Experimental Economics (2019) 22:773–793
https://doi.org/10.1007/s10683-018-09597-5

ORIGINAL PAPER

Multiple hypothesis testing in experimental economics

John A. List[1] · Azeem M. Shaikh[1] · Yang Xu[1]

Received: 9 October 2017 / Revised: 10 November 2018 / Accepted: 19 November 2018 /
Published online: 29 January 2019

Abstract

The analysis of data from experiments in economics routinely involves testing multiple null hypotheses simultaneously. These different null hypotheses arise naturally in this setting for at least three different reasons: when there are multiple outcomes of interest and it is desired to determine on which of these outcomes a treatment has an effect; when the effect of a treatment may be heterogeneous in that it varies across subgroups defined by observed characteristics and it is desired to determine for which of these subgroups a treatment has an effect; and finally when there are multiple treatments of interest and it is desired to determine which treatments have an effect relative to either the control or relative to each of the other treatments. In this paper, we provide a bootstrap-based procedure for testing these null hypotheses simultaneously using experimental data in which simple random sampling is used to assign treatment status to units. Using the general results in Romano and Wolf (Ann Stat 38:598–633, 2010), we show under weak assumptions that our procedure (1) asymptotically controls the familywise error rate—the probability of one or more false rejections—and (2) is asymptotically balanced in that the marginal probability of rejecting any true null hypothesis is approximately equal in large samples. Importantly, by incorporating information about dependence ignored in classical multiple testing procedures, such as the Bonferroni and Holm corrections, our procedure has much greater ability to detect truly false null hypotheses. In the presence of multiple treatments, we additionally show how to exploit logical restrictions across null hypotheses to further improve power. We illustrate our methodology by revisiting the study by Karlan and List (Am Econ Rev 97(5):1774–1793, 2007) of why people give to charitable causes.

Keywords Experiments · Multiple hypothesis testing · Multiple treatments · Multiple outcomes · Multiple subgroups · Randomized controlled trial · Bootstrap · Balance

JEL Classification C12 · C14

Documentation of our procedures and our Stata and Matlab code can be found at https://github.com/seidelj/mht.

Extended author information available on the last page of the article

1 Introduction

Multiple hypothesis testing simply refers to any instance in which more than one null hypothesis is tested simultaneously. While this problem is pervasive throughout all empirical work in economics, we focus on the analysis of data from experiments in economics. In this setting, different null hypotheses arise naturally for at least three different reasons: when there are multiple outcomes of interest and it is desired to determine on which of these outcomes a treatment has an effect; when the effect of a treatment may be heterogeneous in that it varies across subgroups defined by observed characteristics (e.g., gender or age) and it is desired to determine for which of these subgroups a treatment has an effect; and finally when there are multiple treatments of interest and it is desired to determine which treatments have an effect relative to either the control or relative to each of the other treatments.

Testing multiple null hypotheses for each of these three reasons is ubiquitous in the analysis of experimental data. Anderson (2008), for example, reports that 84% of experiments published from 2004 to 2006 in a set of social sciences field journals examine five or more outcomes simultaneously and 61% examine ten or more outcomes simultaneously. Specific examples include many studies of early childhood interventions, such as the Abecedarian and Perry pre-school programs, which collected data on a large variety of outcomes pertaining to educational attainment, employment, and criminal behavior, among others. Similarly, Fink et al. (2014) report that 76% of field experiments published in leading economics journals examine multiple subgroups and 29% examine ten or more subgroups. Specific examples include analyses of how the effects of competition may vary by gender (Gneezy et al. 2003; Niederle and Vesterlund 2007; Flory et al. 2015b) or age (Sutter and Glätzle-Rützler 2014; Flory et al. 2015a). Multiple treatments are also commonplace in experiments. For instance, the recent economics literature has studied how different incentive schemes affect a variety of outcomes including worker productivity (Hossain and List 2012), child food choice and consumption (List and Samek 2015), and educational performance (Levitt et al. 2012).

With a few exceptions, some of which we note below, it is uncommon for the analyses of these data to account for the multiple hypothesis testing. As a result, the probability of a false rejection may be much higher than desired. To illustrate this point, consider testing N null hypotheses simultaneously. Suppose that for each null hypothesis a p value is available whose distribution is uniform on the unit interval when the corresponding null hypothesis is true. Suppose further that all null hypotheses are true and that the p values are independent. In this case, if we were to test each null hypothesis in the usual way at level $\alpha \in (0, 1)$, then the probability of one or more false rejections equals $1 - (1 - \alpha)^N$, which may be much greater than α and in fact tends rapidly to one as N increases. For instance, with $\alpha = 0.05$, it equals 0.226 when $N = 5$, equals 0.401 when $N = 10$ and 0.994 when $N = 100$. In order to control the probability of a false rejection, it is therefore important to account appropriately for multiplicity of null hypotheses being tested.

In this paper, we provide a bootstrap-based procedure for testing these null hypotheses simultaneously using experimental data in which simple random sampling is used to assign treatment status to units. Formally, we establish our results by applying the general results in Romano and Wolf (2010). In particular, we show under weak assumptions that our procedure (1) asymptotically controls the familywise error rate—the probability of one or more false rejections—and (2) is asymptotically balanced in that the the marginal probability of rejecting any true null hypothesis is approximately equal in large samples. Importantly, by incorporating information about dependence ignored in classical multiple testing procedures, such as the Bonferroni (1935) and Holm (1979) corrections, our procedure has much greater ability to detect truly false null hypotheses. In the presence of multiple treatments, we additionally show how to exploit logical restrictions across null hypotheses to further improve power. See Remark 3.7 for further discussion of this point.

As mentioned previously, it is uncommon in the experimental economics literature for authors to account for the multiplicity of null hypotheses being tested. Some notable exceptions include Kling et al. (2007), who use a more restrictive resampling-based multiple testing procedure due to Westfall and Young (1993) and Anderson (2008), Heckman et al. (2010), Heckman et al. (2011), and Lee and Shaikh (2014), who combine randomization methods with results in Romano and Wolf (2005) to construct multiple testing procedures with finite-sample validity for testing a more restrictive family of null hypotheses. Perhaps most importantly, none of these papers consider null hypotheses emerging due to multiple treatments, which, as noted above, is a very common occurrence in experiments in economics.

The remainder of our paper is organized as follows. In Sect. 2, we introduce our setup and notation as well as the assumptions under which we will establish the validity of our multiple testing procedure. Section 3 describes our multiple testing procedure and establishes its validity. In Sect. 4, we apply our methodology to data originally presented in Karlan and List (2007), who study the economics of charity by measuring, among other things, the effectiveness of a matching grant on charitable giving. Section 5 concludes. Proofs of all results can be found in "Appendix".

2 Setup and notation

For $k \in \mathcal{K}$, let $Y_{i,k}$ denote the (observed) kth outcome of interest for the ith unit, D_i denote treatment status for the ith unit, and Z_i denote observed, baseline covariates for the ith unit. Further denote by $\mathcal{D}$ and $\mathcal{Z}$ the supports of D_i and Z_i, respectively. For $d \in \mathcal{D}$, let $Y_{i,k}(d)$ be the kth potential outcome for the ith unit if treatment status were (possibly counterfactually) set equal to d. As usual, the kth observed outcome and kth potential outcome are related to treatment status by the relationship

$$Y_{i,k} = \sum_{d \in \mathcal{D}} Y_{i,k}(d) I\{D_i = d\}.$$

It is useful to introduce the shorthand notation $Y_i = (Y_{i,k} : k \in \mathcal{K})$ and $Y_i(d) = (Y_{i,k}(d) : k \in \mathcal{K})$. We assume that $((Y_i(d) : d \in \mathcal{D}), D_i, Z_i), i = 1, \ldots, n$ are i.i.d. with distribution $Q \in \Omega$, where our requirements on Ω are specified below.

It follows that the observed data $(Y_i, D_i, Z_i), i = 1, \dots, n$ are i.i.d. with distribution $P = P(Q)$. Denote by $\hat{P}_n$ the empirical distribution of the observed data.

The family of null hypotheses of interest is indexed by

$$s \in \mathcal{S} \subseteq \{(d, d', z, k) : d \in \mathcal{D}, d' \in \mathcal{D}, z \in \mathcal{Z}, k \in \mathcal{K}\}.$$

For each $s \in \mathcal{S}$, define

$$\omega_s = \{Q \in \Omega : E_Q\big[Y_{i,k}(d) - Y_{i,k}(d') | Z_i = z\big] = 0\}.$$

Using this notation, the family of null hypotheses of interest is given by

$$H_s : Q \in \omega_s \text{ for } s \in \mathcal{S}. \tag{1}$$

In other words, the sth null hypothesis specifies that the average effect of treatment d on the kth outcome of interest for the subpopulation where $Z_i = z$ equals the average effect of treatment d' on the kth outcome of interest for the subpopulation where $Z_i = z$. For later use, denote by $\mathcal{S}_0(Q)$ the subset of $\mathcal{S}$ corresponding to true null hypotheses, i.e.,

$$\mathcal{S}_0(Q) = \{s \in \mathcal{S} : Q \in \omega_s\}.$$

Our goal is to construct a procedure for testing these null hypotheses in a way that ensures asymptotic control of the familywise error rate for each $Q \in \Omega$. More precisely, we require for each $Q \in \Omega$ that

$$\limsup_{n \to \infty} FWER_Q \leq \alpha \tag{2}$$

for a pre-specified value of $\alpha \in (0, 1)$, where

$$FWER_Q = Q\{\text{reject any } H_s \text{ with } s \in \mathcal{S}_0(Q)\}. \tag{3}$$

The notation $FWER_Q$ is intended to reflect the fact that the quantity on the right-hand-side of (3) is the familywise error rate computed under Q. We additionally require that the testing procedure is "balanced" in that for each $Q \in \Omega$,

$$\lim_{n \to \infty} Q\{\text{reject } H_s\} = \lim_{n \to \infty} Q\{\text{reject } H_{s'}\} \text{ for any } s \text{ and } s' \text{ in } \mathcal{S}_0(Q). \tag{4}$$

We impose the requirement of "balance" to avoid situations where some (true) null hypotheses may be more likely to be rejected than other (true) null hypotheses for reasons that are viewed as undesirable, such as some outcomes taking on much larger values than other outcomes.

We now describe our main requirements on Ω. The assumptions make use of the notation

$$\begin{aligned} \mu_{k|d,z}(Q) =& E_Q[Y_{i,k}(d) | D_i = d, Z_i = z] \\ \sigma^2_{k|d,z}(Q) =& \mathrm{Var}_Q[Y_{i,k}(d) | D_i = d, Z_i = z]. \end{aligned}$$

Assumption 2.1 For each $Q \in \Omega$,

$$(Y_i(d) : d \in \mathcal{D}) \perp\!\!\!\perp D_i | Z_i$$

under Q.

Assumption 2.2 For each $Q \in \Omega$, $k \in \mathcal{K}$, $d \in \mathcal{D}$ and $z \in \mathcal{Z}$,

$$0 < \sigma^2_{k|d,z}(Q) = \mathrm{Var}_Q[Y_{i,k}(d)|D_i = d, Z_i = z] < \infty.$$

Assumption 2.3 For each $Q \in \Omega$, there is $\epsilon > 0$ such that

$$Q\{D_i = d, Z_i = z\} > \epsilon \tag{5}$$

for all $d \in \mathcal{D}$ and $z \in \mathcal{Z}$.

Assumption 2.1 simply requires that treatment status was randomly assigned. Assumption 2.2 is a mild non-degeneracy requirement. Assumption 2.3 simply requires that both D_i and Z_i are discrete random variables (with finite supports).

Remark 2.1 Note that we have assumed in particular that treatment status $D_i, i = 1, \ldots, n$ is i.i.d. While this assumption accommodates situations in which treatment status is assigned according to simple random sampling, it does not accommodate more complicated treatment assignment rules, such as those in which treatment status is assigned in order to "balance" baseline covariates among the subsets of individuals with different treatment status. For a discussion of such treatment assignment rules and the implications for inference about the average treatment effect, see Bugni et al. (2015). □

Remark 2.2 When $\mathcal{S}$ is very large, requiring control of the familywise error rate may significantly limit the ability to detect genuinely false null hypotheses. For this reason, it may be desirable in such situations to relax control of the familywise error rate in favor of generalized error rates that penalize false rejections less severely. Examples of such error rates include: the m-familywise error rate, defined to be the probability of m or more false rejections; the tail probability of the false discovery proportion, defined to be the fraction of false rejections (understood to be zero if there are no rejections at all); and the false discovery rate, defined to be the expected value of false discovery proportion. Control of the m-familywise error rate and the tail probability of the false discovery proportion using resampling are discussed in Romano et al. (2008b) and Romano and Wolf (2010). For procedures based only on (multiplicity-unadjusted) p values, see Lehmann and Romano (2005), Romano and Shaikh (2006a, b). For resampling-based control of the false discovery rate, see Romano et al. (2008a). □

3 A stepwise multiple testing procedure

In this section, we describe a stepwise multiple testing procedure for testing (1) in a way that satisfies (2) and (4) for any $Q \in \Omega$. In order to do so, we first require some additional notation. To this end, first define the "unbalanced" test statistic for H_s,

$$T_{s,n} = \sqrt{n}\left| \frac{1}{n_{d,z}} \sum_{1 \le i \le n : D_i = d, Z_i = z} Y_{i,k} - \frac{1}{n_{d',z}} \sum_{1 \le i \le n : D_i = d', Z_i = z} Y_{i,k} \right|, \tag{6}$$

and its re-centered version

$$\tilde{T}_{s,n}(P) = \sqrt{n}\left|\frac{1}{n_{d,z}}\sum_{1\le i\le n:D_i=d,Z_i=z}(Y_{i,k}-\tilde{\mu}_{k|d,z}(P)) - \frac{1}{n_{d',z}}\sum_{1\le i\le n:D_i=d',Z_i=z}(Y_{i,k}-\tilde{\mu}_{k|d',z}(P))\right|, \tag{7}$$

where

$$\tilde{\mu}_{k|d,z}(P) = E_P[Y_{i,k}|D_i = d, Z_i = z].$$

Next, for $s \in \mathcal{S}$, define

$$J_n(x, s, P) = P\left\{\tilde{T}_{s,n}(P) \le x\right\}.$$

Note that $J_n(x, s, P)$ is simply the distribution of (6) when H_s is true. In order to achieve "balance," rather than reject H_s for large values of $T_{s,n}$, we reject H_s for large values of

$$J_n(T_{s,n}, s, \hat{P}_n). \tag{8}$$

Note that (8) is simply one minus a (multiplicity-unadjusted) bootstrap p value for testing H_s based on $T_{s,n}$. Finally, for $\mathcal{S}' \subseteq \mathcal{S}$, let

$$L_n(x, \mathcal{S}', P) = P\left\{\max_{s\in\mathcal{S}'} J_n(\tilde{T}_{s,n}(P), s, P) \le x\right\}.$$

Note that $L_n(x, \mathcal{S}', P)$ is the distribution the maximum of (8) over $s \in \mathcal{S}'$ when H_s is true for all $s \in \mathcal{S}'$. Using this notation, we may describe our proposed stepwise multiple testing procedure as follows:

Algorithm 3.1

Step 0. *Set* $\mathcal{S}_1 = \mathcal{S}$.

$\vdots$

Step *j. If* $\mathcal{S}_j = \emptyset$ *or*

$$\max_{s\in\mathcal{S}_j} J_n(T_{s,n}, s, \hat{P}_n) \le L_n^{-1}(1-\alpha, \mathcal{S}_j, \hat{P}_n),$$

then stop. Otherwise, reject any H_s *with* $J_n(T_{s,n}, s, \hat{P}_n) > L_n^{-1}(1-\alpha, \mathcal{S}_j, \hat{P}_n)$, *set*

$$\mathcal{S}_{j+1} = \{s \in \mathcal{S}_j : J_n(T_{s,n}, s, \hat{P}_n) \le L_n^{-1}(1-\alpha, \mathcal{S}_j, \hat{P}_n)\},$$

and continue to the next step.

$\vdots$

The following theorem describes the asymptotic behavior of our proposed multiple testing procedure.

Theorem 3.1 *Consider the procedure for testing* (1) *it given by Algorithm* 3.1. *Under Assumptions* 2.1–2.3, *Algorithm* 3.1 *satisfies* (2) *and* (4) *for any* $Q \in \Omega$.

Remark 3.1 If $\mathcal{S} = \{s\}$, i.e., $\mathcal{S}$ is a singleton, then the familywise error rate is simply the usual probability of a Type I error. Hence, Algorithm 3.1 provides asymptotic control of the probability of a Type I error. In this case, Algorithm 3.1 is equivalent to the usual bootstrap test of H_s, i.e., the test that rejects H_s whenever $T_{s,n} > J_n^{-1}(1-\alpha, s, \hat{P}_n)$. □

Remark 3.2 As noted above, $\hat{p}_{s,n} = 1 - J_n(T_{s,n}, s, \hat{P}_n)$ may be interpreted as a bootstrap p value for testing H_s. Indeed, for any $Q \in \omega_s$, it is possible to show that

$$\limsup_{n\to\infty} Q\{\hat{p}_{s,n} \le u\} \le u$$

for any $0 < u < 1$. A crude solution to the multiplicity problem would therefore be to apply a Bonferroni or Holm correction to these p values. By replacing $L_n^{-1}(1-\alpha, \mathcal{S}_j, \hat{P}_n)$ with a suitable choice of critical value, it is possible to describe both the Bonferroni and Holm corrections in terms of Algorithm 3.1. The Bonferroni correction may be obtained by applying Algorithm 3.1 with $1 - \frac{\alpha}{|\mathcal{S}|}$ in place of $L_n^{-1}(1-\alpha, \mathcal{S}_j, \hat{P}_n)$, whereas the Holm correction, first described in Holm (1979), may be obtained by applying Algorithm 3.1 with $1 - \frac{\alpha}{|\mathcal{S}_j|}$ in place of $L_n^{-1}(1-\alpha, \mathcal{S}_j, \hat{P}_n)$. Such approaches would indeed satisfy (2), as desired, but implicitly rely upon a "least favorable" dependence structure among the p values. To the extent that the true dependence structure differs from this "least favorable" one, improvements may be possible. Algorithm 3.1 uses the bootstrap to incorporate implicitly information about the dependence structure when deciding which null hypotheses to reject. In fact, Algorithm 3.1 will always reject at least as many null hypotheses as these procedures. □

Remark 3.3 Implementation of Algorithm 3.1 typically requires approximating the quantities $J_n(x, s, \hat{P}_n)$ and $L_n(x, \mathcal{S}', \hat{P}_n)$ using simulation. As noted by Romano and Wolf (2010), doing so does not require nested bootstrap simulations. To explain further, for $b = 1, \ldots, B$, draw a sample of size n from $\hat{P}_n$ and denote by $\tilde{T}^{*,b}_{s,n}(\hat{P}_n)$ the quantity $\tilde{T}_{s,n}(P)$ using the bth resample and $\hat{P}_n$ as an estimate of P. Then, $J_n(x, s, \hat{P}_n)$ may be approximated as

$$\hat{J}_n(x, s, \hat{P}_n) = \frac{1}{B} \sum_{1 \le b \le B} I\{\tilde{T}^{*,b}_{s,n}(\hat{P}_n) \le x\}$$

and $L_n(x, \mathcal{S}', \hat{P}_n)$ may be approximated as

$$\hat{L}_n(x, \mathcal{S}', \hat{P}_n) = \frac{1}{B} \sum_{1 \le b \le B} I\left\{\max_{s \in \mathcal{S}'} \hat{J}_n(T^{*,b}_{s,n}(\hat{P}_n), s, \hat{P}_n) \le x\right\}.$$

In particular, the same set of bootstrap resamples may be used in the two approximations. □

Remark 3.4 In terms of higher-order asymptotic properties, it is often desirable to studentize, i.e., to replace $T_{s,n}$ and $\tilde{T}_{s,n}(P)$, respectively, with

$$T^{\text{stud}}_{s,n} = \frac{T_{s,n}}{\sqrt{n \cdot \left(\frac{\tilde{\sigma}^2_{k|d,z}(\hat{P}_n)}{n_{d,z}} + \frac{\tilde{\sigma}^2_{k|d',z}(\hat{P}_n)}{n_{d',z}} \right)}}$$

$$\tilde{T}^{\text{stud}}_{s,n}(P) = \frac{\tilde{T}_{s,n}(P)}{\sqrt{n \cdot \left(\frac{\tilde{\sigma}^2_{k|d,z}(\hat{P}_n)}{n_{d,z}} + \frac{\tilde{\sigma}^2_{k|d',z}(\hat{P}_n)}{n_{d',z}} \right)}},$$

where

$$\tilde{\sigma}^2_{k|d,z}(P) = \text{Var}_P[Y_{i,k}|D_i = d, Z_i = z].$$

Theorem 3.1 continues to hold with these changes. □

Remark 3.5 In some cases, it may be of interest to consider one-sided null hypotheses, e.g., $H^-_s : P \in \omega^-_s$, where

$$\omega^-_s = \{Q \in \Omega : E_Q[Y_{i,k}(d) - Y_{i,k}(d')|Z_i = z] \le 0\} \tag{9}$$

In this case, it suffices simply to replace T_{s_n} and $\tilde{T}_{s_n}(P)$, respectively, with $T^-_{s,n}$ and $\tilde{T}^-_{s,n}(P)$, which are, respectively, defined as in (6) and (7), but without the absolute values. An analogous modification can be made for null hypotheses $H^+_s : P \in \omega^+_s$, where ω^+_s is defined as in (9), but with the inequality reversed. □

Remark 3.6 Note that a multiplicity-adjusted p value for H_s, $\hat{p}^{\text{adj}}_{s,n}$, may be computed simply as the smallest value of α for which H_s is rejected in Algorithm 3.1. □

Remark 3.7 It is possible to improve Algorithm 3.1 by exploiting transitivity (i.e., $\mu_{k|d,z}(Q) = \mu_{k|d',z}(Q)$ and $\mu_{k|d',z}(Q) = \mu_{k|d'',z}(Q)$ implies that $\mu_{k|d,z}(Q) = \mu_{k|d'',z}(Q)$). To this end, for $\mathcal{S}' \subseteq \mathcal{S}$, define

$$\mathbb{S}(\mathcal{S}') = \{\mathcal{S}'' \subseteq \mathcal{S}' : \exists\, Q \in \Omega \text{ s.t. } \mathcal{S}'' = \mathcal{S}_0(Q)\}$$

and replace $L^{-1}_n(1-\alpha, \mathcal{S}_j, \hat{P}_n)$ in Algorithm 3.1 with

$$\max_{\tilde{\mathcal{S}} \in \mathbb{S}(\mathcal{S}_j)} L^{-1}_n(1-\alpha, \tilde{\mathcal{S}}, \hat{P}_n).$$

With this modification to Algorithm 3.1, Theorem 3.1 remains valid. Note that this modification is only non-trivial when there are more than two treatments and may be computationally prohibitive when there are more than a few treatments. □

Remark 3.8 Note that we only require that the familywise error rate is asymptotically no greater than α for each $Q \in \Omega$. By appropriately strengthening the assumptions of Theorem 3.1, it is possible to show that Algorithm 3.1 satisfies

$$\limsup_{n \to \infty} \sup_{Q \in \Omega} FWER_Q \leq \alpha.$$

In particular, it suffices to replace Assumption 2.2 with a mild uniform integrability requirement and require in Assumption 2.3 that there exists $\epsilon > 0$ for which (5) holds for all $Q \in \Omega$, $d \in \mathcal{D}$ and $z \in \mathcal{Z}$. Relevant results for establishing this claim can be found in Romano and Shaikh (2012), Bhattacharya et al. (2012), and Machado et al. (2013). □

4 Empirical applications

In this section, we apply our methodology to data originally presented in Karlan and List (2007), who use direct mail solicitations targeted to previous donors of a nonprofit organization to study the effectiveness of a matching grant on charitable giving. The sample includes all 50,083 individuals who had given to the organization at least once since 1991. Each individual was independently assigned with probability two-thirds to a treatment group (resulting in 33,396, or 67 percent of the sample, being treated) and with probability one-third to a control group (resulting in 16,687 subjects, or 33 percent of the sample, being untreated). Individuals in the treatment group were offered independently and with equal probability one of 36 possible matching grants whose terms varied along three dimensions: three possible values for the price ratio of the match, four possible values for the maximum size of the matching gift across all donations, and three possible values for the suggested donation amount. The possible values for the price ratio of the match were \$1:\$1, \$2:\$1, and \$3:\$1. Here, an \$X:\$1 ratio means that for every dollar the individual donates, the matching donor also contributes \$X. Hence, the charity receives \$X+1 for every \$1 the individual donates (subject to the maximum size of the matching gift across all donations). The possible values for the maximum matching grant amount were \$25,000, \$50,000, \$100,000, and "unstated." The possible values for the (individual-specific) suggested donation amounts were the individual's highest previous contribution, 1.25 times the highest previous contribution, and 1.50 times the highest previous contribution.

In the following three subsections, we first consider testing families of null hypotheses that emerge in this application due to multiple outcomes alone, multiple subgroups alone and multiple treatments alone. In the final subsection, we then consider testing the family of null hypotheses that emerges by combining all three considerations at the same time. In each case, we consider inference based on Theorem 3.1 using the studentized test statistics described in Remark 3.4. We also compare our results with those obtained using the classical Bonferroni and Holm multiple testing procedures. Stata and Matlab code used to produce these results can be found at the following address: https://github.com/seidelj/mht.

4.1 Multiple outcomes

Four outcomes of interest in Karlan and List (2007) are the response rate, dollars given not including the matching amount, dollars given including the matching amount, and the change in the amount given (not including the matching amount). A more detailed description of these variables can be found in Karlan and List (2007). Table 1 displays for each of these four outcomes of interest, the following five quantities: difference in means between treated and untreated groups, a (multiplicity-unadjusted) *p* value computed using Remark 3.1, a (multiplicity-adjusted) *p* value computed using Theorem 3.1, a (multiplicity-adjusted) *p* value obtained by applying Bonferroni to the (multiplicity-unadjusted) *p* values, a (multiplicity-adjusted) *p* value obtained by applying Holm to the (multiplicity-unadjusted) *p* values. Following Remark 3.2, the (multiplicity-adjusted) *p* values obtained by applying Bonferroni can be calculated simply by multiplying the (multiplicity-unadjusted) *p* values by the total number of hypotheses in Table 1. Similarly, the (multiplicity-adjusted) *p* values obtained by applying Holm can be calculated by the multiplying the smallest (multiplicity-unadjusted) *p* value (corresponding in this case to response rate) by the total number of hypotheses in Table 1, multiplying the second smallest (multiplicity-unadjusted) *p* value (corresponding in this case to dollars given including match) by one less than the total number of hypotheses in Table 1, and continuing in this fashion until multiplying the largest (multiplicity-unadjusted) *p* value (corresponding in this case to amount change) by one.

Before adjusting for the multiplicity of null hypotheses being tested, we find that the treatment has an effect on the response rate, dollars given not including the matching amount, and dollars given including the matching amount at the 5% significance level. Here, by treatment, we mean receiving any of the 36 possible matching grants. After adjusting for the multiplicity of null hypotheses being tested, however, we find that the effect of the treatment on dollars given not including the matching is no longer significant at the 5% significance level—instead, it is only significant at the 10% significance level. By comparing the last three columns in Table 1, we additionally see that the *p* values obtained by applying Theorem 3.1 are an improvement upon those obtained by applying Bonferroni or Holm.

Table 1 Multiple outcomes

Outcome	DI	*p* values			
		Unadj.	Multiplicity adj.		
		Remark 3.1	Theorem 3.1	Bonf.	Holm
Response rate	0.0042	0.0003***	0.0003***	0.0013***	0.0013***
Dollars given not including match	0.1536	0.0500*	0.0967*	0.2000	0.1000
Dollars given including match	2.0876	0.0003***	0.0003***	0.0013***	0.0010***
Amount change	6.3306	0.7200	0.7200	1.0000	0.7200

DI refers to difference in means. * and *** indicates that the corresponding *p* values less than 10% and 1%, respectively

4.2 Multiple subgroups

Four subgroups of interest in Karlan and List (2007) are red county in a red state, blue county in a red state, red county in a blue state, and blue county in a blue state. Red states are defined as states that voted for George W. Bush in the 2004 Presidential election and blue states are defined as states that voted for John Kerry in the same election. Red and blue counties are defined analogously. In this subsection, we examine how the effect of the treatment on the response rate varies across these subgroups. Table 2 displays for each of the four subgroups of interest, the same five quantities found in Table 1. Note that 105 out of the 50,083 individuals in our dataset do not have complete subgroup information. We treat these 105 individuals as a subgroup of no interest for our analysis.

Before adjusting for the multiplicity of null hypotheses being tested, we find that the treatment has an effect on two of the four subgroups at the 10% significance level. As before, here, by treatment, we mean receiving any of the 36 possible matching grants. After adjusting for the multiplicity of null hypotheses being tested, however, we find that the treatment only has an effect on one subgroup at the same significance level. By comparing the last three columns in Table 1, we again see that the *p* values obtained by applying Theorem 3.1 are an improvement upon those obtained by applying Bonferroni or Holm.

4.3 Multiple treatments

We now consider null hypotheses that emerge due to multiple treatments. We define three treatments corresponding to different values for the price ratio of the match: \$1:\$1, \$2:\$1, and \$3:\$1. Each treatment is understood to mean any of the 12 possible treatments with the same value for the price ratio of the match. We focus on dollars given not including the matching amount as the outcome of interest.

We first consider testing three null hypotheses corresponding to comparing each treatment with the control group. Table 3 displays for each of these three null hypotheses the same five quantities found in Table 1. Before adjusting for the multiplicity of null hypotheses being tested, we find that the treatment \$2:\$1

Table 2 Multiple subgroups

Subgroup	DI	*p* values			
		Unadj.	Multiplicity adj.		
		Remark 3.1	Theorem 3.1	Bonf.	Holm
Red county in a red state	0.0095	0.0003***	0.0003***	0.0013***	0.0013***
Blue county in a red state	0.0070	0.0503*	0.1427	0.2013	0.1510
Red county in a blue state	0.0016	0.4560	0.7017	1.0000	0.9120
Blue county in a blue state	0.0000	0.9920	0.9920	1.0000	0.9920

DI refers to difference in means. * and *** indicates that the corresponding *p* values less than 10% and 1%, respectively

Table 3 Multiple treatments (Comparing multiple treatments with a control)

Treatment/control groups	DI	*p* values			
		Unadj.	Multiplicity adj.		
		Remark 3.1	Theorem 3.1	Bonf.	Holm
Control versus 1:1	0.1234	0.2627	0.2627	0.7880	0.2627
Control versus 2:1	0.2129	0.0477**	0.1297	0.1430	0.1430
Control versus 3:1	0.1245	0.2060	0.3537	0.6180	0.4120

DI refers to difference in means. ** indicates that the corresponding *p* value less than 5%

has an effect at the 5% significance level on the outcome of interest. After adjusting for the multiplicity of null hypotheses being tested, however, we find that this effect is no longer significant even at the 10% significance level. By comparing the last three columns in Table 3, we again see that the *p* values obtained by applying Theorem 3.1 are an improvement upon those obtained by applying Bonferroni or Holm.

Next, we consider testing the six null hypotheses corresponding to all pairwise comparisons across the three treatments and the control group. Table 4 displays for each of these three null hypotheses the same five quantities found in Table 1. The results are qualitatively similar to those described above. Table 4 also displays a sixth quantity corresponding to the improvement in *p* values described in Remark 3.7 obtained by exploiting the logical restrictions among null hypotheses when there are multiple treatments. While in this case the improvement does not lead to any additional rejections of null hypotheses, we see that the difference in *p* values can in some cases be large. Note that this column is omitted from the previous table because Remark 3.7 results in no further improvements when solely comparing each treatment with the control group.

Table 4 Multiple treatments (All pairwise comparisons across multiple treatments and a control)

Treatment/control groups	DI	*p* values				
		Unadj.	Multiplicity adj.			
		Remark 3.1	Theorem 3.1	Remark 3.7	Bonf.	Holm
Control versus 1:1	0.1234	0.2627	0.5810	0.4973	1.0000	1.0000
Control versus 2:1	0.2129	0.0477**	0.1930	0.1930	0.2860	0.2860
Control versus 3:1	0.1245	0.2060	0.5533	0.4167	1.0000	1.0000
1:1 versus 2:1	0.0895	0.4627	0.7467	0.7467	1.0000	1.0000
1:1 versus 3:1	0.0011	0.9920	0.9920	0.9920	1.0000	0.9920
2:1 versus 3:1	0.0883	0.4633	0.6963	0.4633	1.0000	0.9267

DI refers to difference in means. ** indicates that the corresponding *p* value less than 5%

4.4 Multiple outcomes, subgroups, and treatments

More often than not, it is desired to test null hypotheses stemming from all three considerations above: multiple outcomes, multiple subgroups, and multiple treatments simultaneously. In this subsection, we consider the four outcome variables described in Sect. 4.1, the four subgroups described in Sect. 4.2, and the three treatments described in Sect. 4.3. Here, we only consider comparing each treatment with the control group. As a result there are 48 null hypotheses of interest.

For each of the 48 null hypotheses, Table 5 displays the same five quantities found in Table 1. Before adjusting for the multiplicity of null hypotheses being tested, we reject 21 null hypotheses at the 10% significance level. After adjusting for the multiplicity of null hypotheses being tested, however, we find that only 9 null hypotheses are rejected at the same significance level. It is worth noting that 7 of these 9 null hypotheses are related to the same outcome—dollars given including the matching amount. By comparing the last three columns in Table 5, we again see that the p values obtained by applying Theorem 3.1 are an improvement upon those obtained by applying Bonferroni or Holm.

5 Conclusion

In this paper, we have developed a procedure for testing simultaneously null hypotheses that emerge naturally when analyzing data from experiments because of some combination of the presences of multiple outcomes of interest, multiple subgroups of interest or multiple treatments. Using the general results in Romano and Wolf (2010), we have shown that our approach applies under weak assumptions to experiments in which individuals are assigned to treatments and control using simple random sampling. Notably, we show not only that our procedure has greater power than classical multiple testing procedures like Bonferroni and Holm, but have also shown how further improvements can be obtained in the presence of multiple treatments by exploiting the logical restrictions among null hypotheses. We have applied our methodology to data originally presented in Karlan and List (2007), who studied the effectiveness of a matching grant on charitable giving.

As we have argued in the introduction, it is commonplace to consider multiple null hypotheses when analyzing experimental data for one or more of the reasons mentioned above. It is, however, uncommon to account correctly for the multiplicity of null hypotheses under consideration, and, as a result, the probability of a false rejection may be much higher than desired. This failure to adjust inference procedures is almost certainly related to the “credibility” and “reproducibility” crises that plague not just the social sciences, but the sciences more generally. See, for example, Jennions and Moller (2002), Ioannidis (2005), Nosek et al. (2012), Bettis (2012), Maniadis et al. (2014) and Camerer et al. (2016). We believe the adoption of testing procedures like the one described in this paper will help address these concerns and, with this in mind, advocate that researchers at the very least report multiplicity-adjusted results alongside conventional multiplicity-unadjusted results. In some cases, such as when evaluating existing studies, it may be more convenient to compute a Bonferroni or Holm correction,

Table 5 Multiple outcomes, subgroups, and treatments

Outcome	Subgroup	Treatment/control groups	DI	*p* values			
				Unadj.	Multiplicity adj.		
				Remark 3.1	Theorem 3.1	Bonf.	Holm
Response rate	Red county in a red state	Control versus 1:1	0.0079	0.0217**	0.4643	1.0000	0.7583
Response rate	Red county in a red state	Control versus 2:1	0.0100	0.0017***	0.0480**	0.0800*	0.0733*
Response rate	Red county in a red state	Control versus 3:1	0.0107	0.0017***	0.0470**	0.0800*	0.0717*
Response rate	Blue county in a red state	Control versus 1:1	0.0024	0.5973	1.0000	1.0000	1.0000
Response rate	Blue county in a red state	Control versus 2:1	0.0080	0.0987*	0.8997	1.0000	1.0000
Response rate	Blue county in a red state	Control versus 3:1	0.0108	0.0247**	0.5000	1.0000	0.8387
Response rate	Red county in a blue state	Control versus 1:1	0.0003	0.9060	0.9990	1.0000	1.0000
Response rate	Red county in a blue state	Control versus 2:1	0.0010	0.7190	1.0000	1.0000	1.0000
Response rate	Red county in a blue state	Control versus 3:1	0.0034	0.2290	0.9953	1.0000	1.0000
Response rate	Blue county in a blue state	Control versus 1:1	0.0006	0.8667	1.0000	1.0000	1.0000
Response rate	Blue county in a blue state	Control versus 2:1	0.0026	0.5033	1.0000	1.0000	1.0000
Response rate	Blue county in a blue state	Control versus 3:1	0.0032	0.3740	1.0000	1.0000	1.0000
Dollars given not including match	Red county in a red state	Control versus 1:1	0.4260	0.0903*	0.9027	1.0000	1.0000
Dollars given not including match	Red county in a red state	Control versus 2:1	0.4097	0.0557*	0.7860	1.0000	1.0000
Dollars given not including match	Red county in a red state	Control versus 3:1	0.3214	0.0710*	0.8483	1.0000	1.0000
Dollars given not including match	Blue county in a red state	Control versus 1:1	0.0374	0.8950	1.0000	1.0000	1.0000
Dollars given not including match	Blue county in a red state	Control versus 2:1	0.4325	0.1853	0.9853	1.0000	1.0000
Dollars given not including match	Blue county in a red state	Control versus 3:1	0.5728	0.0933*	0.8983	1.0000	1.0000
Dollars given not including match	Red county in a blue state	Control versus 1:1	0.0256	0.8683	1.0000	1.0000	1.0000
Dollars given not including match	Red county in a blue state	Control versus 2:1	0.0928	0.5893	1.0000	1.0000	1.0000
Dollars given not including match	Red county in a blue state	Control versus 3:1	0.0243	0.8847	1.0000	1.0000	1.0000
Dollars given not including match	Blue county in a blue state	Control versus 1:1	0.0074	0.9747	0.9747	1.0000	0.9747
Dollars given not including match	Blue county in a blue state	Control versus 2:1	0.0380	0.8650	1.0000	1.0000	1.0000

Table 5 (continued)

Outcome	Subgroup	Treatment/control groups	DI	*p* values			
				Unadj.	Multiplicity adj.		
				Remark 3.1	Theorem 3.1	Bonf.	Holm
Dollars given not including match	Blue county in a blue state	Control versus 3:1	0.2173	0.2847	0.9997	1.0000	1.0000
Dollars given including match	Red county in a red state	Control versus 1:1	1.5042	0.0080***	0.2170	0.3840	0.3120
Dollars given including match	Red county in a red state	Control versus 2:1	2.5335	0.0010***	0.0247**	0.0480**	0.0450**
Dollars given including match	Red county in a red state	Control versus 3:1	3.2419	0.0003***	0.0003***	0.0160**	0.0160**
Dollars given including match	Blue county in a red state	Control versus 1:1	0.8370	0.0603*	0.8037	1.0000	1.0000
Dollars given including match	Blue county in a red state	Control versus 2:1	2.8217	0.0080***	0.2163	0.3840	0.3040
Dollars given including match	Blue county in a red state	Control versus 3:1	4.5776	0.0023***	0.0667*	0.1120	0.0957*
Dollars given including match	Red county in a blue state	Control versus 1:1	0.7737	0.0087***	0.2290	0.4160	0.3207
Dollars given including match	Red county in a blue state	Control versus 2:1	1.9283	0.0007***	0.0133**	0.0320**	0.0307**
Dollars given including match	Red county in a blue state	Control versus 3:1	2.5722	0.0003***	0.0003***	0.0160**	0.0157**
Dollars given including match	Blue county in a blue state	Control versus 1:1	0.9967	0.0133**	0.3240	0.6400	0.4800
Dollars given including match	Blue county in a blue state	Control versus 2:1	2.1371	0.0023***	0.0653*	0.1120	0.0933*
Dollars given including match	Blue county in a blue state	Control versus 3:1	2.1658	0.0020***	0.0577*	0.0960*	0.0840*
Amount change	Red county in a red state	Control versus 1:1	1.8252	0.1310	0.9497	1.0000	1.0000
Amount change	Red county in a red state	Control versus 2:1	0.5491	0.6443	1.0000	1.0000	1.0000
Amount change	Red county in a red state	Control versus 3:1	0.0681	0.9593	0.9987	1.0000	1.0000
Amount change	Blue county in a red state	Control versus 1:1	92.3221	0.4410	1.0000	1.0000	1.0000
Amount change	Blue county in a red state	Control versus 2:1	93.7227	0.4410	1.0000	1.0000	1.0000
Amount change	Blue county in a red state	Control versus 3:1	94.2640	0.4410	1.0000	1.0000	1.0000
Amount change	Red county in a blue state	Control versus 1:1	51.9652	0.4530	1.0000	1.0000	1.0000
Amount change	Red county in a blue state	Control versus 2:1	0.4450	0.6817	1.0000	1.0000	1.0000
Amount change	Red county in a blue state	Control versus 3:1	1.1372	0.2593	0.9973	1.0000	1.0000

Table 5 (continued)

Outcome	Subgroup	Treatment/control groups	DI	*p* values			
				Unadj.	Multiplicity adj.		
				Remark 3.1	Theorem 3.1	Bonf.	Holm
Amount change	Blue county in a blue state	Control versus 1:1	0.9294	0.4617	1.0000	1.0000	1.0000
Amount change	Blue county in a blue state	Control versus 2:1	0.2938	0.8277	1.0000	1.0000	1.0000
Amount change	Blue county in a blue state	Control versus 3:1	0.5147	0.6577	1.0000	1.0000	1.0000

DI refers to difference in means. *, **, and *** indicate that the corresponding *p* values less than 10%, 5%, and 1%, respectively which only requires knowledge of conventional multiplicity-unadjusted p values, rather than apply the methodology in this paper.

Acknowledgements We would like to thank Joseph P. Romano for helpful comments on this paper. We also thank Joseph Seidel for his excellent research assistance. The research of the second author was supported by National Science Foundation Grants DMS-1308260, SES-1227091, and SES-1530661.

Appendix

Proof of Theorem 3.1

First note that under Assumption 2.1, $Q \in \omega_s$ if and only if $P \in \tilde{\omega}_s$, where

$$\tilde{\omega}_s = \{P(Q) : Q \in \Omega, E_P[Y_{i,k}|D_i = d, Z_i = z] = E_P[Y_{i,k}|D_i = d', Z_i = z]\}.$$

The proof of this result now follows by verifying the conditions of Corollary 5.1 in Romano and Wolf (2010). In particular, we verify Assumptions B.1–B.4 in Romano and Wolf (2010).

In order to verify Assumption B.1 in Romano and Wolf (2010), let

$$T^*_{s,n}(P) = \sqrt{n}\left(\frac{1}{n_{d,z}} \sum_{1 \le i \le n : D_i = d, Z_i = z} (Y_{i,k} - \tilde{\mu}_{k|d,z}(P)) - \frac{1}{n_{d',z}} \sum_{1 \le i \le n : D_i = d', Z_i = z} (Y_{i,k} - \tilde{\mu}_{k|d',z}(P))\right),$$

and note that

$$T^*_n(P) = (T^*_{s,n}(P) : s \in \mathcal{S}) = f(A_n(P), B_n),$$

where

$$A_n(P) = \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} A_{n,i}(P),$$

with $A_{n,i}(P)$ equal to the $2|\mathcal{S}|$-dimensional vector formed by stacking vertically for $s \in \mathcal{S}$ the terms

$$\begin{pmatrix} (Y_{i,k} - \tilde{\mu}_{k|d,z}(P))I\{D_i = d, Z_i = z\} \\ (Y_{i,k} - \tilde{\mu}_{k|d',z}(P))I\{D_i = d', Z_i = z\} \end{pmatrix}, \tag{10}$$

and B_n is the $2|\mathcal{S}|$-dimensional vector formed by stacking vertically for $s \in \mathcal{S}$ the terms

$$\begin{pmatrix} \frac{1}{\frac{1}{n}\sum_{1 \le i \le n} I\{D_i = d, Z_i = z\}} \\ -\frac{1}{\frac{1}{n}\sum_{1 \le i \le n} I\{D_i = d', Z_i = z\}} \end{pmatrix}. \tag{11}$$

and $f : \mathbf{R}^{2|\mathcal{S}|} \times \mathbf{R}^{2|\mathcal{S}|} \to \mathbf{R}^{2|\mathcal{S}|}$ is the function of $A_n(P)$ and B_n whose sth argument for $s \in \mathcal{S}$ is given by the inner product of the sth pair of terms in $A_n(P)$ and the sth pair of terms in B_n, i.e., the inner product of (10) and (11). The weak law of large numbers and central limit theorem imply that

$$B_n \xrightarrow{P} B(P),$$

where $B(P)$ is the $2|\mathcal{S}|$-dimensional vector formed by stacking vertically for $s \in \mathcal{S}$ the terms

$$\begin{pmatrix} \frac{1}{P\{D_i=d,Z_i=z\}} \\ -\frac{1}{P\{D_i=d',Z_i=z\}} \end{pmatrix}.$$

Next, note that $E_P[A_{n,i}(P)] = 0$. Assumption 2.3 and the central limit theorem therefore imply that

$$A_n(P) \xrightarrow{d} N(0, V_A(P))$$

for an appropriate choice of $V_A(P)$. In particular, the diagonal elements of $V_A(P)$ are of the form

$$\tilde{\sigma}^2_{k|d,z}(P)P\{D_i = d, Z_i = z\}.$$

The continuous mapping theorem thus implies that

$$T_n^*(P) \xrightarrow{d} N(0, V(P))$$

for an appropriate variance matrix $V(P)$. In particular, the sth diagonal element of $V(P)$ is given by

$$\frac{\tilde{\sigma}^2_{k|d,z}(P)}{P\{D_i = d, Z_i = z\}} + \frac{\tilde{\sigma}^2_{k|d',z}(P)}{P\{D_i = d', Z_i = z\}}. \tag{12}$$

In order to verify Assumptions B.2–B.3 in Romano and Wolf (2010), it suffices to note that (12) is strictly greater than zero under our assumptions. Note that it is not required that $V(P)$ be non-singular for these assumptions to be satisfied.

In order to verify Assumption B.4 in Romano and Wolf (2010), we first argue that

$$T_n^*(P_n) \xrightarrow{d} N(0, V(P)) \tag{13}$$

under P_n for an appropriate sequence of distributions P_n for (Y_i, D_i, Z_i). To this end, assume that

(a) $P_n \xrightarrow{d} P$.
(b) $\tilde{\mu}_{k|d,z}(P_n) \to \tilde{\mu}_{k|d,z}(P)$.
(c) $B_n \xrightarrow{P_n} B(P)$.
(d) $\mathrm{Var}_{P_n}[A_{n,i}(P_n)] \to \mathrm{Var}_P[A_{n,i}(P)]$.

Under (a) and (b), it follows that $A_{n,i}(P_n) \xrightarrow{d} A_{n,i}(P)$ under P_n. By arguing as in Theorem 15.4.3 in Lehmann and Romano (2006) and using (d), it follows from the Lindeberg–Feller central limit theorem that

$$A_n(P_n) \xrightarrow{d} N(0, V_A(P))$$

under P_n. It thus follows from (c) and the continuous mapping theorem that (13) holds under P_n. Assumption B.4 in Romano and Wolf (2010) now follows simply by nothing that the Glivenko-Cantelli theorem, strong law of large numbers and continuous mapping theorem ensure that $\hat{P}_n$ satisfies (a)–(d) with probability one under P.

References

Anderson, M. (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*(484), 1481–1495.

Bettis, R. A. (2012). The search for asterisks: Compromised statistical tests and flawed theories. *Strategic Management Journal*, *33*(1), 108–113.

Bhattacharya, J., Shaikh, A. M., & Vytlacil, E. (2012). Treatment effect bounds: An application to swanganz catheterization. *Journal of Econometrics*, *168*(2), 223–243.

Bonferroni, C. E. (1935). *Il calcolo delle assicurazioni su gruppi di teste*. Rome: Tipografia del Senato.

Bugni, F., Canay, I., & Shaikh, A. (2015). Inference under covariate-adaptive randomization. Technical report, cemmap working paper, Centre for Microdata Methods and Practice.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Fink, G., McConnell, M., & Vollmer, S. (2014). Testing for heterogeneous treatment effects in experimental data: False discovery risks and correction procedures. *Journal of Development Effectiveness*, *6*(1), 44–57.

Flory, J. A., Gneezy, U., Leonard, K. L., & List, J. A. (2015a). Gender, age, and competition: The disappearing gap. Unpublished Manuscript.

Flory, J. A., Leibbrandt, A., & List, J. A. (2015b). Do competitive workplaces deter female workers? A large-scale natural field experiment on job-entry decisions. *The Review of Economic Studies*, *82*(1), 122–155.

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, *118*(3), 1049–1074.

Heckman, J., Moon, S. H., Pinto, R., Savelyev, P., & Yavitz, A. (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the highscope perry preschool program. *Quantitative Economics*, *1*(1), 1–46.

Heckman, J. J., Pinto, R., Shaikh, A. M., & Yavitz, A. (2011). Inference with imperfect randomization: The case of the perry preschool program. National Bureau of Economic Research Working Paper w16935.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hossain, T., & List, J. A. (2012). The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science*, *58*(12), 2151–2167.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), e124.

Jennions, M. D., & Moller, A. P. (2002). Publication bias in ecology and evolution: An empirical assessment using the 'trim and fill' method. *Biological Reviews of the Cambridge Philosophical Society*, *77*(02), 211–222.

Karlan, D., & List, J. A. (2007). Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *The American Economic Review*, *97*(5), 1774–1793.

Kling, J., Liebman, J., & Katz, L. (2007). Experimental analysis of neighborhood effects. *Econometrica*, *75*(1), 83–119.

Lee, S., & Shaikh, A. M. (2014). Multiple testing and heterogeneous treatment effects: Re-evaluating the effect of progresa on school enrollment. *Journal of Applied Econometrics*, *29*(4), 612–626.

Lehmann, E., & Romano, J. (2005). Generalizations of the familywise error rate. *The Annals of Statistics*, *33*(3), 1138–1154.

Lehmann, E. L., & Romano, J. P. (2006). *Testing statistical hypotheses*. Berlin: Springer.

Levitt, S. D., List, J. A., Neckermann, S., & Sadoff, S. (2012). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. National Bureau of Economic Research w18165.

List, J. A., & Samek, A. S. (2015). The behavioralist as nutritionist: Leveraging behavioral economics to improve child food choice and consumption. *Journal of Health Economics*, *39*, 135–146.

Machado, C., Shaikh, A., Vytlacil, E., & Lunch, C. (2013). Instrumental variables, and the sign of the average treatment effect. Unpublished Manuscript, Getulio Vargas Foundation, University of Chicago, and New York University. [2049].

Maniadis, Z., Tufano, F., & List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *The American Economic Review*, *104*(1), 277–290.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, *122*(3), 1067–1101.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia ii. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

Romano, J. P., & Shaikh, A. M. (2006a). On stepdown control of the false discovery proportion. In *Lecture Notes-Monograph Series* (pp. 33–50).

Romano, J. P., & Shaikh, A. M. (2006b). Stepup procedures for control of generalizations of the familywise error rate. *The Annals of Statistics*, *34*, 1850–1873.

Romano, J. P., & Shaikh, A. M. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *The Annals of Statistics*, *40*(6), 2798–2822.

Romano, J. P., Shaikh, A. M., & Wolf, M. (2008a). Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, *17*(3), 417–442.

Romano, J. P., Shaikh, A. M., & Wolf, M. (2008b). Formalized data snooping based on generalized error rates. *Econometric Theory*, *24*(02), 404–447.

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, *73*(4), 1237–1282.

Romano, J. P., & Wolf, M. (2010). Balanced control of generalized error rates. *The Annals of Statistics*, *38*, 598–633.

Sutter, M., & Glätzle-Rützler, D. (2014). Gender differences in the willingness to compete emerge early in life and persist. *Management Science*, *61*(10), 2339–23354.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p value adjustment* (Vol. 279). New York: Wiley.

Affiliations

John A. List[1] · Azeem M. Shaikh[1] · Yang Xu[1]

✉ Yang Xu
yangxu@uchicago.edu

John A. List
jlist@uchicago.edu

Azeem M. Shaikh
amshaikh@uchicago.edu

[1] Department of Economics, University of Chicago, 5757 S University Ave, Chicago, IL 60637, USA