1 - \alpha/2)) \right) = \alpha/2,
$$

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 (12), we have

$$
\lim_{n \to \infty} P \left( L_{n,1}(B_n \Phi^{-1}(1 - \alpha/2)) > U_{n,0}(B_n \Phi^{-1}(1 - \alpha/2)) \right) = \alpha/2.
$$
Thus, since the events within (12) and (13) are disjoint, we have

$$\lim_{n \to \infty} P \left( \max_{s \in \{0,1\}} L_{n,s}(B_n \Phi^{-1}(1 - \alpha/2)) > \min_{s \in \{0,1\}} U_{n,s}(B_n \Phi^{-1}(1 - \alpha/2)) \right) = \alpha.$$ 

Accordingly, rather than choosing for $\gamma$ via re-sampling, we could form uncertainty intervals for $\beta_0$ and $\beta_1$ as

$$C_{n,0}(B_n \Phi^{-1}(1 - \alpha/2)) \quad (14)$$

and

$$C_{n,1}(B_n \Phi^{-1}(1 - \alpha/2)), \quad (15)$$

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, $C_{n,0}(\Phi^{-1}(1 - \alpha/2))$ and $C_{n,0}(\Phi^{-1}(1 - \alpha/2))$, 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 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.,

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

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

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

Finally, it can be seen from (11) that inferences based on the non-overlap of the un-
certainty intervals in (14) and (15) will be identical to those based on a hypothesis testing procedure which rejects $\beta_s = \beta_t$ in favour of $\beta_s > \beta_t$ if

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

**Appendix B: Ignoring Treatment Effect Comparisons**

We now consider narrowing our problem to focus solely on whether or not any of the treatment effects is different from zero. That is, we ignore all pairwise comparisons of the treatment effects.

As with our more general problem, our approach is based on the construction of a collection of uncertainty intervals. Here, however, it will be much more convenient to proceed directly from model (1). That is, we construct uncertainty intervals for each $\delta_s$, $s \in \{1, \ldots, k\}$, as

$$D_{n,s}(\lambda) = \left[ \hat{\delta}_{n,s} \pm \lambda \times se(\hat{\delta}_{n,s}) \right].$$

These uncertainty intervals can be used to make inferences as follows. We infer that $\delta_s > 0$ if the lower endpoint of $D_{n,s}$ is greater than zero, i.e., if

$$\hat{\delta}_{n,s} - \lambda \times se(\hat{\delta}_{n,s}) > 0.$$  

Similarly, we infer that $\delta_s < 0$ if the upper endpoint of $D_{n,s}$ is less than zero, i.e., if

$$\hat{\delta}_{n,s} + \lambda \times se(\hat{\delta}_{n,s}) < 0.$$  

---

26 Notice that, for $s \in \{1, \ldots, k\}$, comparing $\delta_s$ to zero is equivalent to comparing $\beta_s \equiv \beta_0 + \delta_s$ to $\beta_0$. Hence, if we were to proceed from model (4), we would need to construct uncertainty intervals for each $\beta_s$, $s \in K$, and then determine whether or not any of the uncertainty intervals for $\beta_1, \ldots, \beta_k$ overlap the uncertainty interval for $\beta_0$. The uncertainty intervals that we construct here for $\delta_1, \ldots, \delta_k$ can be viewed as “absorbing” the uncertainty around $\beta_0$. 

---

29
Accordingly, a feasible choice of $\lambda$ at the nominal FWER $\alpha$ is the smallest value satisfying

$$\frac{1}{B} \sum_{b=1}^{B} I \left( \max_{s \in \{1, \ldots, k\}} \left| \hat{\delta}_{n,s} - \hat{\delta}_{n,s} \right| - \lambda \times \text{se} \left( \hat{\delta}_{n,s} \right) > 0 \right) \leq \alpha,$$

where, for $b \in \{1, \ldots, B\}$, $\hat{\delta}_{n,s}^b$ is the $b$th replicate of $\hat{\delta}_{n,s}^*$, the bootstrap counterpart of $\hat{\delta}_{n,s}$. Notice that, with $k = 1$, $D_{n,1}$ is an asymptotically-valid $(1 - \alpha)$-level confidence interval for $\delta_1$ (see Appendix A).

It is interesting to note that inferences based on this procedure will be identical to those based on a multiple hypothesis testing procedure which rejects $\delta_s = 0$ in favour of $\delta_s > 0$ if $T_{n,s} > \lambda$, and $\delta_s < 0$ if $T_{n,s} < -\lambda$, where

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

Moreover, since the choice of $\lambda$ suggested above is just the $1 - \alpha$ quantile of $\max_{s \in \{1, \ldots, k\}} |T_{n,s}^*|$, where

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

such a testing procedure can be seen to be analogous to the max-$T$ procedure introduced in Section 4.27

References


27Lee & Shaikh (2014) utilize a similar multiple hypothesis testing procedure, coupled with a stepwise refinement, to decide if any of the treatment effects (which, in their setup, vary based on the type of individual rather than on the treatment received) differ from zero.


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


Supplementary Appendix

I: Computational Details for Performance Pay Example

The computational steps involved in choosing $\gamma$ for the example in Section 3.2 are outlined below. 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 (8). Store the estimates $\hat{\gamma}_n = (\hat{\beta}_{n,0}, \hat{\beta}_{n,1}, \hat{\beta}_{n,2})'$ and their corresponding standard errors (clustered by school in this example), as well as the fitted values $\{\hat{\text{Score}}_i\}_{i=1}^n$ and the residuals $\{\hat{V}_i\}_{i=1}^n$.

2. For each $i \in \{1, \ldots, n\}$, generate

$$\text{Score}_i^* = \hat{\text{Score}}_i + \hat{V}_i \sum_{g=1}^G W_g^* I(\text{School}_i = g),$$

where School is a categorical variable taking values in $\{1, \ldots, G\}$ (here, $G$ is the number of schools) and $\{W_g^*\}_{g=1}^G$ 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 & Flachaire, 2008).  

3. Use OLS to estimate

$$\text{Score}_i^* = \beta_0 \text{Control}_i + \beta_1 \text{Group}_i + \beta_2 \text{Individual}_i + X_i' \eta + \text{errors}.$$  

Store the estimates $\hat{\gamma}_n^* = (\hat{\beta}_{n,0}^*, \hat{\beta}_{n,1}^*, \hat{\beta}_{n,2})'$ and their corresponding standard errors.

---

28See also Webb (2014) if $G \leq 12$.  

SA.1
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 (5). This can be accomplished using the following algorithm:

5.1. Pick a candidate value for \( \gamma \).

5.2. For each \( b \in \{1, \ldots, B\} \), calculate the maximum lower endpoint

\[
\overline{L}^{eb} = \max_{s \in K} \left\{ \left( \hat{\beta}_{n,s}^{eb} - \hat{\beta}_{n,s} \right) - \gamma \times se \left( \hat{\beta}_{n,s}^{eb} \right) \right\}
\]

and the minimum upper endpoint

\[
\underline{U}^{eb} = \min_{s \in K} \left\{ \left( \hat{\beta}_{n,s}^{eb} - \hat{\beta}_{n,s} \right) + \gamma \times se \left( \hat{\beta}_{n,s}^{eb} \right) \right\},
\]

where \( \hat{\beta}_{n,s}^{eb} \) is the \( b^{th} \) replicate of \( \hat{\beta}_{n,s} \).

5.3. Calculate \( \hat{\alpha} = \# \left( \overline{L}^{eb} > \underline{U}^{eb} \right) \)/\( B \).

5.4. Iterate over Steps 5.2–5.3, using a root finder to solve \((\alpha - \hat{\alpha}) - I(\alpha < \hat{\alpha}) = 0\).\(^{29}\)

Figure C1 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 \( \gamma \).\(^{30}\) Using the algorithm in Step 5 above, we obtain a value 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.

\(^{29}\)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 C1b), 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 C1b), 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 \( \alpha < \hat{\alpha} \).

\(^{30}\)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 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).
Figure C1: Frequency of observing at least one pair of non-overlapping uncertainty intervals across bootstrap replications ($\hat{\alpha}$) as a function of $\gamma$ in performance pay example

**II: Test Statistics for Charitable Giving Example**

Table D1 displays $T$-statistics of the form (9) for each of the $\binom{36+1}{2} = 666$ relevant pairwise parameter comparisons for the example in Section 5.1. The 17 $T$-statistics that are greater than 1.960 in absolute value are in bold.
### Table D1: T-statistics in charitable giving example

<table>
<thead>
<tr>
<th>Control</th>
<th>1.1/25/1</th>
<th>1.1/50/1</th>
<th>1.1/100/1</th>
<th>1.1/NM/1</th>
<th>2.1/25/1</th>
<th>2.1/50/1</th>
<th>2.1/100/1</th>
<th>2.1/NM/1</th>
<th>3.1/25/1</th>
<th>3.1/50/1</th>
<th>3.1/100/1</th>
<th>3.1/NM/1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>-1.302</td>
<td>0.099</td>
<td>0.014</td>
<td>0.684</td>
<td>-0.055</td>
<td>-2.096</td>
<td>0.750</td>
<td>0.047</td>
<td>0.141</td>
<td>0.011</td>
<td>0.070</td>
<td>-0.337</td>
</tr>
<tr>
<td>1/1/25/1</td>
<td>0.705</td>
<td>0.697</td>
<td>1.097</td>
<td>1.385</td>
<td>0.871</td>
<td>1.923</td>
<td>0.652</td>
<td>0.137</td>
<td>0.031</td>
<td>0.185</td>
<td>1.555</td>
<td>0.053</td>
</tr>
<tr>
<td>1/1/50/1</td>
<td>-0.154</td>
<td>0.185</td>
<td>1.155</td>
<td>0.903</td>
<td>0.519</td>
<td>0.954</td>
<td>0.053</td>
<td>0.926</td>
<td>0.099</td>
<td>0.733</td>
<td>1.286</td>
<td>0.080</td>
</tr>
<tr>
<td>1/1/100/1</td>
<td>0.398</td>
<td>1.428</td>
<td>-0.606</td>
<td>-0.955</td>
<td>0.755</td>
<td>1.234</td>
<td>0.125</td>
<td>1.234</td>
<td>0.047</td>
<td>0.087</td>
<td>0.875</td>
<td>0.290</td>
</tr>
<tr>
<td>1/1/NM/1</td>
<td>1.095</td>
<td>-0.992</td>
<td>-1.353</td>
<td>0.391</td>
<td>0.875</td>
<td>-0.290</td>
<td>0.845</td>
<td>0.287</td>
<td>0.054</td>
<td>-0.672</td>
<td>1.196</td>
<td>0.097</td>
</tr>
<tr>
<td>2/1/25/1</td>
<td>-1.896</td>
<td>2.179</td>
<td>-0.701</td>
<td>-0.259</td>
<td>-1.362</td>
<td>-0.345</td>
<td>-1.304</td>
<td>0.047</td>
<td>0.109</td>
<td>0.008</td>
<td>-0.672</td>
<td>1.196</td>
</tr>
<tr>
<td>2/1/50/1</td>
<td>-0.389</td>
<td>1.284</td>
<td>1.747</td>
<td>0.761</td>
<td>1.793</td>
<td>0.583</td>
<td>0.137</td>
<td>0.031</td>
<td>0.185</td>
<td>1.555</td>
<td>0.053</td>
<td></td>
</tr>
<tr>
<td>2/1/100/1</td>
<td>1.598</td>
<td>2.059</td>
<td>1.100</td>
<td>2.153</td>
<td>0.873</td>
<td>0.137</td>
<td>0.031</td>
<td>0.185</td>
<td>1.555</td>
<td>0.053</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2/1/NM/1</td>
<td>0.465</td>
<td>0.608</td>
<td>0.407</td>
<td>-0.606</td>
<td>-0.654</td>
<td>-0.109</td>
<td>-1.108</td>
<td>0.008</td>
<td>0.137</td>
<td>0.031</td>
<td>0.185</td>
<td>1.555</td>
</tr>
<tr>
<td>3/1/25/1</td>
<td>-1.160</td>
<td>-0.080</td>
<td>0.031</td>
<td>0.097</td>
<td>0.047</td>
<td>0.054</td>
<td>0.054</td>
<td>0.008</td>
<td>0.137</td>
<td>0.031</td>
<td>0.185</td>
<td>1.555</td>
</tr>
<tr>
<td>3/1/50/1</td>
<td>1.155</td>
<td>-0.062</td>
<td>0.137</td>
<td>0.031</td>
<td>0.185</td>
<td>1.555</td>
<td>0.053</td>
<td>0.137</td>
<td>0.031</td>
<td>0.185</td>
<td>1.555</td>
<td>0.053</td>
</tr>
<tr>
<td>3/1/100/1</td>
<td>1.051</td>
<td>0.008</td>
<td>0.137</td>
<td>0.031</td>
<td>0.185</td>
<td>1.555</td>
<td>0.053</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Control**

Control T-statistics greater than 1.960 in absolute value are in bold. Treatment labels are formed as follows: matching ratio / maximum matching grant in thousands of dollars (NM = no maximum) / size of illustrative contribution relative to the donor's maximum previous donation.

**Notes:**

- Treatment labels are formed as follows: matching ratio / maximum matching grant in thousands of dollars (NM = no maximum) / size of illustrative contribution relative to the donor’s maximum previous donation.

**SA.4**