A Simple, Graphical Procedure for Comparing Multiple Treatment Effects

Brennan S. Thompson† and Matthew D. Webb‡

February 6, 2016

<<< PRELIMINARY AND INCOMPLETE >>>

Abstract

In this paper, we utilize a new graphical procedure to show how multiple treatment effects can be compared while controlling the familywise error rate (the probability of finding one or more spurious differences between the parameters of interest). Monte Carlo simulations suggest that this procedure adequately controls the familywise error rate in finite samples, and has average power nearly identical to a max-T procedure. We illustrate our proposed approach using three different empirical examples chosen to demonstrate the flexibility of the procedure.

Keywords: multiple comparisons; familywise error rate; average treatment effect; bootstrap

*Thanks to audience members at the University of Calgary Empirical Microeconomics Workshop, Camp Econometrics X, the 32nd Canadian Econometric Study Group Meeting, Queen’s University, the 85th Southern Economics Association Meeting. Thanks also to Otavio Camargo-Bartalotti for helpful comments. We also thank the authors of the empirical papers examined for making their data publicly available. This research was supported in part by a grant from Social Sciences and Humanities Research Council.

†Department of Economics, Ryerson University. E-mail: brennan@ryerson.ca
‡Department of Economics, Carleton University. E-mail: matt.webb@carleton.ca
1 Introduction

Many empirical questions are being answered by researchers in experimental settings. In many cases, an ‘experiment’ involves many specific treatments.\footnote{A recent examination of articles in top economics journals by Young (2015) found that regressions in these articles contain on average 5.8 treatment measures.} This paper proposes a graphical technique to evaluate the statistical significance of multiple treatments. In the case of comparing multiple treatments, the problem at hand is two-fold: (A) we want to know whether or not the effect of each treatment is different from zero, and (B) we want to know whether or not the effect of each treatment is different from that of any of the other treatments.

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

\[
Y_t = \beta_0 + \sum_{i=1}^{k} \delta_i T_{i,t} + Z_t \eta + U_t, \quad (t = 1, \ldots, n),
\]

where \(T_{i,t}\) equals one if individual \(t\) receives treatment \(i \in \{1, \ldots, k\}\) and zero otherwise; \(Z_t\) is a vector of other explanatory variables (e.g., age, gender, etc.) for individual \(t\); and \(U_t\) is an idiosyncratic error term for individual \(t\). We assume that each individual receives only one of the \(k\) treatments or is in a control group.\footnote{If there are individuals receiving a combination of treatments, such individuals could be included in a separate treatment group; see Section 4.1 for such an example.}

With \(\delta_i\) representing the (average) treatment effect of the \(i\)th treatment, the first part of our problem requires testing

\[
\delta_i = 0, \quad \text{for each } i \in \{1, \ldots, k\}, \quad (1.1)
\]
while the second part of our problem requires testing

$$\delta_i = \delta_j, \quad \text{for each unique } (i, j) \in \{1, \ldots, k\} \times \{1, \ldots, k\}. \quad (1.2)$$

Thus, our problem involves testing a total of $\binom{k+1}{2}$ restrictions: the $k$ restrictions in (1.1), plus the $\binom{k}{2}$ restrictions in (1.2).

In order to make it easier to keep track of all of these restrictions, we can re-write the above model as

$$Y_t = \beta_0 C_t + \sum_{i=1}^{k} \beta_i T_{i,t} + Z'_t \eta + U_t, \quad (1.3)$$

where $C_t$ equal one if individual $t$ belongs to the control group and zero otherwise (so that $C_t + \sum_{i=1}^{k} T_{i,t} = 1$ for all $t$); and $\beta_i \equiv \beta_0 + \delta_i$, for $i \in \{1, \ldots, k\}$. Thus, since $\delta_i = 0$ is equivalent to $\beta_i = \beta_0$, and $\delta_i = \delta_j$ is equivalent to $\beta_i = \beta_j$, our problem boils down to testing the following $\binom{k+1}{2}$ restrictions:

$$\beta_i = \beta_j, \quad \text{for each unique } (i, j) \in K \times K, \quad (1.4)$$

where $K = \{0, \ldots, k\}$.

This is quite different than testing for the joint significance of all treatments, as was considered in Young (2015). Of course, when (non-jointly) testing more than one restriction at a given nominal level, the probability of rejecting at least one true restriction, i.e., the familywise error rate (FWER), is typically well in excess of that given nominal level. To illustrate the severity of this issue, we generate, for $k \in \{2, \ldots, 5\}$, one million samples from the model in (1.3) as follows. We set $\beta_0 = \cdots = \beta_k = 0$, and assign 100 observations to the control group (i.e., $\sum_{t=1}^{n} C_t = 100$) and 100 observations to each of the $k$ treatment groups (i.e., $\sum_{t=1}^{n} T_{i,t} = 100$ for each $i \in \{1, \ldots, k\}$), so that $n = 100(k + 1)$. For each $t$, $Z_t = 0$ and $U_t$ is an independent
Within each sample, we test (A) each of the \( k \) restrictions in (1.1), and (B) each of the \( \binom{k}{2} \) restrictions in (1.2), using conventional \( t \)-tests at the 5% nominal level. The rejection frequencies for these tests are shown in Figure 1. Specifically, the dash-dotted line shows the frequency of rejecting at least one of the \( k \) restrictions in (1.1), while the dashed line shows the frequency of rejecting at least one of the \( \binom{k}{2} \) restrictions in (1.2). The solid line shows the frequency of rejecting at least one of the \( \binom{k+1}{2} \) restrictions in (1.4) (i.e., the empirical FWER). Notice that, even with \( k = 2 \), the empirical FWER is approximately 0.122; the frequency of rejecting (A) \( \beta_1 = \beta_0 \) and/or \( \beta_2 = \beta_0 \) is approximately 0.091, and (B) \( \beta_1 = \beta_2 \) is approximately 0.050. With \( k = 5 \), the empirical FWER is approximately 0.364.\(^3\)

In recognition of such problems, a wide variety of multiple testing procedures,  

---

\(^3\)In general, the FWER is less than or equal to \( 1 - (1 - \alpha_1)^m \), where \( m \) is the number of restrictions being tested, and \( \alpha_1 \) is the nominal level that each restriction is tested at. However, the FWER will be strictly less than \( 1 - (1 - \alpha_1)^m \) when the \( m \) tests are not independent from one another, as is the case here (to see this, note that, for example, the estimates of \( \beta_1 - \beta_0 \) and \( \beta_2 - \beta_0 \) are correlated).
ranging from simple (but overly-conservative) Bonferroni-type corrections to resampling-based stepwise procedures (Romano and Wolf, 2005a,b), have been developed to control the FWER or other generalized error rates, such as the false discovery rate (Benjamini and Hochberg, 1995). While such procedures have been extensively used for comparing multiple treatments in biostatistics (Dunnett, 1955; Dunnett and Tamhane, 1991, 1997), the economics literature has, to the best of our knowledge, ignored the problem of multiple comparisons whenever multiple treatment effects are estimated.\footnote{Recently, some researchers in economics have used multiple-testing procedures when comparing heterogeneous treatment effects, in which different types of individuals may respond differently to the same treatment; see Section 5 for further discussion.}

For the general problem of making multiple pairwise comparisons that we are interested in, Bennett and Thompson (forthcoming), hereafter BT, have recently proposed a graphical procedure that identifies differences between a pair of parameters through the non-overlap of the so-called uncertainty intervals (see Section 2) for those parameters. This graphical procedure, which can be seen as a resampling-based generalization of Tukey’s (1953) graphical procedure, is appealing as it allows users to easily determine both statistical and practical significance of pairwise differences, while also providing them with a measure of uncertainty concerning the locations of the individual parameters. The procedure is also extremely general, in that no strict distributional assumptions are imposed, and can be applied to a wide variety of statistical models; beyond linear regression models such as that in (1.3), it can also be applied to other models that may be of use in the estimation of treatment effects, such as the logistic regression model (see Section 4.2 for such an application).

The remainder of this paper is organized as follows. In Section 2, we review the basic graphical procedure of BT, as well as an extension of this procedure designed to identify the “best” treatment under consideration. Section 2 also includes a running
illustration using data from a field experiment in which different types of performance pay for teachers are compared. Section 3 describes the results of a set of Monte Carlo simulations designed to examine the finite-sample performance of the procedures. In Section 4, we present the results of two additional empirical examples chosen to demonstrate the flexibility of the procedures. Section 5 concludes.

2 The Overlap Procedure

The graphical procedure of BT involves presenting each of the parameter estimates $\hat{\beta}_{n,i}$, $i \in K$, together with a corresponding uncertainty interval,

$$\left[ \hat{\beta}_{n,i} \pm \gamma \times se(\hat{\beta}_{n,i}) \right],$$

whose length is determined by the parameter $\gamma$ (discussed below) and $se(\hat{\beta}_{n,i})$, the standard error of $\hat{\beta}_{n,i}$ (high-level assumptions on the large-sample behaviour of these estimators are given at the end of this section). We denote the lower and upper endpoints of the uncertainty interval for $\beta_i$ by $L_{n,i} = L_{n,i}(\gamma)$ and $U_{n,i} = U_{n,i}(\gamma)$, respectively.

These uncertainty intervals are used to make inferences about the ordering of the parameters of interest as follows. We infer that $\beta_i > \beta_j$ if $L_{n,i} > U_{n,j}$. If the uncertainty intervals for $\beta_i$ and $\beta_j$ overlap one another, we can make no such inference. For this reason, we refer to this procedure as the overlap procedure.

Note that, if all $k+1$ parameters are equal, the “ideal” choice of $\gamma$ at the nominal FWER $\alpha$ is the smallest value satisfying

$$\Pr \left\{ \max_{i \in K} \{ L_{n,i}(\gamma) \} > \min_{i \in K} \{ U_{n,i}(\gamma) \} \right\} \leq \alpha,$$

(2.1)
where \( P \) denotes the unknown probability mechanism from which our data is generated (the probability in (2.1) is just the probability that at least one pair of uncertainty intervals is overlapping). Of course, because \( P \) is unknown, this choice is infeasible. Instead, we choose \( \gamma \) to satisfy the bootstrap analogue of (2.1):

\[
\text{Prob}_{\hat{P}_n}\left\{ \max_{i \in K} \{L_{n,i}^*(\gamma)\} > \min_{i \in K} \{U_{n,i}^*(\gamma)\} \right\} \leq \alpha,
\]

where \( \hat{P}_n \) is an estimate of \( P \), and \( L_{n,i}^* \) and \( U_{n,i}^* \) are, respectively, the lower and upper endpoints of

\[
\left[ (\hat{\beta}_{n,i}^* - \hat{\beta}_{n,i}) \pm \gamma \times \text{se}(\hat{\beta}_{n,i}^*) \right],
\]

with \( \hat{\beta}_{n,i}^* \) and \( \text{se}(\hat{\beta}_{n,i}^*) \) obtained under sampling from \( \hat{P}_n \). As in BT, we make only the following high-level assumptions:

- For each \( i \in K \), \( \sqrt{n}(\hat{\beta}_{n,i} - \beta_i) \) and \( \sqrt{n}(\hat{\beta}_{n,i}^* - \hat{\beta}_{n,i}) \) both have the same (continuous and strictly increasing) limiting distribution.

- For each \( i \in K \), \( \sqrt{n} \times \text{se}(\hat{\beta}_{n,i}) \) and \( \sqrt{n} \times \text{se}(\hat{\beta}_{n,i}^*) \) both converge in probability to the same (positive) constant.

Under these conditions, BT show that the overlap procedure described above (1) controls the FWER 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 3 below suggest that the overlap procedure provides satisfactory control of the FWER and has good (average) power properties in finite samples.
2.1 A Simple Empirical Example

In order to illustrate the overlap procedure described above, we utilize data from Muralidharan and Sundararaman (2011), hereafter MS. This paper describes the results of a field experiment in India designed to examine the effects of offering teachers performance pay conditional upon students’ academic performance. 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 (see the fifth column of their Table 8). To make these comparisons, they first estimate the following model:

$$\text{Score}_t = \beta_0 + \delta_1 \text{Group}_t + \delta_2 \text{Individual}_t + Z_t' \eta + U_t,$$

(2.2)

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 $Z$ contains the combined math and language score in year 0, as well as a set of indicator variables for subdistricts. There are $n = 29760$ observations, and the model is estimated using ordinary least squares. Standard errors are clustered by school. Results are shown in the second column of Table 1.

Next, MS test the following three restrictions:

---

5The data used in this example is publicly-available at:
Table 1: Performance pay example: Parameter estimates

<table>
<thead>
<tr>
<th></th>
<th>Model (2.2)</th>
<th>Model (2.3)</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></td>
<td>0.286</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.172)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td></td>
<td>0.415</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.168)</td>
</tr>
<tr>
<td>$\beta_2 - \beta_0$</td>
<td>0.283</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.058)</td>
<td></td>
</tr>
<tr>
<td>$\beta_2 - \beta_1$</td>
<td>0.129</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.068)</td>
<td></td>
</tr>
</tbody>
</table>

Note: Standard errors in brackets.

MS1: $\delta_1 = 0$

MS2: $\delta_2 = 0$

MS3: $\delta_2 = \delta_1$

The $T$-statistics corresponding to the tests of these three restrictions 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 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 the model in (1.3), i.e.,

$$\text{Score}_t = \beta_0 \text{Control}_t + \beta_1 \text{Group}_t + \beta_2 \text{Individual}_t + Z'_t \eta + U_t, \quad (2.3)$$

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 the model in (2.2) are the same as the estimates of $\beta_1 - \beta_0$, $\beta_2 - \beta_0$, and $\beta_2 - \beta_1$, respectively, arising from the model in (2.3).

With a nominal FWER of $\alpha = 0.05$, we obtain a value of 0.495 for $\gamma$ with the overlap procedure. This choice is based on 999 wild cluster bootstrap (Cameron et al., 2008) replications. 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 can not infer that $\beta_1 > \beta_0$ (or, equivalently, that $\delta_1 > 0$).
- Since $L_{n,2} > U_{n,0}$, we infer that $\beta_2 > \beta_0$ (or, equivalently, that $\delta_2 > 0$).
- Since the uncertainty intervals for $\beta_2$ and $\beta_1$ overlap, we can not infer that $\beta_2 > \beta_1$ (or, equivalently, that $\delta_2 - \delta_1 > 0$).

Thus, while our conclusion on the second restriction (MS2) is consistent with MS, our conclusions on the first and third restrictions (MS1 and MS3) are not.\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 asymptotic $p$-value of just under 0.06).}

We also consider an alternative – and arguably more useful – method of visualizing these results in Figure 2b. Here, the quantities are re-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 both $\hat{\beta}_{n,1}$ and $\hat{\beta}_{n,2}$, and from the endpoints of the uncertainty intervals for $\beta_1$ and $\beta_2$.\footnote{Although we refer to these re-centered uncertainty intervals as uncertainty intervals for $\delta_1$ and $\delta_2$, it is important to emphasize that their lengths are computed using the standard errors of the estimates of $\beta_1$ and $\beta_2$, respectively, rather than the standard errors of the estimates of $\delta_1$ and $\delta_2$.}

Moreover, we include a dotted horizontal line at the the value of $U_{n,0} - \hat{\beta}_{n,0} = 0.083$, the upper endpoint of the re-centered version of the uncertainty interval for $\beta_0$ (this dotted horizontal line can be thought of as the upper endpoint of the “uncertainty interval for zero”; if the vertical axis extended far enough below zero, we would include another dotted horizontal line representing the the lower endpoint of the “uncertainty interval for zero”). Given that the uncertainty interval for $\delta_2$ (the treatment effect of for the individual incentive) lies entirely above this dotted horizontal line, for example, one can quickly infer that $\delta_2 > 0$ (which is equivalent to $\beta_2 > \beta_0$, as was seen from Figure 2a).

Figure 2: Performance pay example: Basic overlap procedure
2.2 A Modification for Multiple Comparisons with the Best

The procedure described above is designed to control the FWER across all pairwise comparisons. This approach allows for a (potentially complete) ranking of all the treatments under consideration. For example, assuming (without loss of generality) that a larger value of the outcome variable is “better”, one could infer that treatment $i$ is the “best”, i.e., $\beta_i > \beta_j$ for all $j \in K \setminus \{i\}$, if $L_{n,i} > U_{n,j}$ for all $j \in K \setminus \{i\}$.

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 treatments effects (and zero). Such a problem is the focus of so-called “multiple comparisons with the best” procedures (Hsu, 1981, 1984; Horrace and Schmidt, 2000). Below, we follow BT in developing a modification of the graphical procedure described in the previous section to focus on this problem.\(^8\) 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

\(^8\)In 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 $r$. \[12\]
largest parameter value. Moreover,

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

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 a smaller point estimate.

Similar to what is done in the basic overlap procedure, we infer that \( \beta_{[1]} \) is the largest parameter value in the collection if \( L_{n,[1]} > U_{n,[j]} \) for all \( j > 1 \). Thus, since our “ideal” choice of \( \gamma \) is the smallest value satisfying

\[
\text{Prob}_P \left\{ L_{n,[1]}(\gamma) > \max_{j \in \{2, \ldots, k+1\}} \{U_{n,[j]}(\gamma)\} \right\} \leq \alpha \quad (2.4)
\]

when all \( k + 1 \) parameters are equal, a feasible choice of \( \gamma \) is the smallest value satisfying

\[
\text{Prob}_{\hat{P}_n} \left\{ L_{n,[1]^*}(\gamma) > \max_{i \in \{2, \ldots, k+1\}} \{U_{n,[i]^*}(\gamma)\} \right\} \leq \alpha, \quad (2.5)
\]

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

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

That is, \( L_{n,[1]^*} \) is the lower endpoint of the uncertainty interval for the parameter which is estimated to be largest (after re-centering) when sampling from \( \hat{P}_n \).

Simulation evidence presented in the next section suggests that the choice of \( \gamma \) resulting from this modification may be substantially smaller than the choice resulting from the basic overlap procedure, thereby greatly increasing the power of the
Before moving on, however, we illustrate the modified overlap procedure using the performance pay example introduced in Section 2.1 (in practice, one should decide which procedure will be used \textit{a priori} so as to avoid “cherry picking” their results). That is, we seek to determine whether or not the treatment effect for the individual incentive (the treatment effect which is estimated to be the largest) is statistically distinguishable from both zero and from the other treatments effects. Using the modified overlap procedure, we obtain a value of 0.34 for $\gamma$, leading to substantially narrower uncertainty intervals; see Figure 3. Note that we only plot the lower endpoint of the uncertainty interval for $\beta[1] = \beta_2$ and the upper endpoints of the uncertainty intervals for $\beta[2] = \beta_1$ and $\beta[3] = \beta_0$ in Figure 3a (and similarly in Figure 3b, which is re-centered around the treatment effects, $\delta[1] = \delta_2$ and $\delta[2] = \delta_1$). This is done in order to emphasize the fact that the modified overlap procedure can not be used to make inferences about the ordering of the parameter pair ($\beta[i], \beta[j]$) whenever $\min\{i, j\} > 1$, i.e., whenever neither of the parameters is estimated to be largest.

Here, our conclusions are more in line with MS. Specifically, we infer that $\beta_2 > \beta_0$ ($\delta_2 > 0$) and that $\beta_2 > \beta_1$ ($\delta_2 > \delta_1$). That is, our conclusions on the second and third restrictions (MS2 and MS3) are consistent with MS. However, since the modified overlap procedure focuses solely on comparisons with the best, it does not allow us to infer anything about the ordering of $\beta_1$ and $\beta_0$ (or, equivalently, anything about the sign of $\delta_2$). That is, we can not say anything about MS1 either way. 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 the “second best”, “third best”, etc., is of no concern, the modified overlap procedure can be recommended on the grounds of
Figure 3: Performance pay example: Modified overlap procedure
potentially much higher power.

3 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 simple max-$T$ procedure as a benchmark for the basic overlap procedure. Specifically, this procedure rejects the restriction $\beta_i = \beta_j, i \neq j$, whenever

$$T_{n,(i,j)} = \left| \frac{\hat{\beta}_{n,i} - \hat{\beta}_{n,j}}{\text{se} \left( \hat{\beta}_{n,i} - \hat{\beta}_{n,j} \right)} \right|. \quad (3.1)$$
Table 2: Empirical FWERs for max-$T$ and basic and modified overlap procedures

<table>
<thead>
<tr>
<th>$n$</th>
<th>Homoskedasticity</th>
<th>Heteroskedasticity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>max-$T$</td>
<td>Overlap</td>
</tr>
<tr>
<td>300</td>
<td>0.063</td>
<td>0.057</td>
</tr>
<tr>
<td>600</td>
<td>0.060</td>
<td>0.056</td>
</tr>
<tr>
<td>1200</td>
<td>0.053</td>
<td>0.048</td>
</tr>
</tbody>
</table>

is greater than $1 - \alpha$ quantile of $\max_{i \neq j} T_{n,(i,j)}^*$, where

$$T_{n,(i,j)}^* = \frac{\left(\hat{\beta}_{n,i}^* - \hat{\beta}_{n,i}\right) - \left(\hat{\beta}_{n,j}^* - \hat{\beta}_{n,j}\right)}{\text{se}\left(\hat{\beta}_{n,i}^* - \hat{\beta}_{n,j}^*\right)}.$$

Note that, if there are no other explanatory variables (i.e., if the $Z_i \eta$ term is excluded from the model in (1.3)), as in the design of our simulations in Section 1 and below, then the estimate of $\text{Cov}\left(\hat{\beta}_{n,i}, \hat{\beta}_{n,j}\right)$, which depends on, will be zero for all $i \neq j$. Nonetheless, $T_{n,(i,j)}$ and $T_{n,(i,l)}$ will be correlated, since $\hat{\beta}_{n,i} - \hat{\beta}_{n,j}$ and $\hat{\beta}_{n,i} - \hat{\beta}_{n,l}$ are correlated.

In what follows, we estimate $\beta_i$, $i \in K$, using ordinary least squares and obtain heteroskedasticity-consistent standard errors using the method of White (1980). The bootstrap counterparts of these objects are obtained using the wild bootstrap (Mammen, 1993) with 199 replications. We set the nominal FWER $\alpha$ equal to 0.05.

The design of our simulations is the same as the one described in Section 1, but with several variations. First, we fix $k = 5$ and generate 10,000 samples of size $n$, with $n \in \{300, 600, 1200\}$ (i.e., when $n = 300$, there are 50 observations in the control group and 50 observations in each of the 5 treatment groups, and so on). Second, we set $\beta_i = \theta(i + 1)$, with $\theta \in \{0, 0.01, 0.02, \ldots, 1\}$, 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 $i \in \{1, \ldots, k\}$ are normal
with mean zero and variance $i + 1$.

Table 2 shows that control of the FWER for the max-$T$ procedure and both the
basic and modified overlap procedures is adequate at all of the sample sizes consid-
ered in both the homoskedastic and heteroskedastic cases. Interestingly, although
the differences are quite small, the empirical FWERs for the overlap procedures are
uniformly smaller than those for the max-$T$ procedure.

In order to compare the power of the max-$T$ procedure and the basic overlap
procedure, we follow BT and Romano and Wolf (2005b) in examining average power,
which is the proportion of false restrictions (of the form $\beta_i = \beta_j$ 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, respec-
tively. Within these figures, black lines correspond to the basic overlap procedure
and red lines correspond to the max-$T$ procedure, while lines that are solid, dashed,
and dotted correspond to $n = 300$, $n = 600$, and $n = 1200$, respectively. Evidently,
both procedures have nearly identical average power over $\theta$ and at all of the sample
sizes considered in both the homoskedastic and heteroskedastic cases.

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 param-
eter 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. Within these figures, the black line corresponds
Figure 4: Empirical average power for overlap and max-$T$ procedures

(a) Homoskedastic errors
(b) Heteroskedastic errors

Figure 5: Empirical probability of identifying the best for the basic and modified overlap procedures

(a) Homoskedastic errors
(b) Heteroskedastic errors
to the basic overlap procedure while the blue line corresponds to the modified overlap procedure. To reduce clutter, we only plot the results for $n = 600$ (where there are 100 observations in the control group and 100 observations in each of the 5 treatment groups), 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 considered 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 basic 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).

4 Additional Empirical Examples

4.1 Student Achievement Programs

Angrist et al. (2009), hereafter ALO, conducted a field experiment at a large university in Canada in order to examine programs aimed at improving students' first year academic performance and second year re-enrollment.\footnote{This dataset is publicly-available from: https://www.aeaweb.org/aej-applied/data/2007-0062_data.zip} The experiment involved sorting students into a control group and three treatment groups (i.e., $k = 3$). Students in the first treatment group were offered support services (supplemental instruction and peer advising), while students in the second treatment group were offered finan-
cial 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 here on just first semester grades for simplicity. Moreover, ALO present results for males and females (both separately and together), while we only consider females. We do so because ALO do not find any “significant” treatment effects for males or for males and females combined (see the first two columns of their Table 5). In order to examine the effects of the different treatments in this case, ALO estimate the following model:

\[
\text{Grade}_t = \beta_0 + \delta_1 \text{Support}_t + \delta_2 \text{Financial}_t + \delta_3 \text{Combined}_t + Z_t \eta + U_t,
\]

where Grade is a credit-weighted average (on a 0 – 100 grading scale) for the first semester; Support, Financial, Combined are indicator variables indicating membership in the support services only treatment group, the financial incentives only treatment group, and the combined support services and financial incentives treatment group, respectively; and \(Z\) 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. There are \(n = 729\) observations, and the model is estimated using ordinary least squares. Heteroskedasticity-consistent standard errors are obtained using the method of White (1980).

While ALO do not explicitly test for differences between the different treatment effects, they do directly compare the parameter estimates in the text. For example, on p. 14 the authors state that “the estimates for women suggest the combination of services and fellowships . . . had a larger impact than fellowships alone.” If they had formally tested for the differences between the coefficients, the \((3+1)\) = 6 restrictions
would be \((T\text{-statistics of the form given in (3.1) for the tests of these restrictions are given in brackets):}

ALO1: \( \delta_1 = 0 \ (T_{n,(1,0)} = 0.578) \)

ALO2: \( \delta_2 = 0 \ (T_{n,(2,0)} = 2.212) \)

ALO3: \( \delta_3 = 0 \ (T_{n,(3,0)} = 3.173) \)

ALO4: \( \delta_1 = \delta_2 \ (T_{n,(1,2)} = 1.215) \)

ALO5: \( \delta_1 = \delta_3 \ (T_{n,(1,3)} = 2.103) \)

ALO6: \( \delta_2 = \delta_3 \ (T_{n,(2,3)} = 1.007) \)

Thus, at the 5% nominal level, one would reject the restrictions ALO2, ALO3, and ALO5 if the tests were conducted separately (i.e., without controlling the FWER). In other words, one could conclude that only the treatment effects for the financial group and the combined group are statistically different from zero (ALO do test the first three restrictions and come to the same conclusions about them), and that the treatment effect for the combined group is statistically different from the treatment effect for the support group. Again, while ALO do not explicitly test ALO4 – ALO6, their statement quoted above (about the treatment effect for the combined group being larger than the treatment effect for the financial group) is not statistically supported, even if one ignores the multiple testing issue.

We now turn our attention to the overlap procedures. As in the performance pay example explored in Section 2, we first need to re-write the model above in the form of the model in (1.3), where \(\beta_0\) is multiplied by an indicator variable for membership in the control group. However, for simplicity, we present our uncertainty intervals
only in the re-centered form introduced at the end of Section 2.1 (i.e., in terms of the
treatment effects, δ₁, δ₂, and δ₃, rather than β₀, β₁, β₂, and β₃).

Specifically, given a nominal FWER of α = 0.05 and 999 wild bootstrap replications,
we obtain a value of 0.636 for γ with the basic overlap procedure; see Figure 6a. Since the lower endpoint of the uncertainty interval for δ₃ is above the dotted line
(the “uncertainty interval for zero”), we infer that δ₃ > 0. However, we can not infer
anything else. That is, we can reject ALO3, but none of the other restrictions. This
means that our conclusions on two of the six restrictions (ALO2 and ALO5) are not
consistent with separate tests conducted at the 5% nominal level.

For illustrative purposes, we next apply the modified overlap procedure, obtaining
a value of 0.345 for γ; see Figure 6b. Here, we can infer that δ₃ > 0 and that δ₃ > δ₁,
but can not infer that δ₃ > δ₂. That is, we can conclude that the treatment effect
for the combined group (the largest treatment effect estimate) is both positive and
larger than the treatment effect for the support group, but not that it is larger than
the treatment effect for the financial group. This reinforces our finding above that
the claim made by ALO that the treatment effect for the combined group is larger
than the treatment effect for the financial group is not statistically supported.

Overall, the conclusions from Figure 6b differ from those based on Figure 6a only
in that we can conclude that the treatment effect for the combined group is larger
than the treatment effect for the support group. Similar to what was noted in the
performance pay example, the increased power of the modified overlap procedure
comes at the cost of not being able to make comparisons between the treatment
effects for the financial group and the support group, or between either of these two
groups and zero.

4.2 Logit Example - Matching Grants in Charitable Giving

Logit setup

We now consider an application of the overlap procedure to Logit Regression. With
a slight abuse of notation, we define

\[
\beta_0 \equiv \text{Prob}(Y_t = 1|C_t = 1, T_{1,t} = \cdots = T_{k,t} = 0)
\]

and, for \(i \in \{1, \ldots, k\},\)

\[
\beta_i \equiv \text{Prob}(Y_t = 1|C_t = 0, T_{i,t} = 1, T_{j,t} = 0 \text{ for } j \neq i).
\]

Then, as in our usual setup, \(\delta_i \equiv \beta_i - \beta_0\) is the treatment effect of the \(i\)th treatment.

In the logit framework, we have

\[
\text{Prob}(Y_t = 1|C_t, T_{1,t}, \ldots, T_{k,t}) = \frac{1}{1 + \exp[-(\lambda_0 C_t + \sum_{i=1}^{k} \lambda_i T_{i,t})]},
\]
which implies that
\[ \beta_i = \frac{1}{1 + \exp(-\lambda_i)}. \]
for all \( i \in K \).

Now, let \( \hat{\lambda}_{n,i} \) be the (Q)MLE of \( \lambda_i \). Under standard regularity conditions, the limiting distribution of \( \sqrt{n}(\hat{\lambda}_{n,i} - \lambda_i) \) is normal with mean zero and variance \( \sigma_i^2 \). Thus, by the delta method, \( \sqrt{n}(\hat{\beta}_{n,i} - \beta_i) \) is normal with mean zero and variance
\[ \left( \frac{\exp(-\lambda_i)}{[1 + \exp(-\lambda_i)]^2} \right)^2 \sigma_i^2 \]
Accordingly,
\[ \text{se}\left(\hat{\beta}_{n,i}\right) = \frac{\exp(-\hat{\lambda}_{n,i})}{[1 + \exp(-\hat{\lambda}_{n,i})]^2} \times \text{se}\left(\hat{\lambda}_{n,i}\right), \]
where \( \text{se}\left(\hat{\lambda}_{n,i}\right) \) is consistent in the sense that \( \sqrt{n} \times \text{se}\left(\hat{\lambda}_{n,i}\right) \xrightarrow{P} \sigma_i \).

Notice here that the overlap procedures only require the standard errors for each parameter estimate of interest, rather than the entire estimate covariance matrix. Calculation of the uncertainty intervals follows Section 2 rather closely, with the necessary changes to both \( \beta_i \) and \( \text{se}\left(\hat{\beta}_{n,i}\right) \). Additionally, we use the bootstrap as described in Davidson and MacKinnon (2004) for Binary Dependent Variables. The bootstrap procedure involves three steps.

1. Run the logit regression of interest, and for each observation estimate the probability, call this \( \hat{y}_i \)

2. Within each bootstrap replication generate \( n \) uniform(0, 1) values, \( u_i \)

3. Compare the random values, \( u_i \), to \( \hat{y}_i \), and generate the bootstrap \( y^* \) values such that \( y^* = 1 \) if \( u_i \leq \hat{y}_i \) and \( y^* = 0 \) if \( u_i > \hat{y}_i \).

We consider an example using data from the experiment analyzed in Karlan and
The paper itself evaluates whether a matching grant increases charitable giving. Matching grants are schemes in which a donor’s donation to a charity is ‘matched’ by a third party, thus resulting in a charity receiving $1 + x$ dollars for every 1 dollar donated. The experiment involved sending letters to 50,000 previous donors of a charity asking them to donate. The 50,000 letters varied in three dimensions: the price (match) ratio, the maximum size of the matching grant, and the ask amount.

The price ratio is the primary level of the treatment. The ratio determines how much the third party would contribute for each dollar donated: there are three treatment ratios: $1:1$, $2:1$, and $3:1$, those in the control group were not offered a matching grant. The maximum size identified how much the third party was willing to donate in total matching grants. There were four maximum amounts: $25,000, 50,000, 100,000$, and no stated maximum. Finally, the ask amount suggested a donation amount for each donor, as a function of their maximum previous donation. There were three ask amounts: 1.00 times the max, 1.25 times the max, and 1.50 times the max. All subjects in the control received a letter with no mention of a matching grant.

The 50,000 letters were randomized so that 1/3rd of letters would not involve a matching grant and be the control group, and 2/3rds of letters would involve a matching grant and be treated. The letters were sent so that all subtreatments were equally likely, for instance matching ratios of $3:1$ were as likely as ratios of $1:1$. This design resulted in $36 = 3 \text{prices} \times 4 \text{max} \times 3 \text{ask}$ different types of letters being sent. Each of these 36 letters were sent to between 925-929 previous donors.

When controlling for the types of letters, the authors often treated the subtreatments as continuous variables in regression analysis. While the authors were inter-

---

10 Data for this paper is available for download from: http://www.aeaweb.org/aer/data/dec07/20060421_data.zip

---

25
ested in whether the price ratio matters, a charity might be interested in determining which specific letters were more fruitful in soliciting donations. To answer this question precisely, would involve analyzing whether any of the 36 treatment letters had a higher response rate than the control group, and whether a particular treatment letter had a higher response rate than all other treatment letters. This would involve making $37 \binom{37}{2} = 666$ unique comparisons. It is non-trivial to display that many p-values.

We analyze whether any particular letter resulted in recipients being more likely to donate, regardless of amount, than recipients of the control group. While it is impractical to highlight all 666 unique comparisons without controlling for the familywise error rate, we report each of the type A comparisons (or $\delta_i$ coefficients) in Table 3. The table estimates the coefficients from a logit regression of whether an individual gave to the charity, relative to the control group, using the following model:

$$\text{Prob}\left((\text{Gave}_i = 1)|\left(\delta_0 + \sum_{t=1}^{36} \delta_t \text{treatment}_t\right)\right).$$

The table shows that seven of the treatments are estimated to be statistically significant at the 5% level. Moreover, one of the treatments with an ask ratio of 1.25, a match ratio of 1:1, and a total matching grant of $100,000 is estimated to be significant at the 1% level. A naive interpretation would suggest that this treatment was the most significant, as it has both the largest coefficient and the smallest p-value.

We then estimate the effects of the various treatments using the overlap procedure. The resulting overlap plot can be found in Figure 7. The figure shows, that when controlling the FWER at $\alpha = 0.05$ none of the individual treatments is statistically significant. This is perhaps not surprising, as the number of comparisons results in $\gamma = 1.956$. The overlap procedure suggests that none of the letters had a statistically
significant impact on giving. One could also ask whether there was an individual
treatment which significantly increased the likelihood of giving. This would involve
the modified procedure, for multiple comparisons with the best.

Table 3: Gave by Treatment Type

<table>
<thead>
<tr>
<th>treatment</th>
<th>coeff</th>
<th>std err</th>
<th>p-value</th>
<th>treatment</th>
<th>coeff</th>
<th>std err</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=1,R=1,S=1</td>
<td>0.084</td>
<td>0.245</td>
<td>0.731</td>
<td>A=2,R=2,S=3</td>
<td>0.335</td>
<td>0.219</td>
<td>0.126</td>
</tr>
<tr>
<td>A=1,R=1,S=2</td>
<td>-0.103</td>
<td>0.267</td>
<td>0.701</td>
<td>A=2,R=2,S=4</td>
<td>0.461</td>
<td>0.207</td>
<td><strong>0.026</strong></td>
</tr>
<tr>
<td>A=1,R=1,S=3</td>
<td>0.025</td>
<td>0.252</td>
<td>0.921</td>
<td>A=2,R=3,S=1</td>
<td>0.192</td>
<td>0.233</td>
<td>0.411</td>
</tr>
<tr>
<td>A=1,R=1,S=4</td>
<td>0.191</td>
<td>0.233</td>
<td>0.414</td>
<td>A=2,R=3,S=2</td>
<td>0.336</td>
<td>0.219</td>
<td>0.125</td>
</tr>
<tr>
<td>A=1,R=2,S=1</td>
<td>0.290</td>
<td>0.224</td>
<td>0.194</td>
<td>A=2,R=3,S=3</td>
<td>0.027</td>
<td>0.252</td>
<td>0.914</td>
</tr>
<tr>
<td>A=1,R=2,S=2</td>
<td>-0.032</td>
<td>0.259</td>
<td>0.900</td>
<td>A=2,R=3,S=4</td>
<td>-0.036</td>
<td>0.259</td>
<td>0.890</td>
</tr>
<tr>
<td>A=1,R=2,S=3</td>
<td>-0.035</td>
<td>0.259</td>
<td>0.893</td>
<td>A=3,R=1,S=1</td>
<td>0.140</td>
<td>0.239</td>
<td>0.557</td>
</tr>
<tr>
<td>A=1,R=2,S=4</td>
<td>0.419</td>
<td>0.211</td>
<td><strong>0.047</strong></td>
<td>A=3,R=1,S=2</td>
<td>0.293</td>
<td>0.224</td>
<td>0.191</td>
</tr>
<tr>
<td>A=1,R=3,S=1</td>
<td>0.336</td>
<td>0.219</td>
<td>0.125</td>
<td>A=3,R=1,S=3</td>
<td>-0.101</td>
<td>0.267</td>
<td>0.704</td>
</tr>
<tr>
<td>A=1,R=3,S=2</td>
<td>0.193</td>
<td>0.233</td>
<td>0.409</td>
<td>A=3,R=1,S=4</td>
<td>0.289</td>
<td>0.224</td>
<td>0.196</td>
</tr>
<tr>
<td>A=1,R=3,S=3</td>
<td>0.460</td>
<td>0.207</td>
<td><strong>0.027</strong></td>
<td>A=3,R=2,S=1</td>
<td>0.334</td>
<td>0.219</td>
<td>0.128</td>
</tr>
<tr>
<td>A=1,R=3,S=4</td>
<td>0.084</td>
<td>0.245</td>
<td>0.731</td>
<td>A=3,R=2,S=2</td>
<td>0.420</td>
<td>0.211</td>
<td><strong>0.046</strong></td>
</tr>
<tr>
<td>A=2,R=1,S=1</td>
<td>-0.329</td>
<td>0.296</td>
<td>0.267</td>
<td>A=3,R=2,S=3</td>
<td>0.289</td>
<td>0.224</td>
<td>0.196</td>
</tr>
<tr>
<td>A=2,R=1,S=2</td>
<td>0.378</td>
<td>0.215</td>
<td>0.078</td>
<td>A=3,R=2,S=4</td>
<td>-0.101</td>
<td>0.267</td>
<td>0.704</td>
</tr>
<tr>
<td>A=2,R=1,S=3</td>
<td>0.576</td>
<td>0.198</td>
<td><strong>0.004</strong></td>
<td>A=3,R=3,S=1</td>
<td>0.423</td>
<td>0.211</td>
<td><strong>0.045</strong></td>
</tr>
<tr>
<td>A=2,R=1,S=4</td>
<td>0.085</td>
<td>0.245</td>
<td>0.728</td>
<td>A=3,R=3,S=2</td>
<td>0.242</td>
<td>0.228</td>
<td>0.290</td>
</tr>
<tr>
<td>A=2,R=2,S=1</td>
<td>0.083</td>
<td>0.245</td>
<td>0.735</td>
<td>A=3,R=3,S=3</td>
<td>0.139</td>
<td>0.239</td>
<td>0.560</td>
</tr>
<tr>
<td>A=2,R=2,S=2</td>
<td>0.242</td>
<td>0.228</td>
<td>0.290</td>
<td>A=3,R=3,S=4</td>
<td>0.423</td>
<td>0.211</td>
<td><strong>0.045</strong></td>
</tr>
</tbody>
</table>

Notes: Treatment Legend. A = Asked Amount: 1 = 1.00, 2=1.25, 3 = 1.50 times highest previous donation. R = Match Ratio: 1 = 1:1, 2 = 2:1, 3 = 3:1. S = Size of Matching Grant: 1 = $25,000, 2 = $50,000, 3 = $100,000, 4 = no max stated.

5 Conclusion

We propose a procedure for controlling the familywise error rate (FWER) when comparing multiple treatments. Note that this problem is quite distinct from the problem of comparing heterogeneous treatment effects, in which different types of individuals may respond differently to the same treatment; Anderson (2008), Lee and Shaikh (2014) and Gu and Shen (2015) consider this problem using a multiple comparisons
approach. However, these papers only consider identifying those types for which the treatment effect is non-zero; that is, they do not consider whether the differences between the (non-zero) treatments effects are non-zero.

Our proposed technique expands upon the ranking procedure introduced in Bennett and Thompson (forthcoming). Monte Carlo simulations demonstrate that the proposed technique reliably controls the FWER when estimating treatment effects with average power similar to competing techniques. We show in several empirical examples how controlling the FWER can have large implications for inference. The examples we include go beyond OLS, and include Logit estimates.

Beyond effectively controlling the FWER, our proposed procedure simplifies the inference procedure by making it graphical in nature. In one of our examples, there

Figure 7: Charitable Giving example: Overlap procedure
are 36 treatments which would necessitate 666 p-values for the resulting hypothesis
tests. Our technique replaces all of these p-value with a single figure, which involves
an uncertainty interval for each treatment. The hypothesis tests are then conduct-
ing by examining whether two uncertainty intervals overlap. When two intervals
overlap, the difference between those treatments is not statistically significant, when
controlling for a user specified familywise error rate. These figures concisely report
the information that would otherwise be reported in separate hypothesis tests, and
often make immediately clear whether any treatments were statistically significantly
different from all others.
References


