The Econometrics Journal

Econometrics Journal (2019), volume **22**, pp. 188–205.
doi: 10.1093/ectj/utz006

A simple, graphical approach to comparing multiple treatments

BRENNAN S. THOMPSON[†] AND MATTHEW D. WEBB[‡]

[†]*Department of Economics, Ryerson University, Toronto, ON, M5B 2K3, Canada.*
Email: brennan@ryerson.ca

[‡]*Department of Economics, Carleton University, Ottawa, ON, K1S 5B6, Canada.*
Email: matt.webb@carleton.ca

First version received: 13 December 2018; final version accepted: 10 March 2019.

Summary We consider 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. This 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 this approach with three empirical examples.

Keywords: *Multiple hypothesis testing*, *treatment effects*, *bootstrap*.

JEL codes: *C12*, *C15*.

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 not only to 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.[1] In recognition of this issue, a wide variety of multiple-testing procedures, ranging from the simple Bonferroni correction to resampling-based stepwise procedures (Romano and Wolf, 2005a, 2005b), have been developed to control the FWER and other generalized error rates such as the false discovery rate (the expected proportion of true hypotheses rejected; Benjamini and Hochberg, 1995). While such procedures are often used

[1] In 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 α_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 α_m.

when multiple treatments are examined in biostatistics (Dunnett, 1955; Dunnett and Tamhane, 1991), the econometrics literature has, with the exception of the forthcoming paper by List, Shaikh and Xu (2019), hereafter LSX, ignored the problem of multiple testing whenever multiple treatments are considered.[2] 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 consider a graphical approach to comparing multiple treatments that uses the resampling-based procedure of Bennett and Thompson (2016), hereafter BT, to asymptotically control the FWER. Our main contribution here is simply to demonstrate how this (very general) procedure can be adapted to a regression framework to compare treatment effects while controlling for other sources of heterogeneity. We are hopeful that the approach we consider here will be both illuminating and easy to implement for practitioners.

The advantage of the graphical approach we adapt is that it allows users to easily visualize both statistical and practical significance in the (signed) 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 all of the hypotheses of interest. Ultimately, the user is provided with a single, easy to interpret figure that clearly suggests an ordering of the treatment effects rather than a cumbersome $(k+1)$-by-$(k+1)$ table of test statistics (or p-values). In our own practice, we have found that such a figure is particularly useful in summarizing the results of an analysis for a live audience.

The following section describes our approach in more detail and provides a very simple illustration using data from a field experiment in which $k = 2$ types of performance pay for teachers are compared. Two additional empirical examples are provided in Section 3. The first of these examples is of interest because it involves a large number of treatments ($k = 36$). In the second, we consider a case where treatment effects are estimated using an instrumental variables approach. In both examples, controlling for multiple comparisons meaningfully changes the statistical inferences. Section 4 concludes.

2. METHODOLOGY

2.1. Setup

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

$$Y_i = \beta_0 + \sum_{s=1}^{k} \delta_s D_{s,i} + \mathbf{X}_i'\eta + V_i, \tag{2.1}$$

where $D_{s,i}$ equals one if individual $i \in \{1, \ldots, n\}$ participates in treatment $s \in \{1, \ldots, k\}$ and zero otherwise; $\mathbf{X}_i$ is a vector of control variables (e.g., age, gender, etc.); and V_i is an idiosyncratic

[2] Recently, 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 and Shaikh (2014), Lehrer, Pohl, and Song (2018), and Gu and Shen (2018). Young (2019) "Young 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.

error term.[3] 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 3.2 for an example). In what follows, we define the treatment effect of the s^{th} treatment as δ_s.

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

$$\delta_s = 0, \qquad \text{for each } s \in \{1, \ldots, k\}. \tag{2.2}$$

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

$$\delta_s = \delta_t, \qquad \text{for each unique } (s, t) \in \{1, \ldots, k\}^2. \tag{2.3}$$

Hence, as pointed out in the previous section, our problem involves testing a total of $\binom{k+1}{2}$ hypotheses (in contrast to the problem of jointly testing the *single* hypothesis $\delta_1 = \ldots = \delta_k = 0$).[4] Note that we have not introduced any alternatives to these hypotheses at this point; our proposed approach is designed to allow one to infer the *ordering* of any two treatment effects (or of a particular treatment effect and zero), i.e., to infer, for example, that $\delta_s > \delta_t$, for some $(s, t) \in \{1, \ldots, k\}^2$ (the analogue of a one-sided alternative).

2.2. The Overlap Procedure

The procedure of BT is designed to facilitate all pairwise comparisons within a set of parameters that have been $\sqrt{n}$-consistently estimated; their framework is quite general and they do not explicitly consider any regression models (indeed, in their simulations, the parameters of interest are simply the means of a collection of random variables with varying degrees of correlation).

In order to set up our problem in their general framework, we begin by rewriting model (2.1) as follows:

$$Y_i = \sum_{s=0}^{k} \beta_s D_{s,i} + \mathbf{X}_i' \eta + V_i, \tag{2.4}$$

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 $s \in \{1, \ldots, k\}$ (note the absence of a constant term in this model). In what follows, we denote the parameter vector $(\beta_0, \ldots, \beta_k)'$ by β, and assume that it can be estimated $\sqrt{n}$-consistently.

The reason that we rewrite our model in this way is 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:

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

[3] 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 3.2 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.

[4] In Online Supplement Section S1, 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 into our proposed approach. In Section 2.5, we simplify our problem by ignoring the hypotheses in (2.3). Interestingly, in the work on testing for heterogeneous treatment effects cited in footnote 3 above, each treatment effect is compared only to zero (and not to any of the other treatment effects).

where $K = \{0, \ldots, k\}$. That is, we wish to consider all pairwise comparisons between the $k + 1$ parameters $\beta_0, \ldots, \beta_k$. However, since interest ultimately lies in the k treatment effects, $\delta_1, \ldots, \delta_k$, we show how users can make inferences about the hypotheses in (2.2) and (2.3) more directly later in this section.

The procedure of BT, which can be seen as a resampling-based generalization 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 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 β_s lies entirely above the uncertainty interval for β_t (i.e., if $L_{n,s} > U_{n,t}$). If the uncertainty intervals for β_s and β_t overlap one another (i.e., if $C_{n,s} \cap C_{n,t} \neq \emptyset$), we can make no such inference. For this reason, BT refer to their procedure as the *overlap* procedure; it allows users to easily make comparisons between a pair of parameters by visually checking to see whether or not their uncertainty intervals overlap.

It must be emphasized that using *confidence* intervals to make inferences in this manner would be completely inappropriate. Specifically, when $k = 1$, such inferences would be overly conservative (cf. Online Supplement Section S1); as k grows, the FWER would quickly become larger than one minus the nominal confidence level.

The choice of γ is motivated as follows. If all $k + 1$ parameters are equal, then the 'ideal' choice of γ 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 α. That is, the 'ideal' choice of γ 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 γ; values of γ larger than the ideal value—but still satisfying the above condition—will result in a FWER that is weakly further below α). Since this choice is infeasible, we choose γ using the bootstrap analogue of the above.[5]

Towards this end, for $b \in \{1, \ldots, B\}$, let $\hat{\beta}^{*b}_{n,s}$ be the b^{th} replicate of $\hat{\beta}^{*}_{n,s}$, the bootstrap counterpart of $\hat{\beta}_{n,s}$. Then, a feasible choice of γ is the smallest value satisfying

$$\frac{1}{B}\sum_{b=1}^{B} I\left(\max_{s \in K} L^{*b}_{n,s}(\gamma) > \min_{s \in K} U^{*b}_{n,s}(\gamma)\right) \leq \alpha, \tag{2.5}$$

where $I(\cdot)$ is an indicator function, and $L^{*b}_{n,s}$ and $U^{*b}_{n,s}$ are, respectively, the lower and upper endpoints of

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

[5] In Online Supplement Section S1, we show that, when $k = 1$ and the limiting distribution of $\sqrt{n}(\hat{\beta}_n - \beta)$ is known, we can easily choose γ without resorting to the bootstrap.

Only the following high-level assumptions are made:

ASSUMPTION 2.1. *(a)* $\sqrt{n}(\hat{\beta}_n - \beta)$ and $\sqrt{n}(\hat{\beta}^*_n - \hat{\beta}_n)$ *both have the same (continuous and strictly increasing)* $(k+1)$*-variate limiting distribution; (b)* $\sqrt{n} \times \text{se}\left(\hat{\beta}_{n,s}\right)$ *and* $\sqrt{n} \times \text{se}\left(\hat{\beta}^*_{n,s}\right)$ *both converge in probability to the same (positive) constant, for each* $s \in K$.

These assumptions are very mild, and hold if, say, our (OLS, 2SLS, etc.) estimates are asymptotically normal and our resampling procedure is appropriately chosen (i.e., is compatible with the data generating process).

Our main result follows directly from Theorem 3.1 in BT:

THEOREM 2.1. *Let Assumption 2.1 hold. Then the overlap procedure (a) bounds the FWER from above by* α *asymptotically, (b) is consistent, in the sense that any true differences between parameter pairs are inferred with probability one asymptotically, and (c) infers a correct ordering of the parameters (when they are unequal) with probability one asymptotically.*

Simulation evidence presented both in BT and in our Online Supplement Section S2 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

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

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

$$\tilde{C}_{n,s}(\gamma) = \left[\hat{\delta}_{n,s} \pm \gamma \times \text{se}\left(\hat{\beta}_{n,s}\right)\right], \tag{2.6}$$

where $\hat{\delta}_{n,s} \equiv \hat{\beta}_{n,s} - \hat{\beta}_{n,0}$. It is important to emphasize that (2.6) is not an uncertainty interval for δ_s itself, i.e., its width is proportional to $\text{se}\left(\hat{\beta}_{n,s}\right)$ rather than to $\text{se}\left(\hat{\delta}_{n,s}\right)$.[6]

Denoting the lower and upper endpoints of $\tilde{C}_{n,s}$ by $\tilde{L}_{n,s}$ and $\tilde{U}_{n,s}$, 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 $\tilde{L}_{n,s} > \tilde{U}_{n,t}$.

2.3. *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.

[6] In Section 2.5, where we focus only on comparisons between the treatment effects and zero (i.e., where we ignore the comparisons between the different treatment effects), we utilize uncertainty intervals for the treatment effects themselves.

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

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

so that $t \in A_{n,s}$ whenever the uncertainty intervals for β_s and β_t overlap one another (i.e., whenever the pairwise comparison between β_s and β_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 γ 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 γ as the smallest value satisfying

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

Notice that, here, we are concerned only with the non-overlap of (re-centred) uncertainty intervals that 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 (Theorem 4.1) show that, so long as at least one pair of parameters is equal, the stepwise refinement results in an FWER exactly equal to α asymptotically; this is true of the basic procedure only if *all* of the parameters are equal.

2.4. A Small-Scale Empirical Example

In order to provide an extremely simple illustration of our proposed approach, we utilize data from Muralidharan and Sundararaman (2011), hereafter MS (Section 3 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.[7] Specifically, MS analyse 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:

$$\text{Score}_i = \beta_0 + \delta_1 \text{Group}_i + \delta_2 \text{Individual}_i + \mathbf{X}_i'\eta + V_i, \tag{2.7}$$

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 $\mathbf{X}$ 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 one-third to each of the two treatment groups), and the model is estimated using OLS. Standard errors are

[7] The data used in this example are available from: https://www.journals.uchicago.edu/doi/suppl/10.1086/659655

Table 1. Performance pay example: parameter estimates.

	Model (2.7)	Model (2.8)
β_0	0.132	0.132
	(0.168)	(0.168)
δ_1	0.154	
	(0.057)	
δ_2	0.283	
	(0.058)	
$\delta_2 - \delta_1$	0.129	
	(0.068)	
β_1		0.286
		(0.172)
β_2		0.415
		(0.168)
$\beta_1 - \beta_0$		0.154
		(0.057)
$\beta_2 - \beta_0$		0.283
		(0.058)
$\beta_2 - \beta_1$		0.129
		(0.068)

Note. Clustered standard errors are in brackets.

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: $\delta_2 = \delta_1$.

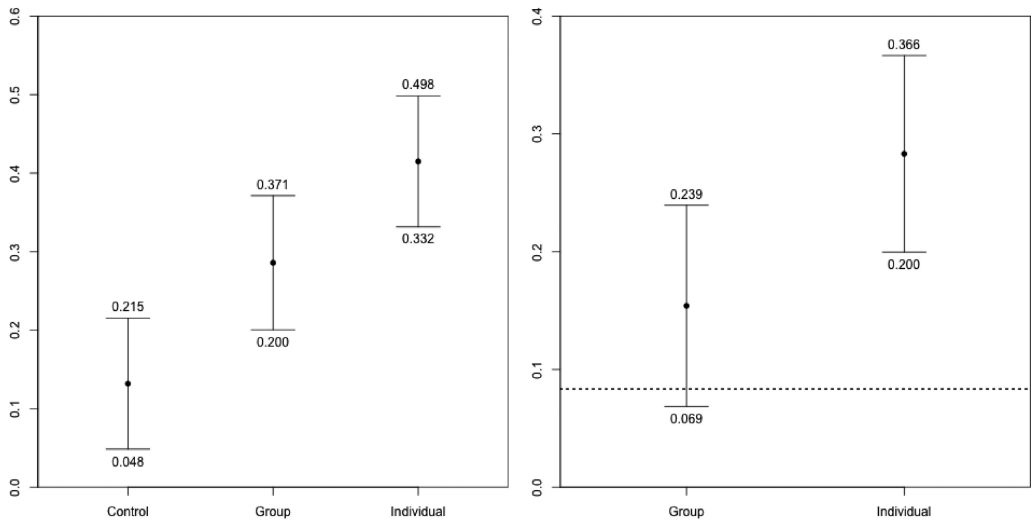
T-statistics corresponding to tests of these three hypotheses are 2.702, 4.879, and 1.897, respectively. Thus, MS conclude that both treatment effects are statistically different from zero, and from one another (even MS3 could be rejected in favour of a two-sided alternative at a nominal level of slightly 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 rewrite the model above in the form of model (2.4), i.e.,

$$\text{Score}_i = \beta_0 \text{Control}_i + \beta_1 \text{Group}_i + \beta_2 \text{Individual}_i + \mathbf{X}_i' \eta + V_i, \tag{2.8}$$

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 δ_1, δ_2, and $\delta_2 - \delta_1$ arising from model (2.7) are identical to the estimates of $\beta_1 - \beta_0$, $\beta_2 - \beta_0$, and $\beta_2 - \beta_1$, respectively, arising from model (2.8).

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 γ (this is the value obtained after the first



(a) Model (2.8) Estimates (b) Model (2.8) Estimates - Recentered

Figure 1. Performance pay example.

iteration; no further refinement was possible).[8] The resulting uncertainty intervals are shown in Figure 1(a), and can be interpreted as follows:

- Since the uncertainty intervals for β_1 and β_0 overlap, we cannot infer anything about their ordering (or, equivalently, anything about the sign of δ_1).
- Since the uncertainty interval for β_2 lies entirely above the uncertainty interval for β_0, we infer that $\beta_2 > \beta_0$ (or, equivalently, that $\delta_2 > 0$).
- Since the uncertainty intervals for β_2 and β_1 overlap, we cannot infer anything about their ordering (or, equivalently, anything about the ordering of δ_2 and δ_1).

Thus, while our results are consistent with rejecting MS2, they are not consistent with rejecting either MS1 or MS3.[9]

Figure 1(b) displays the same uncertainty intervals centred 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 β_1 and β_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 $\delta_2 > 0$. We find that such a figure makes it much easier

[8] Online Supplement Section S3 provides computational details for this example. Moreover, code for the procedure can be found at: https://sites.google.com/site/matthewdwebb/code

[9] We also applied the overlap procedure at a nominal FWER of $\alpha = 0.06$ (obtaining a value of 0.476 for γ), and found that the uncertainty intervals for β_2 and β_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 β_2 and β_1 are non-overlapping is $\alpha = 0.149$ (see BT, Section 3.3, for a discussion of multiplicity-adjusted p-values).

to distinguish the comparisons of each treatment to a control while simultaneously comparing each treatment to all the others.

2.5. *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 (i.e., the $\binom{k}{2}$ hypotheses in (2.3)), and focus on the so-called problem of 'multiple comparisons with a control' (Hsu, 1996) that was first explored by Dunnett (1955).

BT do not consider such a problem, but the procedure we outline here follows their general approach of constructing a set of uncertainty intervals for the parameters of interest. Unlike what was done above, however, it will be much more convenient to proceed directly from model (2.1).[10] That is, we construct uncertainty intervals for each δ_s, $s \in \{1, \ldots, k\}$, as

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

Note that this differs from (2.6) in that its width is proportional to se $\left(\hat{\delta}_{n,s}\right)$ rather than to se $\left(\hat{\beta}_{n,s}\right)$, i.e., $D_{n,s}$ is actually an uncertainty interval for δ_s, unlike $\bar{C}_{n,s}$, which is just a re-centred uncertainty interval for β_s.

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

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

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

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

Accordingly, a feasible choice of λ at the nominal FWER α is the smallest value satisfying

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

where, for $b \in \{1, \ldots, B\}$, $\hat{\delta}^{*b}_{n,s}$ is the b^{th} replicate of $\hat{\delta}^{*}_{n,s}$, the bootstrap counterpart of $\hat{\delta}_{n,s}$.

To illustrate this approach, we return to the performance pay example introduced in the previous section. Following the procedure described above, we obtain a value of 2.53 for λ (recall that the estimates of δ_1 and δ_2 and their standard errors have already been provided in Table 1). Figure 2 displays $D_{n,1}$ and $D_{n,2}$, the uncertainty intervals for δ_1 and δ_2, respectively.

Here, we are able to infer that *both* treatment effects are positive (i.e., $\delta_1 > 0$ and $\delta_2 > 0$), since the uncertainty for each lies entirely above zero (recall that in the previous section, we were not able to infer anything about the sign of δ_1). This example can thus be seen to nicely illustrate the trade-off that is inherent with any multiple-testing procedure: by controlling for a greater number of comparisons (in this case the comparisons between the treatment effects), one may have to sacrifice some power.

[10] Notice that, for $s \in \{1, \ldots, k\}$, comparing δ_s to zero is equivalent to comparing $\beta_s \equiv \beta_0 + \delta_s$ to β_0. Hence, if we were to proceed from model (2.4), we would need to construct uncertainty intervals for each β_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 β_0. The uncertainty intervals that we construct here for $\delta_1, \ldots, \delta_k$ can be viewed as 'absorbing' the uncertainty around β_0.

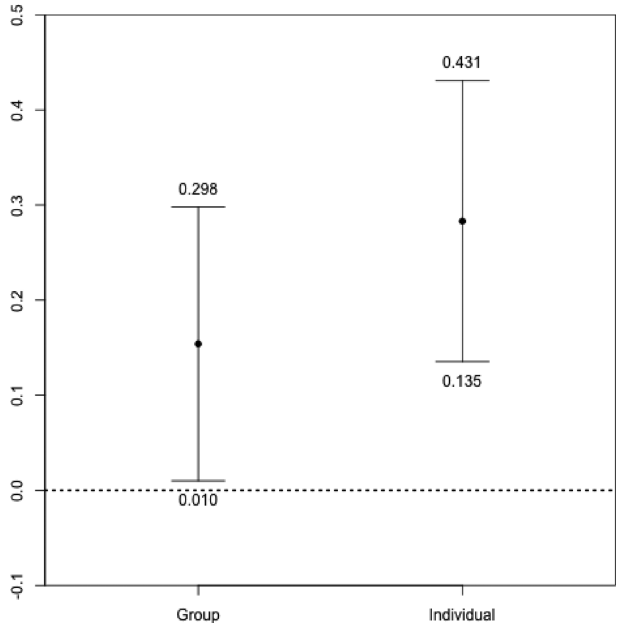


Figure 2. Performance pay example: overlap procedure modified to ignore treatment effect comparisons.

2.6. A Modification for Multiple Comparisons with the Best

Thus far, we have primarily 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 $t \in K\backslash\{s\}$, if $L_{n,s} > U_{n,t}$ for all $t \in K\backslash\{s\}$.[11] 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 centres on identifying only the (first) best treatment. That is, we may only want to know whether or not the treatment effect that 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.[12] 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), the power

[11] Note that $\beta_s > \beta_t$ for all $t \in K\backslash\{s\}$ is equivalent to $\delta_s > 0$ and $\delta_s > \delta_t$ for all $t \in \{1, \ldots, k\}\backslash\{s\}$. 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. Of course, such a ranking is not possible when using the procedure discussed in Section 2.5

[12] In fact, BT introduce a generalization of the 'multiple comparisons with the best' approach that 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.

of the procedure is substantially increased (effectively, the number of comparisons is reduced from $\binom{k+1}{2}$ to k).

We begin by introducing some further notation. Let [1], [2], ..., $[k+1]$ be 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 that is *estimated* to be largest, and not necessarily the 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 highest 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 γ here is the smallest value satisfying

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

where [1*], [2*], ..., [(k + 1)*] are random indices such that

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

That is, $L^{*b}_{n,[1]}$ is the lower endpoint of the uncertainty interval for the parameter that is estimated to be largest in the b^{th} bootstrap sample *after re-centring*.

Simulation evidence presented both in BT and in our Online Supplement Section S2 suggests that the choice of γ 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 2.4. 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 γ, 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 1b) 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 that 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 significance of δ_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

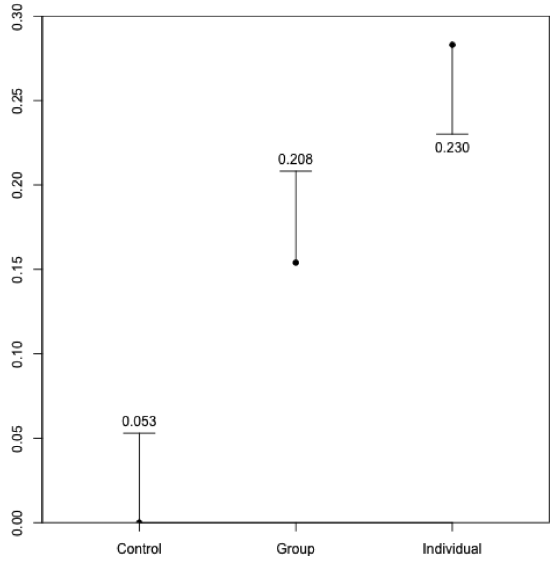


Figure 3. Performance pay example: overlap procedure modified to identify the 'best' treatment.

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 3.1, and only the modified procedure in the empirical example in Section 3.2.

3. ADDITIONAL EMPIRICAL EXAMPLES

3.1. Matching Grants in Charitable Giving

Karlan and List (2007), hereafter KL, conducted a large-scale field experiment to examine the effect of matching grants on charitable giving.[13] Matching grants are schemes in which an individual's 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 that made no mention of a matching grant. The remaining ('treated') donors received letters that 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).

[13] Data for this paper are available for download from: http://www.aeaweb.org/aer/data/dec07/20060421_data.zip

Although KL consider two outcomes, response (a binary variable) and amount given, we focus here solely on the latter.[14] 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 (2.1)—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 and White, 1985). Given a nominal FWER of $\alpha = 0.05$ and 999 replications of the wild bootstrap, we obtain a value of 2.406 for γ using the unmodified overlap procedure (this is the value obtained after the first iteration; no further refinement was possible).

Figure 4 displays our uncertainty intervals centred 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 for each of the $\binom{36+1}{2} = 666$ relevant pairwise parameter comparisons. A histogram of these T-statistics is shown in Figure 5 (a complete listing of these T-statistics is given in Online Supplement Section S4). It is interesting to note that 17 of these T-statistics fall outside of the interval $[-1.960, 1.960]$.[15] 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.

3.2. *Student Achievement Programmes*

Angrist, Lang, and Oreopoulos (2009), hereafter ALO, conducted a field experiment at a large university in Canada in order to examine programmes aimed at improving students' academic performance.[16] 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 (grade point average), which was measured at the end of first-year and again at the end of second-year. In order to examine the effects of the different treatments in this case, ALO estimate

[14] 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.

[15] It is important to emphasize 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.

[16] This dataset is publicly available from: https://www.aeaweb.org/aej-applied/data/2007-0062_data.zip

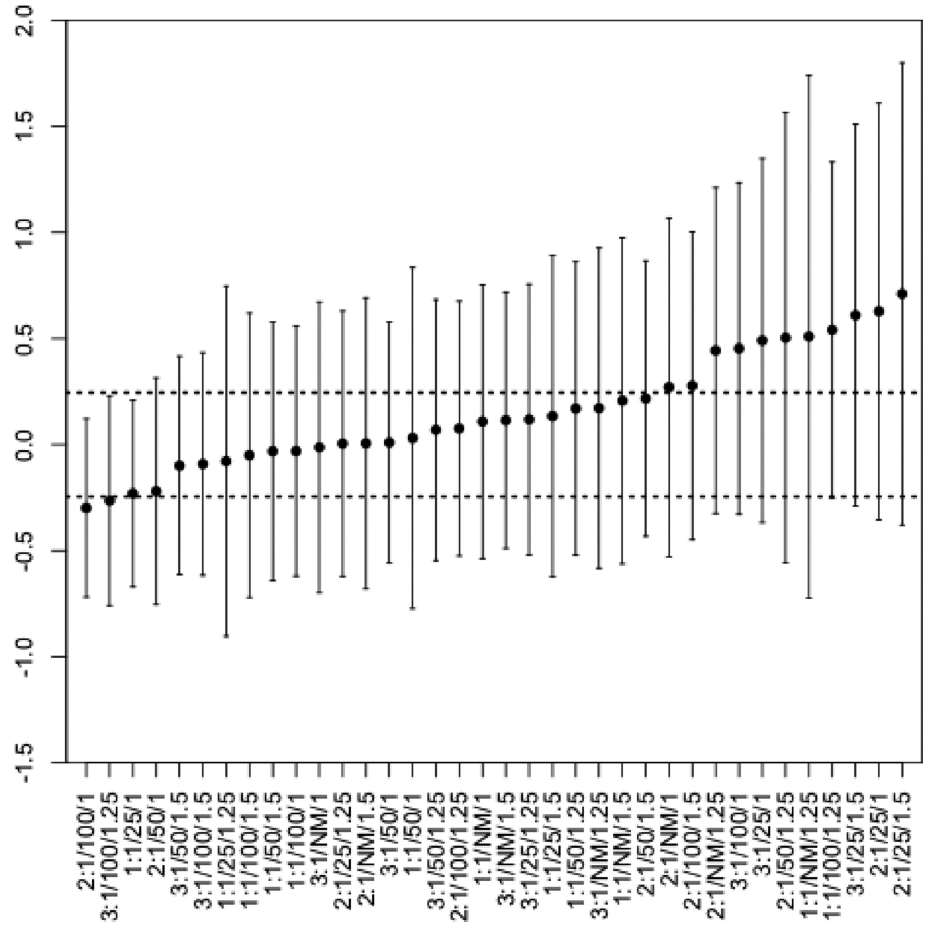


Figure 4. Charitable giving example

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.

the following model:

$$\text{GPA}_i = \beta_0 + \delta_1 \text{PF}_i + \delta_2 \text{PS}_i + \delta_3 \text{PFS}_i + \mathbf{X}_i' \eta + V_i, \tag{3.9}$$

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 **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

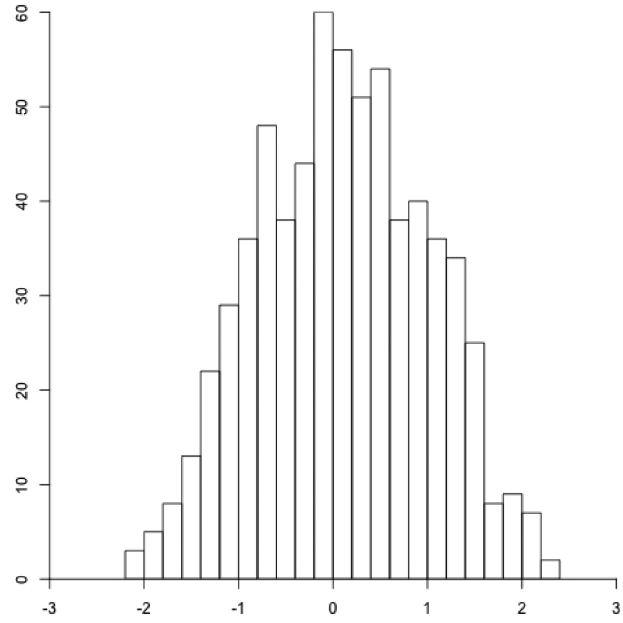


Figure 5. Histogram of T-statistics in charitable giving example.

education. ALO treat participation in the three treatments as endogenous, and use *assignment* to these groups as instruments.[17]

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 rewrite the above model in the form of model (2.4), where β_0 is multiplied by an indicator variable for membership in the control group (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). For simplicity, however, we centre 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 γ using 999 replications of the wild cluster bootstrap (which we modified for 2SLS following the approach of Davidson and MacKinnon, 2010). Figure 6 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 $\tilde{L}_{n,[1]} > \tilde{U}_{n,[s]}$, 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

[17] 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.

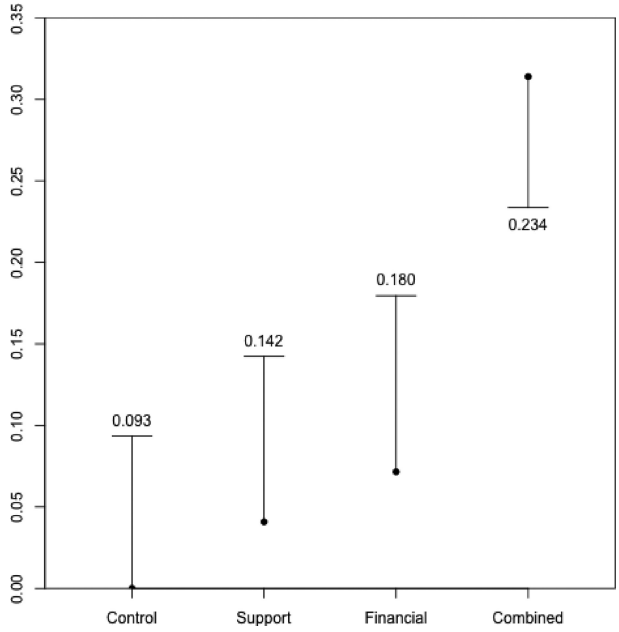


Figure 6. Student achievement example.

treatment effects is zero, and conclude that only δ_3 is positive (each test is conducted at the 5% nominal level, without any consideration of the FWER).[18]

4. 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 heterogeneous treatment effects (i.e., situations where different types of individuals may respond differently to the *same* treatment). A natural extension of our approach would be to incorporate such heterogeneous treatment effects. We leave this to future work.

ACKNOWLEDGEMENTS

The authors would like to thank Co-editor Victor Chernozhukov, as well as Jobu Babin, Chris Bennett, Otavio Camargo-Bartalotti, Steve Lehrer, James MacKinnon, Vincent Pohl, audience members at several seminars and conferences, and, finally, an anonymous referee for helpful comments. We also thank the authors of the empirical papers revisited here for making their data publicly available. This research was supported in part by a grant from Social Sciences and Humanities Research Council.

[18] ALO do informally compare estimates of different treatment effects in other parts of their paper. For example, on p. 14 they state that 'the [intention-to-treat] estimates for women suggest the combination of services and fellowships ...had a larger impact than fellowships alone'.

REFERENCES

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: a reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association 103*, 1481–95.

Angrist, J., D. Lang and P. Oreopoulos (2009). Incentives and services for college achievement: evidence from a randomized trial. *American Economic Journal: Applied Economics 1*, 136–63.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B 57*, 289–300.

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

Cameron, A. C., J. B. Gelbach and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics 90*, 414–27.

Davidson, R. and J. G. MacKinnon (2010). Wild bootstrap tests for IV regression. *Journal of Business and Economic Statistics 28*, 128–144.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association 50*, 1096–121.

Dunnett, C. W. and A. C. Tamhane (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statistics in Medicine 10*, 939–47.

Fink, G., M. McConnell and S. Vollmer (2014). Testing for heterogeneous treatment effects in experimental data: false discovery risks and correction procedures. *Journal of Development Effectiveness 6*, 44–57.

Gu, J. and S. Shen (2018). Oracle and adaptive false discovery rate controlling methods for one-sided testing: theory and application in treatment effect evaluation. *Econometrics Journal 21*, 11–35.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Horrace, W. C. and P. Schmidt (2000). Multiple comparisons with the best, with economic applications. *Journal of Applied Econometrics 15*, 1–26.

Hsu, J. C. (1981). Simultaneous confidence intervals for all distances from the 'best'. *Annals of Statistics 9*, 1026–34.

Hsu, J. C. (1984). Constrained simultaneous confidence intervals for multiple comparisons with the best. *Annals of Statistics 12*, 793–1150.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. Boca Raton, FL: Chapman & Hall/CRC.

Karlan, D. and J. A. List (2007). Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review 97*(5), 1774–93.

Lee, S. and A. M. Shaikh (2014). Multiple testing and heterogenous treatment effects: re-evaluating the effect of Progresa on school enrollment. *Journal of Applied Econometrics 29*(4), 612–26.

Lehrer, S. F., R. V. Pohl and K. Song (2018). Multiple testing and the distributional effects of accountability incentives in education. MPRA Paper 89532, University Library of Munich, Germany.

List, J. A., A. M. Shaikh and Y. Xu (2019).Multiple hypothesis testing in experimental economics. Forthcoming in *Experimental Economics*.

MacKinnon, J. G. and H. White (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics 29*, 305–25.

Muralidharan, K. and V. Sundararaman (2011). Teacher performance pay: experimental evidence from India. *Journal of Political Economy 119*, 39–77.

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

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

Tukey, J. W. (1953). The problem of multiple comparisons. In *The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948–1983*, pp. 1–300. New York, NY: Chapman and Hall.

Young, A. (2019). Channelling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results. Forthcoming in *Quarterly Journal of Economics*, *134*, (2), 557–598.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix
Replication Package

Co-editor Victor Chernozhukov handled this manuscript.