Received: 27 May 2021 | Revised: 18 March 2023

DOI: 10.1002/jae.2985

WILEY Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE

Multiple testing with covariate adjustment in experimental economics

John A. List[1] | **Azeem M. Shaikh**[2] | **Atom Vayalinkal**[3]

[1]Department of Economics, University of Chicago and NBER, Chicago, Illinois, USA

[2]Department of Economics, University of Chicago, Chicago, Illinois, USA

[3]Department of Economics, University of Toronto, Toronto, Ontario, Canada

Correspondence
Azeem M. Shaikh, Department of Economics, University of Chicago and NBER, Chicago, IL, USA.
Email: amshaikh@uchicago.edu

Funding information
National Science Foundation, Grant/Award Number: SES-1530661; Social Sciences and Humanities Research Council of Canada, Grant/Award Number: CGS-767-2022-2529

Summary

This paper provides a framework for testing multiple null hypotheses simultaneously using experimental data in which simple random sampling is used to assign treatment status to units. Using general results from the multiple testing literature, we develop under weak assumptions a procedure that (i) asymptotically controls the familywise error rate—the probability of one or more false rejections—and (ii) is asymptotically balanced in that the marginal probability of rejecting any true null hypothesis is approximately equal in large samples. Our procedure improves upon classical methods by incorporating information about the joint dependence structure of the test statistics when determining which null hypotheses to reject, leading to gains in power. An important point of departure from prior work is that we exploit observed, baseline covariates to obtain further gains in power. The precise way in which we incorporate these covariates is based on recent results from the statistics literature in order to ensure that inferences are typically more powerful in large samples.

KEYWORDS

experiments, familywise error rate, multiple testing, treatment effects

1 | INTRODUCTION

False positives have become the poster child for the credibility revolution of the past two decades in the social sciences. While incenting and improving replications certainly hold an important place in this movement (see, e.g., earlier studies, Butera et al., 2020; Dreber et al., 2015; Maniadis et al., 2014), in this study, we focus instead on false positives that arise because of a failure to account for the testing of multiple null hypotheses simultaneously. As in List et al. (2019), we focus on experimental data in which treatment status is assigned using simple random sampling. In the analysis of experimental data, 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 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 each of the other treatments. Using the general results in Romano and Wolf (2010), we develop under weak assumptions a procedure that (i) asymptotically controls the familywise error rate—the probability of one or more false rejections—and (ii) is asymptotically balanced in that the marginal probability of rejecting any true null hypothesis is approximately equal in large samples. The resulting procedure differs from classical multiple testing procedures, such as Bonferroni (1935) and Holm (1979), in that it incor-

porates information about the joint dependence structure of the test statistics when determining which null hypotheses to reject. This feature leads to important gains in power. An important point of departure, however, is that we exploit observed, baseline covariates to obtain further gains in power. The precise way in which these covariates are incorporated is based upon results in Ye et al. (2022) in order to ensure that inferences are typically more powerful in large samples. We highlight these gains in power by applying our approach to the data in Karlan and List (2007), who examine questions within the economics of charitable giving in a natural field experiment. In particular, they examine how matching grants influence giving rates. In comparison with the reanalysis of this data presented in List et al. (2019), by exploiting observed, baseline covariates, we find even more evidence that the major results in Karlan and List (2007) hold after accounting for the multiplicity of tests under consideration.

The problems stemming from testing multiple null hypotheses simultaneously in empirical research have enjoyed increasing visibility in economics. For some recent applications of multiple testing in empirical research, see Anderson et al. (2003), Lee and Shaikh (2014), Heckman et al. (2010), and Heckman et al. (2020). These papers exploit mainly theoretical results found in Romano and Wolf (2005), upon which Romano and Wolf (2010) build. For an overview of such methods, see Romano et al. (2010b), Romano et al. (2008b), and Romano et al. (2010a). Other closely related methods can be found in Romano and Shaikh (2006) and Romano et al. (2008a).

The remainder of our note proceeds as follows. Section 2 describes our general setting whereas Section 3 describes our procedure algorithmically and our main theoretical result. Section 4 describes our charitable giving case study design and empirical results. Section 5 concludes. Documentation of our procedures and the code to apply our approach in Stata can be found at the following address (https://github.com/vayalinkal/mhtexp2).

2 | SETUP AND NOTATION

For $k \in \mathcal{K}$, let $Y_{i,k}$ denote the kth observed outcome of interest for the ith unit, D_i denote treatment status for the ith unit, X_i denote (a vector of) observed, baseline covariates for the ith unit, and Z_i denote the subgroup that the ith unit belongs to. 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 $Y_i = \left(Y_{i,k} : k \in \mathcal{K}\right)$ and $Y_i(d) = \left(Y_{i,k}(d) : k \in \mathcal{K}\right)$. We assume that $\left((Y_i(d) : d \in \mathcal{D}), D_i, X_i, Z_i\right), i = 1, \ldots, n$ are i.i.d. with distribution $Q \in \Omega$, where our requirements on Ω are specified below. It follows that the observed data $(Y_i, D_i, X_i, Z_i), i = 1, \ldots, n$ 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

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

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

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

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 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, let

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

As in List et al. (2019), our goal is to construct a procedure for testing these null hypotheses in a way that ensures asymptotic control of the familywise error rate uniformly over $Q \in \Omega$. In contrast to List et al. (2019), however, our analysis

exploits the observed, baseline covariates X_i. As explained further in Remark 3.1 in Section 3 below, the precise way in which we do so ensure that the use of such covariates leads to more powerful inferences, at least in large samples.

We require for each $Q \in \Omega$ that

$$\limsup_{n\to\infty} \text{FWER}_Q \leq \alpha \tag{2}$$

for a prespecified value of $\alpha \in (0,1)$, where

$$\text{FWER}_Q = Q\left\{\text{reject any } H_s \text{ with } s \in \mathcal{S}_0(Q)\right\}.$$

We additionally require that the testing procedure is "balanced" in that for each $Q \in \Omega$,

$$\lim_{n\to\infty} Q\left\{\text{reject } H_s\right\} = \lim_{n\to\infty} Q\left\{\text{reject } H_{s'}\right\} \text{ for any } s \text{ and } s' \text{ in } \mathcal{S}_0(Q). \tag{3}$$

Remark 2.1. We emphasize that the family of null hypotheses of interest, $\mathcal{S}$, is only a subset of the set of all possible null hypotheses one could test and that it may not always be desirable to choose $\mathcal{S}$ to be as large as possible. Indeed, the choice of $\mathcal{S}$ may be disciplined, in part, by which decisions one wishes to make and/or what one wishes to learn with confidence from the data. For a discussion, see Viviano et al. (2022). To help illustrate this point, we examine a few different scenarios and consider some potential choices of $\mathcal{S}$ for each of them.

(a) Suppose there is a single treatment, a single outcome, and multiple subpopulations defined according to the joint values of two binary variables z_1 and z_2. Further suppose that it is only feasible to decide to treat everyone with a common value of z_1, regardless of their value of z_2. In this case, if one wants to decide *which* subpopulations to treat, then it is natural to define $\mathcal{S}$ as the family of null hypotheses determined by only the two subpopulations determined by z_1, rather than the four subpopulations determined by (z_1, z_2).

(b) Suppose there is a single treatment, multiple outcomes and a single subpopulation. It is now less clear whether the null hypotheses corresponding to all such outcomes should be included in the family. To guide decision-making about whether to adopt the treatment or not, it may be most desirable (albeit, possibly difficult) to collapse these outcomes into a single outcome and include in $\mathcal{S}$ only those hypotheses involving this aggregate outcome (see, e.g., Heckman et al., 2013). In many cases, however, it may also be of interest to determine, with some confidence, *which* outcomes are affected by the treatment, in which case all such hypotheses should be included. This may also be useful when there is some ambiguity about how best to aggregate these multiple outcomes into a single outcome.

(c) Suppose there are multiple treatments, a single outcome, and a single subpopulation. If one wishes to determine with confidence *which* treatments differ from the control, then all comparisons with the control should be included in $\mathcal{S}$. If one wishes to also rank the treatments, then it is natural to include all pairwise comparisons in $\mathcal{S}$; in this case, however, it may be necessary to control not just the familywise error rate but also the mixed directional familywise error rate, as in Bazylik et al. (2021), Mogstad et al. (2022a), and Mogstad et al. (2022b).

Finally, we note that when the number of null hypotheses included in $\mathcal{S}$ is large, it may be desirable to consider alternative error rates that are less demanding than the FWER; for a discussion, see Remark 3.10 in Section 3. □

Let $P = P(Q)$ be the distribution of the observed data (Y_i, D_i, X_i, Z_i), $i = 1, \ldots, n$. We will often suppress the dependence of P on Q for simplicity. It is useful to introduce the following notation that will be used repeatedly in the sequel: for $k \in \mathcal{K}$, $d \in \mathcal{D}$, and $z \in \mathcal{Z}$, define

$$b_{k|d,z}(P) = \text{Var}_P[X_i|D_i = d, Z_i = z]^{-1}\text{Cov}_P[X_i, Y_{i,k}|D_i = d, Z_i = z],$$
$$\mu_{k|d,z}(P) = E_P\left[Y_{i,k}|D_i = d, Z_i = z\right].$$

We now describe our main requirements on Ω.

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

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

under Q.

Assumption 2.2. For each $Q \in \Omega$, there is $\epsilon > 0$ such that

$$Q\{D_i = d, Z_i = z\} > \epsilon \tag{4}$$

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

Assumption 2.3. For each $Q \in \Omega$, $\text{Var}_Q[X_i|D_i = d, Z_i = z]$ is invertible,

$$E_Q[X_i X_i'|D_i = d, Z_i = z] < \infty \text{ and } E_Q[Y_{i,k}^2(d)|D_i = d, Z_i = z] < \infty,$$

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

Assumption 2.4. For each $Q \in \Omega$,

$$\text{Var}_Q[Y_{i,k}(d) - b_{k|d,z}(P)'X_i|D_i = d, Z_i = z] > 0,$$

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

Assumption 2.1 requires that treatment status was assigned using simple random sampling. Despite its simplicity, this treatment assignment scheme remains widely used. We emphasize, however, that it precludes many other popular treatment assignment schemes, including "matched pairs" and stratified block randomization. For relevant results, see Bugni et al. (2018), Bugni et al. (2019), and Bai et al. (2021). We leave the extension of our results to these other treatment assignment schemes to future work. Assumption 2.2 simply requires that both D_i and Z_i are discrete random variables (with finite supports). Assumption 2.3 requires that there is no perfect multicollinearity among the X_i and that the components of X_i and $Y_i(d)$ have at least two moments. Assumption 2.4 is a mild nondegeneracy requirement.

3 | PROCEDURE

In this section, we describe a stepwise multiple testing procedure for testing (1) in a way that satisfies (2) and (3) for any $Q \in \Omega$. In order to do so, we first require some additional notation.

For $d \in \mathcal{D}$ and $z \in \mathcal{Z}$, define $n_{d,z} = \sum_{1 \le i \le n} I\{D_i = d, Z_i = z\}$, $n_z = \sum_{1 \le i \le n} I\{Z_i = z\}$, and

$$\begin{aligned}\mu_{X|d,z}(P) &= E_P[X_i|D_i = d, Z_i = z],\\ \mu_{X|z}(P) &= E_P[X_i|Z_i = z].\end{aligned}$$

Using this notation and recalling the definitions of $\mu_{k|d,z}(P)$ and $b_{k|d,z}(P)$ in the previous section, we may define, for each $k \in \mathcal{K}$, $d \in \mathcal{D}$, and $z \in \mathcal{Z}$,

$$\theta_{k|d,z}(P) = \mu_{k|d,z}(P) - b_{k|d,z}(P)'(\mu_{X|d,z}(P) - \mu_{X|z}(P))$$

and its sample counterpart

$$\hat{\theta}_{k|d,z} = \bar{Y}_{k|d,z} - \hat{b}_{k|d,z}'(\bar{X}_{d,z} - \bar{X}_z),$$

where $\bar{X}_z = n_z^{-1} \sum_{i=1}^n X_i I\{Z_i = z\}$, $\bar{X}_{d,z} = n_{d,z}^{-1} \sum_{i=1}^n X_i I\{D_i = d, Z_i = z\}$, and

$$\hat{b}_{k|d,z} = \left(\sum_{1 \le i \le n : D_i = d, Z_i = z} (X_i - \bar{X}_{d,z})(X_i - \bar{X}_{d,z})' \right)^{-1} \left(\sum_{1 \le i \le n : D_i = d, Z_i = z} (X_i - \bar{X}_{d,z}) Y_{i,k} \right).$$

Note that $\hat{\theta}_{k|d,z}$ can be obtained as the ordinary least squares estimator of the intercept coefficient in a linear regression of $Y_{i,k}$ on a constant and $X_i - \bar{X}_z$ using only the subsample of data with $D_i = d, Z_i = z$.

Now, for each s, we may define an "unbalanced" test statistic for H_s

$$T_{s,n} = \sqrt{n}\left|\hat{\theta}_{k|d,z} - \hat{\theta}_{k|d',z}\right| \tag{5}$$

and its recentered version

$$\tilde{T}_{s,n}(P) = \sqrt{n}\left|\left(\hat{\theta}_{k|d,z} - \theta_{k|d,z}(P)\right) - \left(\hat{\theta}_{k|d',z} - \theta_{k|d',z}(P)\right)\right|. \tag{6}$$

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

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

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

$$J_n\left(T_{s,n}, s, \hat{P}_n\right). \tag{7}$$

Note that (7) 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

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

Using this notation and following Romano and Wolf (2010), we may describe our proposed stepwise multiple testing procedure according to the algorithm below:

Algorithm 3.1

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

⋮

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

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

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

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

and continue to the next step.

⋮

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

Theorem 3.1. *Consider the procedure for testing (1) given by Algorithm 1. Under Assumptions 2.1–2.4, Algorithm 1 satisfies (2) and (3) for any* $Q \in \Omega$.

In order to better understand Algorithm 1, it is helpful to view Step j of the algorithm as testing the joint null hypothesis that H_s holds for all $s \in \mathcal{S}_j$ using $\max_{s\in\mathcal{S}_j} J_n\left(T_{s,n}, s, \hat{P}_n\right)$ as a test statistic (for which large values provide evidence against the null hypothesis) and the quantity $L_n^{-1}\left(1-\alpha, \mathcal{S}_j, \hat{P}_n\right)$ as a critical value. In fact, for any fixed $s = (d, d', z, k) \in \mathcal{S}_j$,

$$\left\{\theta_{k|d,z} - \theta_{k|d',z} : J_n\left(\sqrt{n}\left|\left(\hat{\theta}_{k|d,z} - \theta_{k|d,z}\right) - \left(\hat{\theta}_{k|d',z} - \theta_{k|d',z}\right)\right|, s, P\right) \leq x\right\}$$

may be viewed as a confidence interval for $\theta_{k|d,z} - \theta_{k|d',z}$ with coverage probability x. The probability that these intervals simultaneously contain the true values of $\theta_{k|d,z} - \theta_{k|d',z}$, for the corresponding $s \in \mathcal{S}_j$, is given by $L_n\left(x, \mathcal{S}_j, P\right)$. This suggests that

$$\left\{\left(\theta_{k|d,z} - \theta_{k|d',z} : s \in \mathcal{S}_j\right) : \max_{s\in\mathcal{S}_j} J_n\left(\sqrt{n}\left|\left(\hat{\theta}_{k|d,z} - \theta_{k|d,z}\right) - \left(\hat{\theta}_{k|d',z} - \theta_{k|d',z}\right)\right|, s, \hat{P}_n\right) \leq L_n^{-1}\left(1-\alpha, \mathcal{S}_j, \hat{P}_n\right)\right\},$$

which employs the bootstrap to estimate $J_n(x, s, P)$ and $L_n(x, \mathcal{S}_j, P)$, is a (simultaneous) confidence region for $(\theta_{k|d,z} - \theta_{k|d',z} : s \in \mathcal{S}_j)$ with coverage probability (approximately) $1 - \alpha$. For further discussion of such simultaneous confidence regions, see Sections 1.4 and 2.2 of Beran (1990) and Section 3 of Romano and Wolf (2010).

Remark 3.1. As mentioned previously, our point of departure from List et al. (2019) is the use of X_i to obtain more powerful inferences. Ye et al. (2022) show that inferences based on $\hat{\theta}_{k|d,z} - \hat{\theta}_{k|d',z}$ are always at least as powerful as those based on a simple difference in means in the following sense: Corollary 1 of Ye et al. (2022) shows that the variance of the limiting distribution of $\sqrt{n}(\hat{\theta}_{k|d,z} - \hat{\theta}_{k|d',z})$ is no greater than the variance of the limiting distribution of $\sqrt{n}(\bar{Y}_{k|d,z} - \bar{Y}_{k|d',z})$ and the comparison is strict unless X_i is uncorrelated with $Y_i(d)$ and $Y_i(d')$. Therefore, our choice of test statistic makes use of covariates in a way that typically leads to substantially more powerful inferences, at least in large samples, when compared with the simple difference-in-means test statistics of List et al. (2019). Negi and Wooldridge (2020) obtain similar results in a binary treatment setting. Earlier antecedents studying regression adjustment of experimental data include Yang and Tsiatis (2001) and Tsiatis et al. (2008). For related results in a finite population setting, see Lin (2013) and Berk et al. (2013). □

Remark 3.2. If $\mathcal{S} = \{s\}$, that is, $\mathcal{S}$ is a singleton, then the familywise error rate is simply the usual probability of a Type I error. Hence, Algorithm 1 provides asymptotic control of the probability of a Type I error. In this case, Algorithm 1 is equivalent to the usual bootstrap test of H_s, that is, the test that rejects H_s whenever $T_{s,n} > J_n^{-1}\left(1 - \alpha, s, \hat{P}_n\right)$. □

Remark 3.3. As noted above, $\hat{p}_{s,n} = 1 - J_n\left(T_{s,n}, s, \hat{P}_n\right)$ 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

$$\limsup_{n\to\infty} Q\left\{\hat{p}_{s,n} \leq u\right\} \leq u$$

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. Such an approach would indeed satisfy (2), as desired, but implicitly relies 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 1 uses the bootstrap to implicitly incorporate information about the dependence structure when deciding which null hypotheses to reject. In fact, since assuming any particular dependence structure is always at least as powerful as assuming a "least favorable" one, Algorithm 1 will always reject at least as many null hypotheses as these procedures, even in finite samples, while still maintaining asymptotic control of the familywise error rate. □

Remark 3.4. Implementation of Algorithm 1 typically requires approximating the quantities $J_n\left(x, s, \hat{P}_n\right)$ and $L_n\left(x, \mathcal{S}', \hat{P}_n\right)$ 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 $\tilde{T}^{*,b}_{s,n}\left(\hat{P}_n\right)$ the quantity $\tilde{T}_{s,n}(P)$ using the b-th resample and $\hat{P}_n$ as an estimate of P. Then, $J_n\left(x, s, \hat{P}_n\right)$ may be approximated as

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

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

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

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

Remark 3.5. It is often desirable to studentize, that is, to replace $T_{s,n}$ and $\tilde{T}_{s,n}(P)$, respectively, with

$$T^{\text{stud}}_{s,n} = \frac{T_{s,n}}{\sqrt{\tilde{\sigma}^2_{s,n}\left(\hat{P}_n\right)}} \text{ and } \tilde{T}^{\text{stud}}_{s,n}(P) = \frac{\tilde{T}_{s,n}(P)}{\sqrt{\tilde{\sigma}^2_{s,n}\left(\hat{P}_n\right)}},$$

where

$$\tilde{\sigma}^2_{s,n}(P) = \frac{S^2_{k|d,z}(P)}{P\{D_i = d, Z_i = z\}} + \frac{S^2_{k|d',z}(P)}{P\{D_i = d', Z_i = z\}} + \left(b_{k|d,z}(P) - b_{k|d',z}(P)\right)' \frac{\mathrm{Var}_P\left[X_i | Z_i = z\right]}{P\{Z_i = z\}} \left(b_{k|d,z}(P) - b_{k|d',z}(P)\right)$$

and $S^2_{k|d,z}(P) = \mathrm{Var}_P\left[Y_{i,k} - b_{k|d,z}(P)'X_i | D_i = d, Z_i = z\right]$. Theorem 3 continues to hold with these changes. □

Remark 3.6. In some cases, it may be of interest to consider one-sided null hypotheses, for example, $H_s^- : P \in \omega_s^-$ where

$$\omega_s^- = \left\{Q \in \Omega : E_Q\left[Y_{i,k}(d) - Y_{i,k}\left(d'\right) | Z_i = z\right] \leq 0\right\}. \tag{8}$$

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 (5) and (6) 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 (8) but with the inequality reversed. □

Remark 3.7. Note that a multiplicity-adjusted p-value for H_s, $\hat{p}^{\mathrm{adj}}_{s,n}$, may be computed simply as the smallest value of α for which H_s is rejected in Algorithm 1.

Remark 3.8. It is possible to improve Algorithm 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

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

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

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

With this modification to Algorithm 1, Theorem 3.1 remains valid. Note that this modification is only nontrivial when there are more than two treatments and may be computationally prohibitive when there are more than a few treatments. □

Remark 3.9. 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 1 satisfies

$$\limsup_{n\to\infty} \sup_{Q\in\Omega} \mathrm{FWER}_Q \leq \alpha.$$

In particular, it suffices to replace Assumption 2.3 with a mild uniform integrability requirement and require in Assumption 2.2 that there exists $\epsilon > 0$ for which (4) 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. (2019). □

Remark 3.10. In some settings, it may be desirable to only control the probability of making k or more false rejections, for some $k > 1$; that is, only require that, for all $Q \in \Omega$,

$$\limsup_{n\to\infty} k\text{-FWER}_Q \leq \alpha,$$

where

$$k\text{-FWER}_Q := Q\left\{\text{reject at least } k \text{ hypotheses } H_s \text{ with } s \in \mathcal{S}_0(Q)\right\}.$$

In this notation, Algorithm 1 controls the 1-FWER. Corollary 5.1 in Romano and Wolf (2010) implies that, under the same assumptions as Theorem 3.1, it is possible to adapt Algorithm 1 to asymptotically control the k-FWER instead, for $k > 1$. For more details on how to modify Algorithm 1 in this way, see Algorithm 4.1 and subsequent discussion in Romano and Wolf (2010). Further, Theorem 8.1 in Romano and Wolf (2010) implies that such an Algorithm 1-based k-FWER control procedure can be sequentially applied to asymptotically control (the tail probability of) the *false discovery proportion* (FDP); that is, to ensure that, for a specific $\gamma \in [0, 1)$ and all $Q \in \Omega$,

$$\limsup_{n\to\infty} Q\{\text{FDP} > \gamma\} \leq \alpha,$$

where

$$\text{FDP} = \begin{cases} 0 & \text{if no hypotheses rejected} \\ \frac{\text{number of rejected hypotheses} H_s \text{ with } s \in S_0(Q)}{\text{total number of rejected hypotheses}} & \text{otherwise} \end{cases}.$$

Note that Algorithm 1 controls the FDP with $\gamma = 0$. For details on how to sequentially apply k-FWER control procedures to control the FDP for $\gamma > 0$, see Algorithm 8.1 in Romano and Wolf (2010).

A related, and popular, approach to controlling false rejections is to control the *false discovery rate* (FDR), which requires that, for all $Q \in \Omega$,

$$\text{FDR} = E_Q[\text{FDP}] \leq \alpha.$$

Unlike error control approaches based on controlling the k-FWER or the FDP, which provide information about the realized number or proportion of false rejections, controlling the FDR only provides such information indirectly through, for example, bounds on the tail probability of the FDP obtained via Markov's inequality: Controlling the FDR at level α implies that $Q\{\text{FDP} > \gamma\} \leq \min\left\{1, \frac{\alpha}{\gamma}\right\}$; see, for example, Romano and Shaikh (2006). Thus, even if the FDR is controlled at level α, the tail probability of the FDP may remain high. For this reason, we do not consider error control based on the FDR in this paper and instead suggest methods based on the k-FWER or the FDP in settings where controlling the FWER may be too demanding. For some relevant results on asymptotic control of the FDR, however, we refer readers to Romano et al. (2008a). □

4 | EMPIRICAL APPLICATIONS

In this section, we apply our testing methodology to the large-scale natural field experiment presented in Karlan and List (2007). Before proceeding, we briefly summarize the main features of the experiment. Using direct mail solicitations targeted to previous donors of a nonprofit organization, Karlan and List (2007) study the effectiveness of a matching grant on charitable giving. The sample consists of 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 (33,396 or 66.68% of the sample) and with probability one third to a control group (16,687 subjects or 33.32% of the sample). Individuals in the treatment group were then offered independently and with equal probability one of 36 possible matching grants, whose specifics varied along three dimensions: the price ratio of the match, the maximum size of the matching gift across all donations, and the example donation amount suggested to the donor. The three possible values for the price ratio of the match were \$1:\$1, \$2:\$1, and \$3:\$1. An \$X:\$1 ratio means that for every dollar, the individual donates, the matching donor also contributes \$X; hence, the charity receives \$X+1 (subject to the maximum amount across all donations). There were four possible values for the maximum matching grant amount: \$25,000, \$50,000, \$100,000, and unstated. The three possible values for the individual-specific suggested 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 emerge 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 3.5 and controlling for a suite of observed, baseline characteristics. Specifically, we control for a number of individual-level characteristics (including sex, number of years since initial donation, and whether they had already donated in 2005),

zip-code level demographics (including average household size, proportion of white and black residents, proportion of residents aged between 18 and 39), and measures of state-level activity of the organization. We also compare our results with those obtained using the classical Bonferroni and Holm multiple testing procedures. In order to emphasize the magnitude of the potential gains in power from properly exploiting observed, baseline covariates, we also present results without exploiting such information, as in List et al. (2019). We also provide an illustration of the potential gains in power from exploiting transitivity restrictions as described in Remark 3.8.

In order to ensure that the results in each table are based on the same sample, we compute our results using the 46,521 individuals for whom all covariates and subgroup identifiers are available. The resulting sample is a subset of the full sample used in List et al. (2019), but maintains very similar treatment proportions, with 31,021 individuals, or 66.68 percent of the remaining sample, being assigned to treatment, and 15,500 individuals, or 33.32 percent of the remaining sample, being assigned to control. The percentages of treatment units assigned a specific match offer are also very similar across the full sample and the selected sample. Stata code used to produce these results, including the sample selection, is available at the following address: https://github.com/vayalinkal/mhtexp2/tree/main/replication.

4.1 | Multiple outcomes

We assess the effects of the treatment on the four outcome variables of interest in Karlan and List (2007): response rate, dollars given not including match, dollars given including match, and amount change. Here, by treatment, we mean receiving any of the 36 possible matching grants. Tables 1 and 2 display, for each of the four outcomes of interest, the following five quantities: column 2 displays the difference in means between the treated and the untreated subjects for the four outcomes. Column 3 displays a (multiplicity-unadjusted) p-value computed using Remark 3.2; column 4 displays a (multiplicity-adjusted) p-value computed using Theorem 3.1. Column 5 displays a (multiplicity-adjusted) p-value obtained by applying a Bonferroni adjustment to the p-values in column 3; column 6 displays a (multiplicity-adjusted) p-value obtained by applying a Holm adjustment to the p-values in column 3. Table 1 reports the results when we do not include controls for any covariates, whereas Table 2 reports results with controls for covariates included.

TABLE 1 Multiple outcomes, without covariates.

		p-values			
		Unadjusted	Multiplicity adjusted		
Outcome	Coeff.	Remark 3.2	Theorem 3.1	Bonferroni	Holm
Response rate	0.0049	0.0003***	0.0003***	0.0013***	0.0010***
Dollars given not including match	0.1831	0.0263**	0.0517*	0.1053	0.0527*
Dollars given including match	2.1740	0.0003***	0.0003***	0.0013***	0.0013***
Amount change	6.8533	0.7150	0.7150	1.0000	0.7150

Note: As noted above, these values are computed using only the subsample with all covariates and subgroup identifiers available and may differ from the corresponding values in List et al. (2019).

*Indicates that the corresponding p-values are less than 10%.

**Indicates that the corresponding p-values are less than 5%.

***Indicates that the corresponding p-values are less than 1%.

TABLE 2 Multiple outcomes, with covariates.

		p-values			
		Unadjusted	Multiplicity adjusted		
Outcome	Coeff.	Remark 3.2	Theorem 3.1	Bonferroni	Holm
Response rate	0.0049	0.0003***	0.0003***	0.0013***	0.0013***
Dollars given not including match	0.1835	0.0260**	0.0510*	0.1040	0.0520*
Dollars given including match	2.1744	0.0003***	0.0003***	0.0013***	0.0010***
Amount change	7.3908	0.6640	0.6640	1.0000	0.6640

*Indicates that the corresponding p-values are less than 10%.

**Indicates that the corresponding p-values are less than 5%.

***Indicates that the corresponding p-values are less than 1%.

Comparing the second column of Tables 1 and 2, we see that for each outcome the estimated treatment effect is weakly larger when we include controls. Moreover, both the unadjusted *p*-values and the adjusted *p*-values based on Theorem 3.1 in Table 2 are weakly smaller than their corresponding *p*-values in Table 1. In both tables, the *p*-values in column 3, which do not take multiplicity into account, indicate that the treatment has a significant effect on response rate, dollars given not including match, and dollars given including match, at the 0.05 level. Once we take multiplicity into account, however, column 4 of Tables 1 and 2 suggests that the effect of the "match" treatment on dollars given not including match is no longer significant at the 0.05 level, only remaining significant at the 0.10 level.

Importantly, because of the incorporation of information about the joint dependence structure of the test statistics when determining which null hypotheses to reject, the *p*-values from Theorem 3.1 are an improvement upon those obtained by applying Bonferroni or Holm adjustments. This feature is evident in both Tables 1 and 2 as the *p*-values in column 4 are always weakly smaller than the *p*-values in columns 5 and 6.

4.2 | Multiple subgroups

The 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. As in the preceding subsection, by treatment, we mean receiving any of the 36 possible matching grants. We focus on a single outcome variable, namely the response rate. Tables 3 and 4 display, for each of the four subgroups of interest, the five quantities similar to those in Tables 1 and 2.

Compared with *p*-values that do not take multiplicity into account, the multiplicity-adjusted *p*-values indicate that, after accounting for multiple testing, the treatment may only have an effect among a smaller number of subgroups. In particular, the values in column 3, which do not account for multiple testing, indicate that the treatment has a significant effect, at the 0.05 level, on response rate among the subgroups "red county in a red state" and "blue county in a red state", with and without controls. After taking the multiplicity issue into account, however, column 4 of each table suggests that the treatment effect for the subgroup "blue county in a red state" only maintains significance at the 0.10 level. Once again, we see that the procedure described in Theorem 3.1 is more powerful than the Bonferroni and Holm procedures, as the *p*-values in column 4 are always weakly smaller than the *p*-values in columns 5 and 6.

TABLE 3 Multiple subgroups, without covariates.

		***p*-values**			
		Unadjusted	**Multiplicity adjusted**		
Subgroup	**Coeff.**	**Remark 3.2**	**Theorem 3.1**	**Bonferroni**	**Holm**
Red county, red state	0.0108	0.0003***	0.0003***	0.0013***	0.0013***
Blue county, red state	0.0083	0.0193**	0.0567*	0.0773*	0.0580*
Red county, blue state	0.0024	0.2913	0.4960	1.0000	0.5827
Blue county, blue state	0.0000	0.9900	0.9900	1.0000	0.9900

*Note:*As noted above, these values are computed using only the subsample with all covariates and subgroup identifiers available and may differ from the corresponding values in List et al. (2019).
*Indicates that the corresponding *p*-values are less than 10%.
**Indicates that the corresponding *p*-values are less than 5%.
***Indicates that the corresponding *p*-values are less than 1%.

TABLE 4 Multiple subgroups, with covariates.

		***p*-values**			
		Unadjusted	**Multiplicity adjusted**		
Subgroup	**Coeff.**	**Remark 3.2**	**Theorem 3.1**	**Bonferroni**	**Holm**
Red county, red state	0.0108	0.0003***	0.0003***	0.0013***	0.0013***
Blue county, red state	0.0083	0.0180**	0.0527*	0.0720*	0.0540*
Red county, blue state	0.0025	0.2797	0.4813	1.0000	0.5593
Blue county, blue state	−0.0000	0.9987	0.9987	1.0000	0.9987

*Indicates that the corresponding *p*-values are less than 10%.
**Indicates that the corresponding *p*-values are less than 5%.
***Indicates that the corresponding *p*-values are less than 1%.

We also see that, with controls included, almost all of the estimated effects are weakly larger and have weakly smaller *p*-values. The only exception is the estimated effect of the treatment on response rate for the subgroup "blue county in a blue state", which is very small and statistically insignificant in both cases but has slightly larger *p*-values (adjusted and unadjusted) and a smaller estimated effect (in both signed and absolute values) when controls are included.

4.3 | Multiple treatments

We first consider the comparison of each treatment with the control. For each of the three treatments, Tables 5 and 6 display the five quantities as described previously. The first table displays the results without controls, while the latter displays results with controls. Comparing these two tables, we see that the estimated treatment effect for the 1:1 match ratio is larger when controls are included, leading to smaller *p*-values, while the estimated treatment effect of the 2:1 and 3:1 match ratios are smaller and the corresponding *p*-values slightly larger when controls are included.

We now consider null hypotheses that emerge due to multiple treatments. We focus on the three treatments on matching-ratio dimension: 1:1, 2:1, and 3:1. We consider "dollars given not including match" as our outcome of interest.

In each table, the values in column 3, which do not take multiplicity into account, suggest that the match ratio 2:1 has a significant effect on the outcome "dollars given not including match", at the 0.05 level. Nonetheless, as shown in

TABLE 5 Multiple treatments with a control, without covariates.

		p-values			
		Unadjusted	Multiplicity adjusted		
Comparison groups	Coeff.	Remark 3.2	Theorem 3.1	Bonferroni	Holm
1:1 versus control	0.1228	0.2547	0.2547	0.7640	0.2547
2:1 versus control	0.2566	0.0273**	0.0747*	0.0820*	0.0820*
3:1 versus control	0.1703	0.1130	0.2050	0.3390	0.2260

Note: As noted above, these values are computed using only the subsample with all covariates and subgroup identifiers available and may differ from the corresponding values in List et al. (2019).

*Indicates that the corresponding *p*-values are less than 10%.

**Indicates that the corresponding *p*-values are less than 5%.

TABLE 6 Multiple treatments with a control, with covariates.

		p-values			
		Unadjusted	Multiplicity adjusted		
Comparison groups	Coeff.	Remark 3.2	Theorem 3.1	Bonferroni	Holm
1:1 versus control	0.1289	0.2347	0.2347	0.7040	0.2347
2:1 versus control	0.2564	0.0277**	0.0760*	0.0830*	0.0830*
3:1 versus control	0.1675	0.1140	0.2067	0.3420	0.2280

*Indicates that the corresponding *p*-values are less than 10%.

**Indicates that the corresponding *p*-values are less than 5%.

TABLE 7 All pairwise comparisons across multiple treatments and a control, no covariates.

		p-values				
		Unadjusted	Multiplicity adjusted			
Comparison groups	Coeff.	Remark 3.2	Theorem 3.1	Remark 3.8	Bonferroni	Holm
1:1 versus control	0.1228	0.2547	0.5790	0.4917	1.0000	1.0000
2:1 versus control	0.2566	0.0273**	0.1210	0.1210	0.1640	0.1640
3:1 versus control	0.1703	0.1130	0.3537	0.2537	0.6780	0.5650
2:1 versus 1:1	0.1338	0.3043	0.5597	0.5597	1.0000	0.9130
3:1 versus 1:1	0.0475	0.7127	0.7127	0.7127	1.0000	0.7127
3:1 versus 2:1	−0.0863	0.4947	0.7190	0.4947	1.0000	0.9893

Note: As noted above, these values are computed using only the subsample with all covariates and subgroup identifiers available and may differ from the corresponding values in List et al. (2019).

**Indicates that the corresponding *p*-values are less than 5%.

TABLE 8 All pairwise comparisons across multiple treatments and a control, with covariates.

		p-values				
		Unadjusted		Multiplicity adjusted		
Comparison groups	Coeff.	Remark 3.2	Theorem 3.1	Remark 3.8	Bonferroni	Holm
1:1 versus control	0.1289	0.2347	0.5473	0.4600	1.0000	0.9387
2:1 versus control	0.2564	0.0277**	0.1233	0.1233	0.1660	0.1660
3:1 versus control	0.1675	0.1140	0.3547	0.2557	0.6840	0.5700
2:1 versus 1:1	0.1275	0.3253	0.5877	0.5877	1.0000	0.9760
3:1 versus 1:1	0.0387	0.7647	0.7647	0.7647	1.0000	0.7647
3:1 versus 2:1	−0.0888	0.4743	0.7013	0.4743	1.0000	0.9487

**Indicates that the corresponding *p*-values are less than 5%.

column 4, the treatment effect is much less significant after applying Theorem 3.1 to this problem; in both cases, it remains significant only at the 0.10 level. Again, in both cases, the empirical results confirm that the *p*-values from Theorem 3.1 improve upon those obtained by applying the Bonferroni or Holm procedure.

We now consider all pairwise comparisons among the treatments and control. For each of the six pairwise comparisons, Tables 7 and 8 present the corresponding five quantities as well as the *p*-values described in Remark 3.8. The first table displays the results without controls, while the latter displays results with controls. Unlike all the other empirical applications, Remark 3.8 becomes nontrivial in this scenario. Recall that Remark 3.8 may improve upon Theorem 3.1 by exploiting transitivity. Indeed, among all of the multiplicity adjustments considered in Tables 7 and 8, the procedure described in Remark 3.8 appears to be the most powerful approach.

Examining the adjusted *p*-values (as given in columns 4 and 5) in the two tables, we see that the treatment effect based on the pairwise comparison between the control and the match ratio 2:1 becomes negligible in both cases. We also notice that the (multiplicity-adjusted) *p*-values in columns 4, 5, and 6 of Tables 5 and 6 are always smaller than their counterparts in Table 7 and 8, respectively, suggesting that the multiple testing problem can often become more severe with a larger number of hypotheses.

4.4 | Multiple outcomes, subgroups, and treatments

In many cases, experimentalists wish to consider families of null hypotheses that involve multiple outcomes, multiple subgroups, and multiple treatments simultaneously (as in Karlan & List, 2007). In this subsection, we simultaneously consider the four outcome variables described in Section 4.1, the four subgroups described in Section 4.2, and the three treatment conditions described in Section 4.3. For each outcome and subgroup, we compare all of the treatments to the control group.

For each of the resulting 48 null hypotheses, Table 9 displays the corresponding five quantities estimated without controlling for covariates, and Table 10 displays the same quantities estimated with controls for covariates included. Similar to our previous discussion, we find that, in both cases, many of the treatment effects are no longer significant after accounting for multiple testing. Given such a large number of null hypotheses, we can see that ignoring the multiplicity of the comparisons being made would deflate the *p*-values by a considerable margin. For instance, the multiplicity-unadjusted *p*-values produced by applying Remark 3.2 suggests that the null hypothesis for "response rate," "red county in a red state," and "2:1 versus control" could be rejected at the 0.01 significance level with the *p*-value being 0.0040 when controls are not included and 0.0060 when controls are included. Yet, by taking multiple testing into account, Theorem 3.1 yields much larger *p*-values of 0.1187 and 0.1710, with and without controls, respectively, suggesting that we cannot reject the hypothesis even at a 0.10 significance level.

Both in the case where controls are included and in the case where they are not, when we do not take multiple testing into account, 21 null hypotheses are rejected at $p < 0.10$ level. Once we adjust for multiplicity, however, Theorem 3.1 indicates that, in both cases, only seven null hypotheses are rejected. Notice that Theorem 3.1 always gives weakly smaller *p*-values than the Bonferroni and Holm procedures and, for many of the hypotheses, generates strictly smaller *p*-values than these two procedures. Indeed, with and without controls, only six null hypotheses are rejected, at the 0.10 level, using *p*-values resulting from either of these two procedures. This difference is even more pronounced if we examine the null hypotheses rejected at the 0.01 level. At the 0.01 level, with and without controls, Theorem 3.1 rejects four null hypotheses, while the Bonferroni and Holm procedures are unable to reject any.

TABLE 9 Multiple outcomes, subgroups, and treatments, without covariates.

				p-values			
				Unadjusted	Multiplicity adjusted		
Outcome	**Subgroup**	**Comp. groups**	**Coeff.**	**Remark 3.2**	**Theorem 3.1**	**Bonferroni**	**Holm**
Response rate	Red county, red state	1:1 versus control	0.0098	0.0077***	0.2067	0.3680	0.2913
Response rate	Red county, red state	2:1 versus control	0.0102	0.0040***	0.1187	0.1920	0.1600
Response rate	Red county, red state	3:1 versus control	0.0123	0.0010***	0.0240**	0.0480**	0.0440**
Response rate	Blue county, red state	1:1 versus control	0.0027	0.5523	1.0000	1.0000	1.0000
Response rate	Blue county, red state	2:1 versus control	0.0098	0.0497**	0.7380	1.0000	1.0000
Response rate	Blue county, red state	3:1 versus control	0.0123	0.0120**	0.2953	0.5760	0.4200
Response rate	Red county, blue state	1:1 versus control	0.0010	0.7017	1.0000	1.0000	1.0000
Response rate	Red county, blue state	2:1 versus control	0.0020	0.5033	1.0000	1.0000	1.0000
Response rate	Red county, blue state	3:1 versus control	0.0042	0.1680	0.9730	1.0000	1.0000
Response rate	Blue county, blue state	1:1 versus control	−0.0006	0.8803	1.0000	1.0000	1.0000
Response rate	Blue county, blue state	2:1 versus control	0.0030	0.4793	1.0000	1.0000	1.0000
Response rate	Blue county, blue state	3:1 versus control	−0.0021	0.5637	1.0000	1.0000	1.0000
Dollars given not incl. match	Red county, red state	1:1 versus control	0.5128	0.0567*	0.7637	1.0000	1.0000
Dollars given not incl. match	Red county, red state	2:1 versus control	0.4567	0.0523*	0.7453	1.0000	1.0000
Dollars given not incl. match	Red county, red state	3:1 versus control	0.3606	0.0610*	0.7713	1.0000	1.0000
Dollars given not incl. match	Blue county, red state	1:1 versus control	−0.0239	0.9393	0.9973	1.0000	1.0000
Dollars given not incl. match	Blue county, red state	2:1 versus control	0.4894	0.1523	0.9633	1.0000	1.0000
Dollars given not incl. match	Blue county, red state	3:1 versus control	0.6128	0.0910*	0.8783	1.0000	1.0000
Dollars given not incl. match	Red county, blue state	1:1 versus control	0.0215	0.8997	1.0000	1.0000	1.0000
Dollars given not incl. match	Red county, blue state	2:1 versus control	0.1466	0.4393	1.0000	1.0000	1.0000
Dollars given not incl. match	Red county, blue state	3:1 versus control	0.0616	0.7263	1.0000	1.0000	1.0000
Dollars given not incl. match	Blue county, blue state	1:1 versus control	−0.0811	0.7207	1.0000	1.0000	1.0000
Dollars given not incl. match	Blue county, blue state	2:1 versus control	0.0745	0.7557	1.0000	1.0000	1.0000
Dollars given not incl. match	Blue county, blue state	3:1 versus control	−0.1303	0.5297	1.0000	1.0000	1.0000
Dollars given incl. match	Red county, red state	1:1 versus control	1.6793	0.0090***	0.2370	0.4320	0.3330
Dollars given incl. match	Red county, red state	2:1 versus control	2.6773	0.0003***	0.0003***	0.0160**	0.0160**
Dollars given incl. match	Red county, red state	3:1 versus control	3.4035	0.0003***	0.0003***	0.0160**	0.0150**
Dollars given incl. match	Blue county, red state	1:1 versus control	0.7603	0.0937*	0.8787	1.0000	1.0000
Dollars given incl. match	Blue county, red state	2:1 versus control	3.0846	0.0037***	0.1103	0.1760	0.1503
Dollars given incl. match	Blue county, red state	3:1 versus control	4.8758	0.0070***	0.1940	0.3360	0.2730
Dollars given incl. match	Red county, blue state	1:1 versus control	0.8424	0.0090***	0.2360	0.4320	0.3240
Dollars given incl. match	Red county, blue state	2:1 versus control	2.0387	0.0003***	0.0003***	0.0160**	0.0157**
Dollars given incl. match	Red county, blue state	3:1 versus control	2.6444	0.0003***	0.0003***	0.0160**	0.0153**
Dollars given incl. match	Blue county, blue state	1:1 versus control	0.8164	0.0263**	0.5237	1.0000	0.8953
Dollars given incl. match	Blue county, blue state	2:1 versus control	2.1807	0.0027***	0.0780*	0.1280	0.1120
Dollars given incl. match	Blue county, blue state	3:1 versus control	2.4146	0.0013***	0.0347**	0.0640*	0.0573*
Amount change	Red county, red state	1:1 versus control	1.9925	0.1050	0.9027	1.0000	1.0000
Amount change	Red county, red state	2:1 versus control	0.1360	0.9070	1.0000	1.0000	1.0000
Amount change	Red county, red state	3:1 versus control	0.4979	0.6983	1.0000	1.0000	1.0000
Amount change	Blue county, red state	1:1 versus control	98.9408	0.4500	1.0000	1.0000	1.0000
Amount change	Blue county, red state	2:1 versus control	101.0547	0.4500	1.0000	1.0000	1.0000
Amount change	Blue county, red state	3:1 versus control	101.4029	0.4500	1.0000	1.0000	1.0000
Amount change	Red county, blue state	1:1 versus control	−55.2064	0.4443	1.0000	1.0000	1.0000
Amount change	Red county, blue state	2:1 versus control	0.1886	0.8600	1.0000	1.0000	1.0000
Amount change	Red county, blue state	3:1 versus control	1.1256	0.2733	0.9970	1.0000	1.0000
Amount change	Blue county, blue state	1:1 versus control	1.1837	0.3410	1.0000	1.0000	1.0000
Amount change	Blue county, blue state	2:1 versus control	−0.0524	0.9717	0.9717	1.0000	0.9717
Amount change	Blue county, blue state	3:1 versus control	0.4172	0.7333	1.0000	1.0000	1.0000

Note: As noted above, these values are computed using only the subsample with all covariates and subgroup identifiers available and may differ from the corresponding values in List et al. (2019).

*Indicates that the corresponding *p*-values are less than 10%.

**Indicates that the corresponding *p*-values are less than 5%.

***Indicates that the corresponding *p*-values are less than 1%.

TABLE 10 Multiple outcomes, subgroups, and treatments, with covariates.

				p-values			
				Unadjusted	Multiplicity adjusted		
Outcome	**Subgroup**	**Comp. groups**	**Coeff.**	**Remark 3.2**	**Theorem 3.1**	**Bonferroni**	**Holm**
Response rate	Red county, red state	1:1 versus control	0.0100	0.0053***	0.1563	0.2560	0.2133
Response rate	Red county, red state	2:1 versus control	0.0100	0.0060***	0.1710	0.2880	0.2340
Response rate	Red county, red state	3:1 versus control	0.0122	0.0013***	0.0373**	0.0640*	0.0587*
Response rate	Blue county, red state	1:1 versus control	0.0029	0.5213	1.0000	1.0000	1.0000
Response rate	Blue county, red state	2:1 versus control	0.0100	0.0460**	0.7107	1.0000	1.0000
Response rate	Blue county, red state	3:1 versus control	0.0124	0.0127**	0.3107	0.6080	0.4433
Response rate	Red county, blue state	1:1 versus control	0.0010	0.7003	1.0000	1.0000	1.0000
Response rate	Red county, blue state	2:1 versus control	0.0022	0.4473	1.0000	1.0000	1.0000
Response rate	Red county, blue state	3:1 versus control	0.0044	0.1570	0.9683	1.0000	1.0000
Response rate	Blue county, blue state	1:1 versus control	−0.0002	0.9413	0.9413	1.0000	0.9413
Response rate	Blue county, blue state	2:1 versus control	0.0031	0.4533	1.0000	1.0000	1.0000
Response rate	Blue county, blue state	3:1 versus control	−0.0025	0.5067	1.0000	1.0000	1.0000
Dollars given not incl. match	Red county, red state	1:1 versus control	0.5161	0.0580*	0.7733	1.0000	1.0000
Dollars given not incl. match	Red county, red state	2:1 versus control	0.4328	0.0610*	0.7763	1.0000	1.0000
Dollars given not incl. match	Red county, red state	3:1 versus control	0.3504	0.0713*	0.8167	1.0000	1.0000
Dollars given not incl. match	Blue county, red state	1:1 versus control	−0.0314	0.9090	0.9940	1.0000	1.0000
Dollars given not incl. match	Blue county, red state	2:1 versus control	0.4814	0.1570	0.9713	1.0000	1.0000
Dollars given not incl. match	Blue county, red state	3:1 versus control	0.6199	0.0980*	0.8953	1.0000	1.0000
Dollars given not incl. match	Red county, blue state	1:1 versus control	0.0254	0.8820	1.0000	1.0000	1.0000
Dollars given not incl. match	Red county, blue state	2:1 versus control	0.1582	0.4023	1.0000	1.0000	1.0000
Dollars given not incl. match	Red county, blue state	3:1 versus control	0.0725	0.6840	1.0000	1.0000	1.0000
Dollars given not incl. match	Blue county, blue state	1:1 versus control	−0.0630	0.7757	1.0000	1.0000	1.0000
Dollars given not incl. match	Blue county, blue state	2:1 versus control	0.0958	0.6943	1.0000	1.0000	1.0000
Dollars given not incl. match	Blue county, blue state	3:1 versus control	−0.1381	0.4987	1.0000	1.0000	1.0000
Dollars given incl. match	Red county, red state	1:1 versus control	1.6947	0.0090***	0.2347	0.4320	0.3240
Dollars given incl. match	Red county, red state	2:1 versus control	2.6234	0.0003***	0.0003***	0.0160**	0.0157**
Dollars given incl. match	Red county, red state	3:1 versus control	3.3892	0.0003***	0.0003***	0.0160**	0.0150**
Dollars given incl. match	Blue county, red state	1:1 versus control	0.7470	0.0997*	0.8913	1.0000	1.0000
Dollars given incl. match	Blue county, red state	2:1 versus control	3.0635	0.0040***	0.1207	0.1920	0.1640
Dollars given incl. match	Blue county, red state	3:1 versus control	4.9089	0.0073***	0.1973	0.3520	0.2787
Dollars given incl. match	Red county, blue state	1:1 versus control	0.8465	0.0087***	0.2287	0.4160	0.3207
Dollars given incl. match	Red county, blue state	2:1 versus control	2.0659	0.0003***	0.0003***	0.0160**	0.0153**
Dollars given incl. match	Red county, blue state	3:1 versus control	2.6770	0.0003***	0.0003***	0.0160**	0.0160**
Dollars given incl. match	Blue county, blue state	1:1 versus control	0.8458	0.0233**	0.4790	1.0000	0.7933
Dollars given incl. match	Blue county, blue state	2:1 versus control	2.2308	0.0030***	0.0880*	0.1440	0.1260
Dollars given incl. match	Blue county, blue state	3:1 versus control	2.3627	0.0013***	0.0363**	0.0640*	0.0573*
Amount change	Red county, red state	1:1 versus control	1.8142	0.1507	0.9697	1.0000	1.0000
Amount change	Red county, red state	2:1 versus control	0.1638	0.8867	0.9987	1.0000	1.0000
Amount change	Red county, red state	3:1 versus control	0.8336	0.5157	1.0000	1.0000	1.0000
Amount change	Blue county, red state	1:1 versus control	118.4810	0.4420	1.0000	1.0000	1.0000
Amount change	Blue county, red state	2:1 versus control	120.4633	0.4390	1.0000	1.0000	1.0000
Amount change	Blue county, red state	3:1 versus control	120.3892	0.4397	1.0000	1.0000	1.0000
Amount change	Red county, blue state	1:1 versus control	−55.7933	0.4443	1.0000	1.0000	1.0000
Amount change	Red county, blue state	2:1 versus control	0.2775	0.8023	1.0000	1.0000	1.0000
Amount change	Red county, blue state	3:1 versus control	1.3375	0.1883	0.9810	1.0000	1.0000
Amount change	Blue county, blue state	1:1 versus control	0.8334	0.5060	1.0000	1.0000	1.0000
Amount change	Blue county, blue state	2:1 versus control	−0.2961	0.8370	1.0000	1.0000	1.0000
Amount change	Blue county, blue state	3:1 versus control	0.2720	0.8263	1.0000	1.0000	1.0000

*Indicates that the corresponding *p*-values are less than 10%.

**Indicates that the corresponding *p*-values are less than 5%.

***Indicates that the corresponding *p*-values are less than 1%.

5 | EPILOGUE

As policymakers continue to increase their demands for evidence and make decisions based on science, the scientific community is relied on to provide sound advice. One core issue that continues to plague the fluid transference of true insights from researchers to policymakers is a large false discovery rate. In this paper, we extend the approach of List et al. (2019) to exploit observed, baseline covariates to obtain more powerful inferences. The methodology builds upon general results in Romano and Wolf (2010) and, with respect to the way in which covariates are included, Ye et al. (2022).

We showcase our methodology by examining how multiple testing affects the insights in Karlan and List (2007). The results are interesting in their own right, as they provide elasticity estimates of import to academics, practitioners, and policymakers. Importantly, we find that multiple testing corrections lead to fewer rejections than when no such adjustments are made, but that using covariates correctly can lead to more powerful inferences.

ACKNOWLEDGMENTS

The research of the second author was supported by the National Science Foundation grant SES-1530661. The third author gratefully acknowledges financial support from the Social Sciences and Humanities Research Council of Canada. We thank the Editor, Edward Vytlacil, and three anonymous referees for their constructive feedback that helped improve and clarify this manuscript. We thank Joe Seidel and Yang Xu for assistance with the code used in List et al. (2019). Joe Romano provided helpful comments concerning the proof of Theorem 3.1. We also thank Alec Brandon, Saminul Haque, David Ledvinka, and David Novgorodsky for comments and suggestions.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at [https://doi.org/10.15456/jae.2023100.1646086306].

REFERENCES

Anderson, L. M., Shinn, C., Fullilove, M. T., Scrimshaw, S. C., Fielding, J. E., Normand, J., & Carande-Kulis, V. G. (2003). The effectiveness of early childhood development programs: A systematic review. *American Journal of Preventive Medicine*, *24*(3, Supplement), 32–46.

Bai, Y., Romano, J., & Shaikh, A. (2021). Inference in experiments with matched pairs. *Journal of the American Statistical Association*, *2021*, 1–37.

Bazylik, S., Mogstad, M., Romano, J. P., Shaikh, A., & Wilhelm, D. (2021). Finite- and large-sample inference for ranks using multinomial data with an application to ranking political parties. National Bureau of Economic Research Working Paper.

Beran, R. (1990). Refining bootstrap simultaneous confidence sets. *Journal of the American Statistical Association*, *85*(410), 417–426.

Berk, R., Pitkin, E., Brown, L., Buja, A., George, E., & Zhao, L. (2013). Covariance adjustments for the analysis of randomized field experiments. *Evaluation Review*, *37*(3-4), 170–196.

Bhattacharya, J., Shaikh, A. M., & Vytlacil, E. (2012). Treatment effect bounds: An application to Swan–Ganz 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. A., Canay, I. A., & Shaikh, A. M. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, *113*(524), 1784–1796.

Bugni, F. A., Canay, I. A., & Shaikh, A. M. (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*, *10*(4), 1747–1785.

Butera, L., Grossman, P. J., Houser, D., List, J. A., & Villeval, M. C. (2020). A new mechanism to alleviate the crises of confidence in science—With an application to the public goods game. (*ID 3598721*). Rochester, NY: Social Science Research Network.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(50), 15343–15347.

Heckman, J. 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., Pinto, R., & Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, *103*(6), 2052–2086.

Heckman, J. J., Pinto, R., & Shaikh, A. (2020). Inference with imperfect randomization: The case of the Perry Preschool Program. (*ID 3656356*). Rochester, NY: Social Science Research Network.

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

Karlan, D., & List, J. A. (2007). Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review*, *97*(5), 1774–1793.
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.
Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, *7*(1), 295–318.
List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, *22*(4), 773–793.
Machado, C., Shaikh, A. M., & Vytlacil, E. J. (2019). Instrumental variables and the sign of the average treatment effect. *Journal of Econometrics*, *212*(2), 522–555.
Maniadis, Z., Tufano, F., & List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review*, *104*(1), 277–290.
Mogstad, M., Romano, J., Shaikh, A., & Wilhelm, D. (2022). Inference for ranks with applications to mobility across neighborhoods and academic achievement across countries. The Review of Economic Studies, forthcoming.
Mogstad, M., Romano, J., Shaikh, A., & Wilhelm, D. (2022). Statistical uncertainty in the ranking of journals and universities. *AEA Papers and Proceedings*, *112*, 630–634.
Negi, A., & Wooldridge, J. (2020). Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*, *40*, 504–534.
Romano, J. P., & Shaikh, A. M. (2006). Stepup procedures for control of generalizations of the familywise error rate. *Annals of Statistics*, *34*(4), 1850–1873.
Romano, J. P., & Shaikh, A. M. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *Annals of Statistics*, *40*(6), 2798–2822.
Romano, J. P., Shaikh, A. M., & Wolf, M. (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling. *TEST. An Official Journal of the Spanish Society of Statistics and Operations Research*, *17*(3), 417.
Romano, J. P., Shaikh, A. M., & Wolf, M. (2008). Formalized data snooping based on generalized error rates. *Econometric Theory*, *24*(2), 404–447.
Romano, J. P., Shaikh, A. M., & Wolf, M. (2010). Hypothesis testing in econometrics. *Annual Review of Economics*, *3*(1), 75–104.
Romano, J. P., Shaikh, A. M., & Wolf, M. (2010). Multiple testing. In Palgrave Macmillan (Ed.), *The new Palgrave dictionary of economics*. London: Palgrave Macmillan UK, pp. 1–5.
Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica : Journal of the Econometric Society*, *73*(4), 1237–1282.
Romano, J. P., & Wolf, M. (2010). Balanced control of generalized error rates. *Annals of Statistics*, *38*(1), 598–633. MR2590052
Tsiatis, A. A., Davidian, M., Zhang, M., & Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, *27*(23), 4658–4677.
Viviano, D., Wuthrich, K., & Niehaus, P. (2022). (When) should you adjust inferences for multiple hypothesis testing? Papers 2104.13367.
Yang, L., & Tsiatis, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *The American Statistician*, *55*(4), 314–321.
Ye, T., Shao, J., Yi, Y., & Zhao, Q. (2022). Toward better practice of covariate adjustment in analyzing randomized clinical trials. *Journal of the American Statistical Association*, *0*(0), 1–13.

How to cite this article: List J. A., Shaikh A. M., & Vayalinkal A. (2023). Multiple testing with covariate adjustment in experimental economics. *Journal of Applied Econometrics*, 1–20. https://doi.org/10.1002/jae.2985

APPENDIX A: 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

$$\begin{aligned}\tilde{\omega}_s &= \left\{P(Q) : Q \in \Omega, E_P[Y_{i,k}|D_i = d, Z_i = z] = E_P[Y_{i,k}|D_i = d', Z_i = z]\right\} \\ &= \left\{P(Q) : Q \in \Omega, \mu_{k|d,z}(P(Q)) = \mu_{k|d',z}(P(Q))\right\} \\ &= \left\{P(Q) : Q \in \Omega, \theta_{k|d,z}(P(Q)) = \theta_{k|d',z}(P(Q))\right\},\end{aligned}$$

since $\mu_{X|d,z}(P(Q)) = \mu_{X|z}(P(Q))$ for any Q satisfying Assumption 2.1.

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 B1, we begin by defining

$$T^*_{s,n}(P) = \sqrt{n}\left(\left(\hat{\theta}_{k|d,z} - \theta_{k|d,z}(P)\right) - \left(\hat{\theta}_{k|d',z} - \theta_{k|d',z}(P)\right)\right),$$

and

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

analogously to List et al. (2019). Next, note that

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_{k|d,z} - \theta_{k|d,z}(P)) &= \sqrt{n}(\bar{Y}_{k|d,z} - \mu_{k|d,z}(P)) - \sqrt{n}\hat{b}'_{k|d,z}(\bar{X}_{d,z} - \bar{X}_z) \\
&\quad + \sqrt{n}b_{k|d,z}(P)'(\mu_{X|d,z}(P) - \mu_{X|z}(P)) \\
&= \sqrt{n}(\bar{Y}_{k|d,z} - \mu_{k|d,z}(P)) - \sqrt{n}\hat{b}'_{k|d,z}\left((\bar{X}_{d,z} - \mu_{X|d,z}(P)) - (\bar{X}_z - \mu_{X|z}(P))\right) \\
&\quad - \sqrt{n}(\hat{b}_{k|d,z} - b_{k|d,z}(P))'(\mu_{X|d,z}(P) - \mu_{X|z}(P)) \\
&= \sqrt{n}(\bar{Y}_{k|d,z} - \mu_{k|d,z}(P)) - \sqrt{n}b_{k|d,z}(P)'\left((\bar{X}_{d,z} - \mu_{X|d,z}(P)) - (\bar{X}_z - \mu_{X|z}(P))\right) \\
&\quad - \sqrt{n}(\hat{b}_{k|d,z} - b_{k|d,z}(P))'\left((\bar{X}_{d,z} - \mu_{X|d,z}(P)) - (\bar{X}_z - \mu_{X|z}(P))\right) \\
&\quad - \sqrt{n}(\hat{b}_{k|d,z} - b_{k|d,z}(P))'(\mu_{X|d,z}(P) - \mu_{X|z}(P)) \\
&= \Delta_{1,k|d,z}(P) - \Delta_{2,k|d,z}(P),
\end{aligned}$$

where

$$\begin{aligned}
\Delta_{1,k|d,z}(P) &= \sqrt{n}(\bar{Y}_{k|d,z} - \mu_{k|d,z}(P)) - \sqrt{n}b_{k|d,z}(P)'\left((\bar{X}_{d,z} - \mu_{X|d,z}(P)) - (\bar{X}_z - \mu_{X|z}(P))\right) \\
\Delta_{2,k|d,z}(P) &= \sqrt{n}(\hat{b}_{k|d,z} - b_{k|d,z}(P))'\left((\bar{X}_{d,z} - \mu_{X|d,z}(P)) - (\bar{X}_z - \mu_{X|z}(P))\right) \\
&\quad + \sqrt{n}(\hat{b}_{k|d,z} - b_{k|d,z}(P))'(\mu_{X|d,z}(P) - \mu_{X|z}(P)).
\end{aligned}$$

For any $P = P(Q)$ such that Q satisfies Assumption 2.1, note that

$$\mu_{X|d,z}(P) - \mu_{X|z}(P) = 0.$$

Furthermore, under Assumptions 2.1–2.3, the WLLN and CMT imply that

$$\hat{b}_{k|d,z} - b_{k|d,z}(P) \xrightarrow{P} 0.$$

Finally, under Assumptions 2.2–2.3, the CLT implies that

$$\sqrt{n}\left(\left(\bar{X}_{d,z} - \mu_{X|d,z}(P)\right) - \left(\bar{X}_z - \mu_{X|z}(P)\right)\right) = O_P(1).$$

Hence,

$$\Delta_{2,k|d,z}(P) \xrightarrow{P} 0.$$

It follows that

$$T^*_n(P) = (\Delta_{1,k|d,z}(P) - \Delta_{1,k|d',z}(P) : s \in \mathcal{S}) + o_P(1).$$

In order to deduce the limiting behavior of $(\Delta_{1,k|d,z}(P) - \Delta_{1,k|d',z}(P) : s \in \mathcal{S})$, note that it may be written as $f(A_n(P), B_n)$, where

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

where $C(P)$ is the $4|\mathcal{S}| \times 2(1 + 2\dim(X_i))|\mathcal{S}|$ matrix formed by stacking diagonally the terms

$$\begin{pmatrix} 1 & -b_{k|d,z}(P)' & 0 & 0 & 0 & 0 \\ 0 & 0 & b_{k|d,z}(P)' & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -b_{k|d',z}(P)' & 0 \\ 0 & 0 & 0 & 0 & 0 & b_{k|d',z}(P)' \end{pmatrix}$$

and $c_i(P)$ is the $2(1 + 2\dim(X_i))|\mathcal{S}|$ dimensional vector formed by stacking vertically for $s \in \mathcal{S}$ the terms

$$\begin{pmatrix} \left(Y_{i,k} - \mu_{k|d,z}(P)\right) I\{D_i = d, Z_i = z\} \\ \left(X_i - \mu_{X|d,z}(P)\right) I\{D_i = d, Z_i = z\} \\ \left(X_i - \mu_{X|z}(P)\right) I\{Z_i = z\} \\ \left(Y_{i,k} - \mu_{k|d',z}(P)\right) I\left\{D_i = d', Z_i = z\right\} \\ \left(X_i - \mu_{X|d',z}(P)\right) I\left\{D_i = d', Z_i = z\right\} \\ \left(X_i - \mu_{X|z}(P)\right) I\{Z_i = z\} \end{pmatrix},$$

B_n is the $4|\mathcal{S}|$-dimensional vector formed by stacking vertically for $s \in \mathcal{S}$ the terms

$$\begin{pmatrix} \frac{n}{n_{d,z}} \\ \frac{n}{n_z} \\ -\frac{n}{n_{d',z}} \\ -\frac{n}{n_z} \end{pmatrix},$$

and $f : \mathbf{R}^{4|\mathcal{S}|} \times \mathbf{R}^{4|\mathcal{S}|} \to \mathbf{R}^{|\mathcal{S}|}$ is the function of $A_n(P)$ and B_n whose sth term is the inner product of sth set of terms defining $A_n(P)$ and B_n. Since $E_P[c_i(P)] = 0$, the CLT, CMT and WLLN allow us to deduce, under Assumptions 2.2–2.4, the limiting behavior of both $A_n(P)$ and B_n. In particular, $B_n(P) \xrightarrow{P} B(P)$, where $B(P)$ is the $4|\mathcal{S}|$-dimensional vector formed by stacking vertically for $s \in \mathcal{S}$ the terms

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

whereas $A_n(P) \xrightarrow{d} N(0, V(P))$ for a suitable variance matrix $V(P)$. It follows from the CMT that

$$f(A_n(P), B_n) \xrightarrow{d} N(0, \Sigma(P)),$$

for a suitable choice of $\Sigma(P)$. Therefore,

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

completing the verification of B1.

To verify B2 and B3, it suffices to show that each of the diagonal terms of $\Sigma(P)$ is non-zero. Some calculation shows that the sth diagonal element of $\Sigma(P)$ is given by

$$\frac{\mathrm{Var}_P[Y_{i,k} - b_{k|d,z}(P)'X_i | D_i = d, Z_i = z]}{P\{D_i = d, Z_i = z\}} + \frac{\mathrm{Var}_P[Y_{i,k} - b_{k|d',z}(P)'X_i | D_i = d', Z_i = z]}{P\{D_i = d', Z_i = z\}} + \frac{\mathrm{Var}_P[\left(b_{k|d,z}(P) - b_{k|d',z}(P)\right)'X_i | Z_i = z]}{P\{Z_i = z\}},$$

which is positive since the last term is non-negative and Assumption 2.4 ensures that the first two terms are positive, completing the verification of B2 and B3.

To verify B4, we first argue that

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

under a suitable sequence of distributions P_n. To this end, assume P_n satisfies:

(a) P_n converges weakly to P
(b) $b_{k|d,z}(P_n) \to b_{k|d,z}(P)$ for all $k \in \mathcal{K}$, $d \in \mathcal{D}$ and $z \in \mathcal{Z}$.
(c) $B_n \xrightarrow{P_n} B(P)$
(d) $\mathrm{Var}_{P_n}[c_i(P_n)] \to \mathrm{Var}_P[c_i(P)]$.
(e) $\Delta_{2,k|d,z}(P_n) \xrightarrow{P_n} 0$ for all $k \in \mathcal{K}$, $d \in \mathcal{D}$ and $z \in \mathcal{Z}$.

Under (b), $C(P_n) \to C(P)$. Using (a) and (d) together with Theorem 15.4.5 in Lehmann and Romano (2006), we have that

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

under P_n. It thus follows from (c) and the CMT that

$$(\Delta_{1,k|d,z}(P_n) - \Delta_{1,k|d',z}(P_n) : s \in \mathcal{S}) \xrightarrow{d} N(0, \Sigma(P)).$$

Finally, (e) implies that

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

under any such sequence P_n. To complete the verification of B4, first note that it suffices to show that for every subsequence n_j there is a further subsequence n_{j_k} such that $\hat{P}_{n_{j_k}}$ satisfies (a) – (e) w.p.1. We provide only a sketch of the argument. To this end, first note that $\hat{P}_n$ satisfies (*a*) by the Glivenko-Cantelli theorem w.p.1 and (b) and (d) w.p.1 by the SLLN and CMT. Arguing element-by-element and applying Lemma 15.4.1 in Lehmann and Romano (2006), it is possible to show that $\hat{P}_n$ satisfies (c) w.p.1. To verify (e), we first argue that it suffices to show for any $\epsilon > 0$ that

$$f(\epsilon, \hat{P}_n) \xrightarrow{P} 0, \tag{A1}$$

where $f(\epsilon, P) = P\{|\Delta_{2,k|d,z}(P)| > \epsilon\}$. To see that the above requirement is sufficient, note that (A1) implies that there is $0 < \epsilon_n \to 0$ sufficiently slowly such that

$$f(\epsilon_n, \hat{P}_n) \xrightarrow{P} 0.$$

The preceding display further implies that for every subsequence n_j there is a further subsequence n_{j_k} such that

$$f(\epsilon_{n_{j_k}}, \hat{P}_{n_{j_k}}) \to 0 \text{ w.p.1},$$

which implies that $\hat{P}_{n_{j_k}}$ satisfies (e) w.p.1. We now return to verifying (A1). By Markov's inequality, it suffices to show that the expected value under P of the left-hand side of (A1) tends to zero. To this end, note that by applying Lemma 15.4.1 in

Lehmann and Romano (2006) it is possible to show that $\hat{b}_{k|d,z}$ tends in probability to $b_{k|d,z}(P)$ under $\hat{P}_n$ w.p.1. The SLLN further implies that $b_{k|d,z}(\hat{P}_n) \to b_{k|d,z}(P)$ w.p.1. It follows that

$$K_{1,n}(\hat{P}_n) \text{ converges weakly to } \delta_0 \text{ w.p.1}, \tag{A2}$$

where $K_{1,n}(P)$ is the distribution of $|\hat{b}_{k|d,z} - b_{k|d,z}(P)|$ under P and δ_0 is the distribution with mass one at zero. Next, let $K_{2,n}(P)$ be the distribution of $|\sqrt{n}((\bar{X}_{d,z} - \mu_{X|d,z}(P)) - (\bar{X}_z - \mu_{X|z}(P)))|$ under P. Note that

$$\sqrt{n}((\bar{X}_{d,z} - \mu_{X|d,z}(P)) - (\bar{X}_z - \mu_{X|z}(P))) = \begin{pmatrix} \frac{n}{n_{d,z}} \\ -\frac{n}{n_z} \end{pmatrix}' \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} \begin{pmatrix} \left(X_i - \mu_{X|d,z}(P)\right) I\{D_i = d, Z_i = z\} \\ \left(X_i - \mu_{X|z}(P)\right) I\{Z_i = z\} \end{pmatrix}.$$

Using this decomposition together with Lemma 15.4.1 and Theorem 15.4.5 in Lehmann and Romano (2006), it is possible to argue that $K_{2,n}(\hat{P}_n)$ converges weakly w.p.1 to the distribution of $|Z|$, where Z is a mean-zero, normally distributed random variable. It follows from the CMT that

$$K_{3,n}(\hat{P}_n) \text{ converges weakly to } \delta_0 \text{ w.p.1}, \tag{A3}$$

where $K_{3,n}(P)$ is the distribution of $|(\hat{b}_{k|d,z} - b_{k|d,z}(P))'\sqrt{n}((\bar{X}_{d,z} - \mu_{X|d,z}(P)) - (\bar{X}_z - \mu_{X|z}(P)))|$ under P. To complete the argument, note that for any constant $c > 0$

$$\begin{aligned} f(\epsilon, P) &= P\{|\Delta_{2,k|d,z}(P)| > \epsilon\} \\ &\le P\left\{|(\hat{b}_{k|d,z} - b_{k|d,z}(P))'\sqrt{n}((\bar{X}_{d,z} - \mu_{X|d,z}(P)) - (\bar{X}_z - \mu_{X|z}(P)))| > \frac{\epsilon}{2}\right\} \\ &\quad + P\left\{|(\hat{b}_{k|d,z} - b_{k|d,z}(P))'\sqrt{n}(\mu_{X|d,z}(P) - \mu_{X|z}(P))| > \frac{\epsilon}{2}\right\} \\ &\le P\left\{|(\hat{b}_{k|d,z} - b_{k|d,z}(P))'\sqrt{n}((\bar{X}_{d,z} - \mu_{X|d,z}(P)) - (\bar{X}_z - \mu_{X|z}(P)))| > \frac{\epsilon}{2}\right\} \\ &\quad + P\left\{|(\hat{b}_{k|d,z} - b_{k|d,z}(P))| > \frac{\epsilon}{2c}\right\} + I\left\{|\sqrt{n}(\mu_{X|d,z}(P) - \mu_{X|z}(P))| > c\right\}. \end{aligned}$$

Hence,

$$\begin{aligned} f(\epsilon, \hat{P}_n) \le &\left(1 - K_{3,n}\left(\frac{\epsilon}{2}, \hat{P}_n\right) + K_{3,n}\left(-\frac{\epsilon}{2}, \hat{P}_n\right)\right) \\ &+ \left(1 - K_{1,n}\left(\frac{\epsilon}{2c}, \hat{P}_n\right) + K_{1,n}\left(-\frac{\epsilon}{2c}, \hat{P}_n\right)\right) + I\left\{|\sqrt{n}(\mu_{X|d,z}(\hat{P}_n) - \mu_{X|z}(\hat{P}_n))| > c\right\}, \end{aligned}$$

where it is understood that $K_{j,n}(x, P)$ is the c.d.f. associated with $K_{j,n}(P)$ for $1 \le j \le 3$. For any fixed $c > 0$, it follows from (A2) – (A3) that the first and second terms on the right-hand side of the preceding display tend to zero w.p.1, while the expected value under P of the last term can be made arbitrarily small by choosing c sufficiently large since, using arguments akin to those given above,

$$\sqrt{n}(\mu_{X|d,z}(\hat{P}_n) - \mu_{X|z}(\hat{P}_n)) = \sqrt{n}((\mu_{X|d,z}(\hat{P}_n) - \mu_{X|d,z}(P)) - (\mu_{X|z}(\hat{P}_n) - \mu_{X|z}(P)) \xrightarrow{d} N(0, W(P))$$

for a suitable choice of $W(P)$, where we have exploited the fact that $\mu_{X|d,z}(P) = \mu_{X|z}(P)$. It now follows from the dominated convergence theorem that the expected value under P of the left-hand side of (A1) tends to zero, as desired.