Statistica Sinica **27** (2017), 1319-1345
doi:https://doi.org/10.5705/ss.202016.0116

MORE POWERFUL MULTIPLE TESTING IN RANDOMIZED EXPERIMENTS WITH NON-COMPLIANCE

Joseph J. Lee[1], Laura Forastiere[2], Luke Miratrix[1] and Natesh S. Pillai[1]

[1]*Harvard University and* [2]*University of Florence*

Abstract: Two common concerns raised in analyses of randomized experiments are (i) appropriately handling issues of non-compliance, and (ii) appropriately adjusting for multiple tests (e.g., on multiple outcomes or subgroups). Although simple intention-to-treat (ITT) and Bonferroni methods are valid in terms of type I error, they can each lead to a substantial loss of power; when employing both simultaneously, the total loss may be severe. Alternatives exist to address each concern. Here we propose an analysis method for experiments involving both features that merges posterior predictive p-values for complier causal effects with randomization-based multiple comparisons adjustments; the results are valid familywise tests that are doubly advantageous: more powerful than both those based on standard ITT statistics and those using traditional multiple comparison adjustments. The operating characteristics and advantages of our method are demonstrated through a series of simulated experiments and an analysis of the United States Job Training Partnership Act (JTPA) Study, where our methods lead to different conclusions regarding the significance of estimated JTPA effects.

Key words and phrases: Causal inference, hypothesis testing, multiple comparisons, posterior predictive p-value, principal stratification, randomization-based inference.

1. Introduction

The United States Job Training Partnership Act (JTPA) Study was a randomized experiment in the 1980s designed to measure the effects of a national, publicly-funded training program. Participants randomly assigned to the treatment group were eligible to receive JTPA services, while participants randomly assigned to the control group were barred from JTPA services for 18 months. Only about 2/3 of the treatment participants, however, actually enrolled and received any JTPA services; the other 1/3 failed to comply with their treatment assignment. Furthermore, because of the fluid nature of the participants' employment, researchers were interested in measuring JTPA effects across several time periods after random assignment, including the in-training period and the first and second post-program years. Analyzing such data requires addressing two

substantial concerns: due to non-compliance, the effects of treatment assignment are not equivalent to the effects of treatment receipt, and conducting tests for multiple time periods without appropriate adjustments may lead to an inflated type I error rate. In this paper, we outline an analysis method that addresses both concerns while maintaining reasonable power to detect treatment effects.

When units in randomized experiments fail to comply with their random assignment, inference for the effects of treatment receipt, rather than of assignment alone, becomes less straightforward. Intention-to-treat (ITT) analyses, which ignore treatment receipt, may have low power when assignment alone has no effect on the experimental outcome. In order to address this loss of power, Rubin (1998) introduced randomization-based posterior predictive p-values for the complier average causal effect (CACE) and showed, through simulation, that they are valid p-values in terms of type I error and that their tests have higher power than tests using ITT p-values under reasonable alternative hypotheses (e.g., hypotheses with non-zero treatment effects for units who are assigned to and receive treatment, but zero treatment effects for units who do not receive it). This framework follows the general approach for Bayesian causal inference in randomized experiments with non-compliance outlined by Imbens and Rubin (1997). Both pieces of work rely on the multiple imputation (Rubin (1987)) of missing compliance statuses; separating the experimental units into principal strata (Frangakis and Rubin (2002)) based on compliance behavior aids inference for the desired causal effect. We use these tools in our approach but adapt them for simultaneous testing of multiple outcomes and subgroups.

Multiple testing issues are common in randomized experiments because multiple outcomes and subgroups of interest are often measured and analyzed for possible effects. Traditionally, practitioners have applied Bonferroni corrections to sets of p-values in order to control their familywise error rate (FWER), i.e., the rate at which at least one type I error is made, in a straightforward manner. Bonferroni corrections, however, tend to be overly conservative, especially when those p-values are correlated (Westfall and Young (1989)). This fact has led many applied researchers to avoid Bonferroni corrections and abandon multiple comparisons adjustments altogether (Cabin and Mitchell (2000); Nakagawa (2004); Perneger (1998); Rothman (1990)). Other avenues exist; randomization-based procedures can provide greater power while maintaining the FWER by accounting for correlated tests. Brown and Fears (1981) and Westfall and Young (1989) first introduced permutation-based multiple testing adjustments, though they did not explicitly motivate them using randomized assignment mechanisms.

Randomization-based procedures are additionally appealing because they do not require any assumptions about the underlying distribution (here, joint) of the data. Furthermore, recent increases in computational power have helped such procedures become more tractable and gain popularity (Good (2005)).

In this article, we connect methodological ideas to appropriately handle both non-compliance and multiple testing in randomized experiments. We build up to this combined approach in stages. In Section 2, we elucidate the method proposed by Rubin (1998) for evaluating meaningful causal effects in the presence of non-compliance. In Section 3, we extend the ideas of Westfall and Young (1989) to fully randomization-based multiple comparisons adjustments and propose such adjustments as a straightforward yet more powerful alternative to Bonferroni corrections. In Section 4, we merge the notions of non-compliance and multiple testing, and outline a combined method of analysis that demonstrates power advantages from both perspectives. In each of Sections 2–4, we empirically show the benefits of the described methods through a series of simulated experiments. In Section 5, we apply traditional methods and our combined method to JTPA data to evaluate the program's effects on employment rate by time period. We illustrate how the methods lead to different conclusions regarding the significance of estimated JTPA effects. Section 6 concludes.

2. Experiments with Non-compliance

2.1. Non-compliance as a missing data problem

Suppose we have a randomized experiment with N units, indexed by i, with observed covariates X_i, randomly assigned to control or active treatment. Let Z_i be a binary indicator for assignment to active treatment, and let $D_i(z)$ be a binary indicator for receipt of active treatment under assignment z. A unit's compliance behavior C_i is defined by the pair of potential outcomes (Neyman (1923); Rubin (1974)) $(D_i(0), D_i(1))$; this notation is adequate under the stable unit treatment value assumption (Rubin (1980, 1986)), which asserts no interference between experimental units, as well as two well-defined outcomes. Each unit then belongs to one of four possible compliance strata:

- Compliers ($C_i = c$), who receive their treatment assignment: $(D_i(0), D_i(1)) = (0, 1)$.
- Never-takers ($C_i = nt$), who never receive the active treatment: $(D_i(0), D_i(1)) = (0, 0)$.

Table 1. Units' possible compliance strata based on observed treatment assignment and receipt.

Assignment Z_i	Receipt D_i^{obs}	Possible C_i Values	
		One-sided Non-compliance	Two-sided Non-compliance
0	0	c, nt	c, nt
0	1	–	at, d
1	0	nt	nt, d
1	1	c	c, at

- Always-takers ($C_i = at$), who always receive the active treatment: $(D_i(0), D_i(1)) = (1, 1)$.
- Defiers ($C_i = d$), who receive the opposite of their treatment assignment: $(D_i(0), D_i(1)) = (1, 0)$.

If non-compliance is one-sided — i.e., units assigned to control are prohibited from receiving the active treatment — then $D_i(0) = 0$ for all i. In such settings, always-takers and defiers do not exist, and two possible strata are left: compliers and never-takers. Real-world scenarios involving one-sided non-compliance include many clinical trials, in which new drugs are unavailable to control patients, and some job training experiments, in which training programs and additional services are unavailable to the control group.

In many practical settings, researchers are most interested in the compliers because the effect of treatment assignment is synonymous with the effect of treatment receipt for those units. Strata membership, however, can never be fully determined for all units because they depend on the two potential outcomes of D, one of which is unobserved. Membership can, on the other hand, be partially determined based on the observed potential outcome, D_i^{obs}. Table 1 outlines the possible compliance strata based on units' observed treatment assignment and receipt. An example "Science" table (Rubin (2005)) under one-sided non-compliance and its observed values under a particular assignment are shown in Table 2.

Because strata memberships are not fully observed, uncertainty with respect to complier-specific effects stems from the missing compliance statuses (i.e., D potential outcomes) in addition to the missing Y potential outcomes. One approach to handling the additional uncertainty is to, in a Bayesian framework, view the missing compliance statuses as random variables. By multiply imputing the missing compliance statuses, e.g., according to a distributional model, they can be integrated out, and we can make inference specific to the compliers.

Table 2. An example Science table under one-sided non-compliance (left) and its corresponding observed and unobserved values under a particular assignment (right).

		$D(z)$		Compliance	$Y(z)$		Assignment	$D(z)$		Compliance	$Y(z)$	
Unit	X_i	$D_i(0)$	$D_i(1)$	C_i	$Y(0)$	$Y(1)$	Z_i	$D_i(0)$	$D_i(1)$	C_i	$Y_i(0)$	$Y_i(1)$
1	X_1	0	0	nt	$Y_1(0)$	$Y_1(1)$	0	0	?	?	Y_1^{obs}	?
2	X_2	0	1	c	$Y_2(0)$	$Y_2(1)$	1	0	1	c	?	Y_2^{obs}
3	X_3	0	1	c	$Y_3(0)$	$Y_3(1)$	1	0	1	c	?	Y_3^{obs}
4	X_4	0	0	nt	$Y_4(0)$	$Y_4(1)$	1	0	0	nt	?	Y_4^{obs}
...				...					...			
N	X_N	0	1	c	$Y_N(0)$	$Y_N(1)$	0	0	?	?	Y_N^{obs}	?

2.2. Randomization-based posterior predictive p-values

As described by Meng (1994), a posterior predictive p-value can be viewed as the posterior mean of a classical p-value, averaging over the posterior distribution of nuisance factors (e.g., missing compliance statuses) under the null hypothesis. Rubin (1998) introduced a randomization-based procedure, which we expound on here, for obtaining posterior predictive p-values for estimated complier-only effects. One posterior predictive p-value is the average of many p-values calculated from multiple "compliance-complete" datasets with imputed compliance statuses; for each compliance-complete dataset, the p-value is obtained through a randomization test (Fisher (1925, 1935)).

Within one randomization test, however, calculations of the test statistic do not use all of the compliance information from the compliance-complete data; rather, they use only the compliance information that would have actually been observed under particular hypothetical randomizations. Though implied, this step of re-observing the data is not explicitly stated by Rubin (1998); we place it in Step 5 of our procedure for emphasis because it is an important prerequisite for conducting a proper test. Unlike discrepancy variables (Meng (1994)), which may depend on unobserved factors (e.g., missing compliance statuses), test statistics must be functions of only the observed data. In order to conduct a proper test, the true observed test statistic value must be measured against the correct distribution, i.e., the distribution of that same test statistic.

In this section, we assume a single outcome for simplicity. The procedure for obtaining a randomization-based posterior predictive p-value is as follows.

1. **Choose a test statistic and calculate its observed value.**
   Choose a test statistic, T, to estimate an effect on the outcome, Y. Calculate T on the observed data to obtain T^{obs}.
   Examples include the maximum-likelihood estimate (MLE) of CACE or the posterior median of CACE, given the observed compliance statuses and

potential outcomes, under the exclusion restriction (see Angrist, Imbens and Rubin (1996); Imbens and Rubin (1997)).

for $m : 1$ to M **do**

2. **Impute missing compliance statuses.**
   Impute the missing compliance statuses, drawing once from their posterior predictive distribution according to a compliance model that assumes the null hypothesis (e.g., of zero effect).
3. **Impute missing potential outcomes.**
   Impute the missing Y potential outcomes under the sharp null hypothesis. Under the typical sharp null hypothesis of zero treatment effect, the missing potential outcome for unit i is imputed exactly as Y_i^{obs}.
4. **Draw a random hypothetical assignment.**
   Draw a random hypothetical assignment vector according to the assignment mechanism used in the original experiment.
5. **Re-observe the data.**
   Treating the imputed compliance statuses, imputed potential outcomes, and hypothetical assignment vector from Steps 2–4 as true, create a corresponding hypothetical observed dataset by masking the potential outcomes and compliance statuses that would not have been observed under the hypothetical assignment.
6. **Calculate the test statistic on these data.**
   Calculate T on the hypothetical observed data to obtain T^{hyp}. Record whether this statistic is at least as extreme as T^{obs}.

end for

7. **Calculate the posterior predictive p-value.**
   The posterior predictive p-value for the null hypothesis with respect to T equals the proportion of the M imputation-randomization sets for which T^{hyp} is as extreme as or more extreme than T^{obs}.

Rubin (1998) discusses several commonly used statistics for evaluating complier causal effects, only some of which tend to estimate the CACE and thus provide suitable power against appropriate alternative hypotheses. As is commonly done in the non-compliance literature, we assume the exclusion restriction (i.e., we assume that treatment assignment has no effect on the outcomes of never-takers and always-takers) for test statistic calculations throughout this paper. Such an assumption is not necessary and does not affect the validity of

the randomization test, but it does facilitate more precise estimation of CACE when true (see Imbens and Rubin (1997)) and is often reasonable.

The imputation in Step 2 is performed probabilistically, using the missing statuses' null posterior predictive distribution, given X, Z, D^{obs}, and Y^{obs}. (Some test statistics, such as the posterior median of CACE, may be computed by multiply imputing the missing compliance statuses. This would be a separate imputation from the one described in Step 2 above. If the test statistic calculation itself involves imputation, such imputation does not need to, and usually does not, assume the null hypothesis.) The repetition of Steps 2–6 is intended to reflect the uncertainty of estimation resulting from the missing compliance statuses; M is a large number (e.g., $10,000$) that controls the Monte Carlo integration error.

Under the null hypothesis, Y is not affected by assignment to or receipt of the active treatment; it is therefore treated like a covariate in the imputation model. Even in the absence of other covariates (X), Y alone may still be successful in stochastically identifying the missing compliance statuses, thus providing tests of CACE with power over ITT tests (see Section 2.3). When additional covariates that affect compliance status supplement Y in the imputation model (e.g., in a Bayesian generalized linear model), the compliance identification tends to sharpen, providing CACE tests with greater power.

In settings with one-sided non-compliance, only the compliance statuses of units assigned to the control group are missing. Let ω_c be the super-population proportion of compliers, and let $\boldsymbol{\eta} = (\eta_c, \eta_n)$ be the parameters that govern the outcome distributions of compliers and never-takers, respectively. Note that under the null hypothesis, these are only two outcome distributions; units within a compliance stratum have the same outcome distributions, regardless of their treatment assignment. The posterior predictive distribution of the missing compliance statuses can be obtained using a two-step data augmentation algorithm (Tanner and Wong (1987)). Using the current (or initial, if starting the algorithm) values of the parameters, the missing compliance statuses are drawn according to Bayes' rule:

$$P(C_i{=}c|Y_i^{\text{obs}}, X_i, Z_i{=}0, D_i^{\text{obs}}{=}0, \omega_c, \boldsymbol{\eta}) = \frac{\omega_c g_c(Y_i^{\text{obs}}; \eta_c)}{\omega_c g_c(Y_i^{\text{obs}}; \eta_c) + (1-\omega_c) g_n(Y_i^{\text{obs}}; \eta_n)}, \tag{2.1}$$

where $g_c(y; \eta_c)$ and $g_n(y; \eta_n)$ are the outcome probabilities (or densities) of y for compliers and never-takers, respectively. Once the missing compliance statuses are drawn, new parameter values are drawn from their compliance-complete-data posterior distributions. These two steps are alternated until distributional

Table 3. The observed values of the Science table from Table 2, with the missing Y potential outcomes imputed under the sharp null hypothesis of zero treatment effect. Imputed Y potential outcomes are in parentheses.

		Assignment	$D(z)$		Compliance status	$Y(z)$	
Unit	X_i	Z_i	$D_i(0)$	$D_i(1)$	C_i	$Y_i(0)$	$Y_i(1)$
1	X_1	0	0	?	?	Y_1^{obs}	(Y_1^{obs})
2	X_2	1	0	1	c	(Y_2^{obs})	Y_2^{obs}
3	X_3	1	0	1	c	(Y_3^{obs})	Y_3^{obs}
4	X_4	1	0	0	nt	(Y_4^{obs})	Y_4^{obs}
...					...		
N	X_N	0	0	?	?	Y_N^{obs}	(Y_N^{obs})

convergence. After convergence, the draws of the missing compliance statuses can be treated as posterior predictive imputations. Obtaining posterior draws of parameters — and consequently, posterior predictive draws of the missing compliance statuses — may be more straightforward if models are conjugate, e.g., Beta-Binomial or Dirichlet-Multinomial models (see Section 2.3).

For each imputation of the missing compliance statuses, a randomization test (here involving only one random hypothetical assignment for computational efficiency) is performed in Steps 3–6. Because p-values are defined as conditional probabilities given that the sharp null hypothesis is true, the imputation of Y potential outcomes in Step 3 must occur under this hypothesis. Table 3 shows the observed values of the Science table from Table 2, with the Y potential outcomes imputed under the sharp null hypothesis of zero treatment effect. For computational efficiency, Step 3 can be performed just once (before the loop) because this imputation is deterministic.

The random draw of a hypothetical assignment vector in Step 4 depends on the specific assignment mechanism used in the experiment, e.g., complete randomization or block randomization. A seemingly alternative procedure to the one described above switches the order of Steps 2 and 4, such that the hypothetical assignment vector is drawn first, and the missing compliance statuses are imputed second. This alternative procedure, however, is exactly equivalent to the one described above because the imputation of the missing compliance statuses under the null hypothesis is influenced by Z only through C^{obs}. Because C^{obs} is fixed by the actual observed data, reversing the order of Steps 2 and 4 does not affect the overall inferential procedure. Intuitively, we can consider the posterior predictive p-value as a double integral over the missing compliance statuses and the randomization; switching the order of integration does not affect the result.

2.3. Illustrative simulations with non-compliance

Consider this modified example from Rubin (1998): suppose a completely randomized double-blind experiment is conducted to investigate the effect of a new drug (provided in addition to standard care) versus standard care alone on Y, which measures the severity of patients' heart attacks in the year following treatment. Y is ordinal, taking on values of 0, 1, and 2 (no, mild, and severe attacks, respectively). We assume that all of the patients survive through the year. We also assume one-sided non-compliance, so our experiment has two groups of patients: compliers and never-takers.

In our simulation, we randomly selected $N = 1,000$ units from a super-population of 10% compliers and 90% never-takers; the compliers tend to be healthier than the never-takers. We randomly assigned $N/2 = 500$ units to control and $N/2$ units to active treatment, observing only the compliance statuses of units assigned to active treatment. For each unit, we generated an observed Multinomial outcome, Y_i^{obs}, according to the specified treatment effect hypothesis. Simulation details are provided in Appendix A.1.

Using the simulated observed data, we calculated two test statistics: the ITT statistic, and the MLE of CACE under the exclusion restriction. We then calculated randomization-based posterior predictive p-values for both test statistics, as described in Section 2.2, under the null hypothesis of zero treatment effect. (For the multiple imputation of the missing compliance statuses, we placed conjugate Beta$(1, 1)$ priors on the parameters governing the complier and never-taker outcome distributions.) To evaluate the frequency characteristics of the posterior predictive p-values, we ran 1,000 replications of the data simulation and p-value procedures. Under the null hypothesis, p-values for the two statistics both appeared valid in terms of type I error; their empirical distributions were approximately uniform. At the $\alpha = .05$ level, tests on ITT and CACE rejected the null hypothesis in 4.5% and 4.1% of simulations, respectively. Under the alternative hypothesis, tests based on the CACE are more powerful (see Figure 1), with tests on ITT and CACE rejecting the null hypothesis in 16.7% and 25.2% of simulations, respectively, at $\alpha = .05$. In a general setting, the magnitude of the power gain from the CACE depends on the proportion of compliers, the magnitude of the treatment effect, and the α level.

3. Experiments with Multiple Testing

3.1. Randomization-based multiple comparisons adjustments

Suppose we have data from a randomized experiment with J estimands and

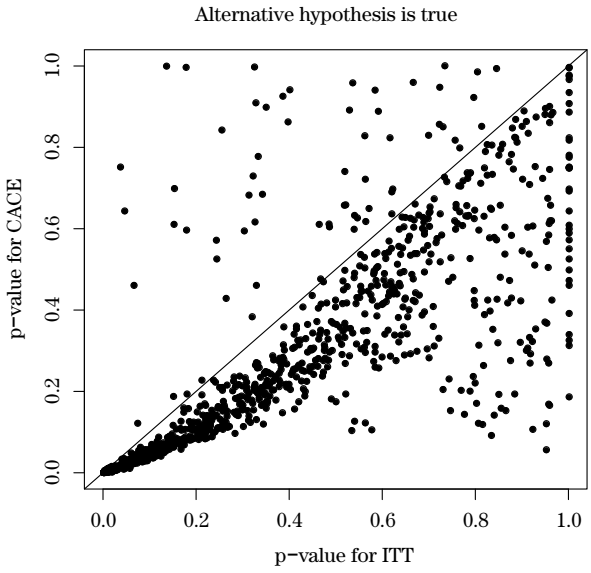

Figure 1. Joint distribution of 1,000 posterior predictive p-values for ITT and CACE estimates under the alternative hypothesis. Tests for CACE are more powerful because p-values for CACE tend to be lower.

are interested in testing whether the active treatment has any non-null effects. The desire for J estimands may result, for example, from multiple outcomes per unit or from multiple, potentially overlapping subgroups of units. Brown and Fears (1981) and Westfall and Young (1989) first proposed permutation-based multiple comparisons adjustments, with the latter showing that such adjustments outperform traditional (e.g., Bonferroni) adjustments in terms of power. They did not, however, explicitly motivate their methods using randomized assignment mechanisms and joint randomization distributions. Furthermore, they assumed specific models that facilitated the calculation of nominal (unadjusted) p-values and implicitly assumed completely randomized assignments throughout.

Here we extend their ideas to a fully randomization-based procedure for multiple comparisons adjustments. Our procedure is connected to — and directly motivated by — the actual randomized assignment mechanism used in the experiment; in addition, both the nominal and adjusted p-values in our procedure are randomization-based, so we do not require any assumptions about the underlying distribution of the data. We calculate fully randomization-based adjusted p-values as follows.

1. **Choose test statistics and calculate their observed values.**
   Choose test statistics, $(T_1, \ldots, T_J)$, and calculate $(T_1^{\text{obs}}, \ldots, T_J^{\text{obs}})$ on the

observed data.

2. **Impute missing potential outcomes.**
   Impute the missing potential outcomes under the sharp null hypothesis.

3. **Calculate nominal p-values for the observed test statistics.**
   For $j = 1, \ldots, J$, calculate the randomization-based p-value for T_j^{obs} by repeatedly (i.e., M' times) drawing a random hypothetical assignment vector according to the assignment mechanism, and calculating the test statistic, T_j^{hyp}, for the corresponding hypothetical observed data. The nominal, marginal randomization-based p-value for T_j^{obs} $(j = 1, \ldots, J)$ equals the proportion of T_j^{hyp} values that are as extreme as or more extreme than T_j^{obs}. Record the $(T_1^{\text{hyp}}, \ldots, T_J^{\text{hyp}})$ values for use in Step 4.

4. **Calculate nominal (marginal) p-values for the hypothetical test statistics.**
   Using the M' sets of $(T_1^{\text{hyp}}, \ldots, T_J^{\text{hyp}})$ values from Step 3, calculate a nominal randomization-based p-value for each T_j^{hyp} and record the minimum of the p-values for each of the M' sets.

5. **Obtain the joint randomization distribution of the nominal p-values.**
   For large M', the repetitions of Step 4 appropriately capture the joint randomization distribution of the test statistics and thus, of the nominal p-values.

6. **Calculate adjusted p-values for the observed test statistics.**
   The adjusted p-value (Westfall and Young (1989)) for T_j^{obs} $(j = 1, \ldots, J)$ equals the proportion of hypothetical observed datasets for which the minimum of the J nominal p-values for $(T_1^{\text{hyp}}, \ldots, T_J^{\text{hyp}})$ is less than or equal to the nominal p-value for T_j^{obs}.

Steps 4–5 essentially represent a translation, i.e., re-scaling, of hypothetical test statistics — which may have different scales — into hypothetical p-values, which share a common 0–1 scale. Our procedure results in individual adjusted p-values that are corrected for the FWER but are also directly interpretable on their own.

Equivalently, to determine α-level significance, we can compare each nominal p-value to the familywise α-level cutoff: the α-th quantile of the minimums recorded in Step 4. The probability that no type I errors are made (i.e., that

we fail to reject all J tests under the null hypothesis) is equivalent to the probability that all J observed marginal p-values are above the cutoff. This equals the probability that the minimum of the J observed p-values is above the cutoff, which is $1-\alpha$ by construction. Thus, the probability of at least one type I error — the FWER — is α, as desired.

Randomization-adjusted p-values are more powerful than traditional Bonferroni-adjusted p-values, especially when the correlations among the J test statistics are high, as shown by simulations. Intuitively, suppose the null hypothesis of zero effects is true and that we have a large number of uncorrelated test statistics; the probability of at least one type I error is quite high because of the number of tests being conducted. Now suppose instead that those test statistics are highly correlated; the probability of at least one type I error is reduced because the tests' type I errors are likely to occur simultaneously, i.e., for the same random assignments. Bonferroni adjustments in these settings are the same, simply counting the number of p-values being examined. In contrast, by utilizing the joint distribution of the nominal p-values, the randomization-based adjustments account for the correlations among test statistics and are less conservative.

3.2. Illustrative simulations with multiple testing

We follow the experimental setup from Section 2.3, modified to include multiple outcomes but without non-compliance. Suppose that researchers now want to investigate the effect of the new drug on three outcomes: $Y_{\cdot 1}, Y_{\cdot 2}$, and $Y_{\cdot 3}$ (with the first subscript denoting the participant), which measure the severity of heart attacks (defined as before) in the first, second, and third year after treatment, respectively. We assume that all of the patients survive through the third year, and we would like to see whether the drug has an effect on heart attack severity at any of the three time points.

To evaluate the frequency characteristics of the adjusted randomization-based p-values, we simulated 1,000 datasets under both null and alternative hypotheses according to each of three outcome correlation structures: zero, partial (approximately 0.5), and perfect correlation. The specific data generation processes are found in in Appendices A.2 and B. The correlations among $Y_{i1}(z)$, $Y_{i2}(z)$, and $Y_{i3}(z)$ $(z = 0, 1)$ are important; however, for a fixed j, the correlation between $Y_{ij}(0)$ and $Y_{ij}(1)$ is inconsequential to the simulation because we only ever observe one of the potential outcomes.

For each simulated dataset, we calculated fully randomization-based adjusted

Table 4. Proportions of multiple testing simulations in which the null hypothesis was rejected, under various data generation processes. Based on 1,000 replications.

	Rejection Rate at $\alpha = .05$			
	Null is true		Alternative is true	
	Bonferroni	Randomization-based	Bonferroni	Randomization-based
Zero correlation	0.042	0.046	0.908	0.919
Partial correlation	0.045	0.053	0.787	0.811
Perfect correlation	0.024	0.045	0.557	0.720

p-values and decided whether or not to reject the null hypothesis of zero treatment effects across the three time periods at $\alpha = .05$. For comparison, we also decided whether or not to reject the null hypothesis using Bonferroni-adjusted p-values. Simulation results under both null and alternative hypotheses are shown in Table 4. Without sacrificing validity under the null hypothesis, the randomization-based adjustment displays greater power than the Bonferroni adjustment under the alternative hypothesis, particularly for scenarios with high correlations among outcomes.

4. Experiments with Both Non-compliance and Multiple Testing

It is natural to merge the analysis methods presented in Sections 2 and 3 — both of which use the randomized assignment mechanism to aid inference — for experiments involving both non-compliance and multiple testing. The results are valid familywise tests that are more powerful from both perspectives: more powerful than both those based on standard ITT statistics and those using traditional multiple comparison adjustments.

Suppose again that we have data from a randomized experiment with J estimands and that we are interested in testing whether the active treatment has any non-null effects. However, not all units comply to their treatment assignments; assume for simplicity that non-compliance is one-sided. In Section 2, Table 2 displays the observed values of a Science table with two Y potential outcomes — one observed and one missing — for each unit. Here, Table 5 shows the corresponding observed values of a Science table with multiple estimands resulting from $J = 3$ outcomes of interest. Each unit has six potential outcomes, only three of which are observed; the other three are missing. Within unit i, we observe the same member of $(Y_{ij}(0), Y_{ij}(1))$ for each outcome j, e.g., if we observe $Y_{i1}(1)$, then we also observe $Y_{i2}(1)$ and $Y_{i3}(1)$.

In experiments with non-compliance and multiple testing, obtaining valid and more powerful familywise tests involves (i) calculating (nominal) posterior

Table 5. Observed and unobserved values of the Science table from Table 2, now with three outcomes of interest. Missing (unobserved) data are denoted by question marks.

		Assignment	$D(z)$		Compliance status	$Y_{\cdot 1}(z)$		$Y_{\cdot 2}(z)$		$Y_{\cdot 3}(z)$	
Unit	X_i	Z_i	$D_i(0)$	$D_i(1)$	C_i	$Y_{i1}(0)$	$Y_{i1}(1)$	$Y_{i2}(0)$	$Y_{i2}(1)$	$Y_{i3}(0)$	$Y_{i3}(1)$
1	X_1	0	0	?	?	Y_{11}^{obs}	?	Y_{12}^{obs}	?	Y_{13}^{obs}	?
2	X_2	1	0	1	c	?	Y_{21}^{obs}	?	Y_{22}^{obs}	?	Y_{23}^{obs}
3	X_3	1	0	1	c	?	Y_{31}^{obs}	?	Y_{32}^{obs}	?	Y_{33}^{obs}
4	X_4	1	0	0	nt	?	Y_{41}^{obs}	?	Y_{42}^{obs}	?	Y_{43}^{obs}
...			...					...			
N	X_N	0	0	?	?	Y_{N1}^{obs}	?	Y_{N2}^{obs}	?	Y_{N3}^{obs}	?

predictive p-values for CACE according to the procedure in Section 2, and (ii) calculating adjusted posterior predictive p-values using the joint randomization distribution of the nominal p-values, according to the procedure in Section 3. Intuitively, this combined method of analysis is preferable because Steps (i) and (ii) provide power gains through distinct and unrelated mechanisms, and neither sacrifices validity in terms of type I error. For the J estimands, we expect each individual (nominal) CACE p-value to be more powerful than its ITT counterpart based on the arguments in Section 2. Furthermore, given a set of J nominal p-values, we expect randomization-adjusted p-values using the nominal p-values' joint randomization distribution to be more powerful than Bonferroni-adjusted p-values, as argued in Section 3. Naturally, adjusting more powerful nominal p-values in a more powerful manner results in adjusted p-values that are doubly advantageous in terms of power. The full procedure is detailed below.

1. **Choose test statistics and calculate their observed values.**
   Choose test statistics, $(T_1, \ldots, T_J)$, and calculate $(T_1^{\text{obs}}, \ldots, T_J^{\text{obs}})$ on the actual observed data.

for $i : 1$ to M **do**

2. **Impute missing compliance statuses.**
   Impute the missing compliance statuses, drawing once from their posterior predictive distribution according to a compliance model that assumes the null hypothesis.
3. **Impute missing potential outcomes.**
   Impute all of the missing $(Y_1, \ldots, Y_J)$ potential outcomes under the sharp null hypothesis.
4. **Draw a random hypothetical assignment.**
   Draw a random hypothetical assignment vector according to the assignment mechanism.

5. **Re-observe the data.**
   Treating the imputed compliance statuses and potential outcomes and the hypothetical assignment vector as true, create a corresponding hypothetical observed dataset by masking the potential outcomes and compliance statuses that would not have been observed under the hypothetical assignment.
6. **Calculate test statistics on the hypothetical observed data.**
   Calculate $(T_1, \ldots, T_J)$ on the hypothetical observed data to obtain $(T_1^{\text{hyp}}, \ldots, T_J^{\text{hyp}})$. For $j = 1, \ldots, J$, record whether T_j^{hyp} is at least as extreme as T_j^{obs}.

end for

7. **Calculate nominal (marginal) posterior predictive p-values for the observed test statistics.**
   For $j = 1, \ldots, J$, the nominal (marginal) posterior predictive p-value for the null hypothesis with respect to the test statistic T_j equals the proportion of the M imputation-randomization sets created by Steps 2–6 for which T_j^{hyp} is as extreme as or more extreme than T_j^{obs}.
8. **Calculate nominal posterior predictive p-values for the hypothetical test statistics and obtain the joint randomization distribution of the nominal posterior predictive p-values.**
   For each of the M imputation-randomization sets, translate the hypothetical test statistics $(T_1^{\text{hyp}}, \ldots, T_J^{\text{hyp}})$ into hypothetical nominal posterior predictive p-values using proportions similar to the one described in Step 7. This step is a computationally efficient way of obtaining the joing distribution of hypothetical test statistics on a common p-value scale, analogous to Steps 4–5 from the procedure in Section 3. Record the minimum of each set of nominal p-values.
9. **Calculate adjusted posterior predictive p-values for the observed test statistics.**
   The adjusted posterior predictive p-value for T_j^{obs} $(j = 1, \ldots, J)$ equals the proportion of the M imputation-randomization sets for which the minimum of the J nominal posterior predictive p-values for $(T_1^{\text{hyp}}, \ldots, T_J^{\text{hyp}})$ is less than or equal to the nominal (marginal) posterior predictive p-value for T_j^{obs}.

Under the null hypothesis, the outcomes $Y_{\cdot 1}, \ldots, Y_{\cdot J}$ inform the multiple imputation of the missing compliance statuses. Posterior predictive imputations

of the missing compliance statuses can be generated using a data augmentation algorithm similar to the one described in Section 2, with Equation (2.1) modified to use the joint set of J observed outcomes.

4.1. Illustrative simulations with both non-compliance and multiple testing

Again consider the heart treatment example from Sections 2.3 and 3.2: we would like to see whether the active treatment has an effect on heart attack severity at any of the three time points after treatment. In these simulations, we assumed one-sided non-compliance, with $N = 1,000$ units randomly sampled from super-populations with 10% and 30% compliers. We also ran simulations with 50% and 70% compliance rates, but almost all of the tests were able to detect treatment effects under the alternative hypotheses, so the comparison tables were less meaningful. Alternative hypotheses 1, 2, and 3, in that order, assumed treatment effects of increasing size. The data generation processes are described in Appendices A.3 and B.

For each simulated dataset, a total of 10 familywise tests were conducted. Five of the tests used the ITT test statistic: one used the Bonferroni correction and one used the randomization-based multiple comparisons adjustment. The other three ITT tests used multiple comparisons adjustments proposed as alternatives to the Bonferroni correction, by Holm (1979), Hochberg (1988), and Hommel (1988). The remaining five tests used the MLE of CACE (under the exclusion restriction) as the test statistic instead of the ITT test statistic. Table 6 displays proportions of simulations in which the null hypothesis was rejected, based on 1,000 replications.

Under the null hypotheses, all 10 familywise tests appear valid in terms of type I error. The randomization-based tests have the rejection rates closest to the nominal rejection rates. As expected, the Bonferroni-adjusted tests are conservative, especially when correlation among outcomes is high. In such settings, there are, in a sense, fewer possible effects to detect, and randomization-adjusted rejection rates are much higher relative to their Bonferroni-adjusted counterparts. The Holm, Hochberg, and Hommel procedures all perform similarly to Bonferroni under the null hypotheses.

Under alternative hypotheses, the CACE tests generally have higher power than the ITT tests. In addition, the randomization-based tests perform very well, especially when correlation among outcomes is high. The Bonferroni and Holm tests perform similarly, while the Hochberg and Hommel tests perform slightly

Table 6. Proportions of simulations in which the null hypothesis was rejected, under various data generation processes. Based on 1,000 replications.

Compliance Rate = 0.1	Rejection Rate at $\alpha = .05$									
	ITT					CACE				
Null is true	Bonferroni	Holm	Hochberg	Hommel	Rand-Based	Bonferroni	Holm	Hochberg	Hommel	Rand-Based
Zero correlation	0.035	0.035	0.037	0.039	0.050	0.033	0.033	0.033	0.034	0.039
Partial correlation	0.025	0.025	0.026	0.030	0.041	0.025	0.025	0.026	0.031	0.039
Perfect correlation	0.012	0.012	0.047	0.047	0.049	0.008	0.008	0.032	0.032	0.033
	ITT					CACE				
Alternative 1 is true	Bonferroni	Holm	Hochberg	Hommel	Rand-Based	Bonferroni	Holm	Hochberg	Hommel	Rand-Based
Zero correlation	0.161	0.161	0.167	0.169	0.189	0.209	0.209	0.217	0.226	0.228
Partial correlation	0.113	0.113	0.118	0.123	0.154	0.179	0.179	0.189	0.196	0.228
Perfect correlation	0.062	0.062	0.139	0.139	0.139	0.094	0.094	0.205	0.205	0.207
	ITT					CACE				
Alternative 2 is true	Bonferroni	Holm	Hochberg	Hommel	Rand-Based	Bonferroni	Holm	Hochberg	Hommel	Rand-Based
Zero correlation	0.303	0.303	0.306	0.310	0.342	0.357	0.357	0.366	0.369	0.380
Partial correlation	0.204	0.204	0.219	0.225	0.273	0.303	0.303	0.314	0.320	0.357
Perfect correlation	0.137	0.137	0.270	0.270	0.270	0.184	0.184	0.364	0.364	0.369
	ITT					CACE				
Alternative 3 is true	Bonferroni	Holm	Hochberg	Hommel	Rand-Based	Bonferroni	Holm	Hochberg	Hommel	Rand-Based
Zero correlation	0.688	0.688	0.702	0.716	0.754	0.710	0.710	0.746	0.756	0.742
Partial correlation	0.370	0.370	0.384	0.398	0.444	0.474	0.474	0.487	0.499	0.518
Perfect correlation	0.297	0.297	0.465	0.465	0.471	0.357	0.357	0.565	0.565	0.570

Compliance Rate = 0.3	Rejection Rate at $\alpha = .05$									
	ITT					CACE				
Null is true	Bonferroni	Holm	Hochberg	Hommel	Rand-Based	Bonferroni	Holm	Hochberg	Hommel	Rand-Based
Zero correlation	0.037	0.037	0.037	0.038	0.046	0.031	0.031	0.032	0.033	0.038
Partial correlation	0.033	0.033	0.035	0.036	0.044	0.025	0.025	0.025	0.026	0.035
Perfect correlation	0.010	0.010	0.032	0.032	0.035	0.007	0.007	0.038	0.038	0.039
	ITT					CACE				
Alternative 1 is true	Bonferroni	Holm	Hochberg	Hommel	Rand-Based	Bonferroni	Holm	Hochberg	Hommel	Rand-Based
Zero correlation	0.529	0.529	0.549	0.557	0.595	0.617	0.617	0.632	0.642	0.670
Partial correlation	0.482	0.482	0.511	0.532	0.571	0.604	0.604	0.625	0.638	0.671
Perfect correlation	0.309	0.309	0.492	0.492	0.497	0.420	0.420	0.606	0.606	0.611
	ITT					CACE				
Alternative 2 is true	Bonferroni	Holm	Hochberg	Hommel	Rand-Based	Bonferroni	Holm	Hochberg	Hommel	Rand-Based
Zero correlation	0.907	0.907	0.914	0.914	0.923	0.969	0.969	0.978	0.979	0.981
Partial correlation	0.859	0.859	0.868	0.872	0.891	0.946	0.946	0.960	0.963	0.968
Perfect correlation	0.752	0.752	0.862	0.862	0.865	0.871	0.871	0.957	0.957	0.957
	ITT					CACE				
Alternative 3 is true	Bonferroni	Holm	Hochberg	Hommel	Rand-Based	Bonferroni	Holm	Hochberg	Hommel	Rand-Based
Zero correlation	0.993	0.993	0.995	0.996	0.996	0.999	0.999	0.999	0.999	0.999
Partial correlation	0.984	0.984	0.986	0.987	0.990	0.997	0.997	0.998	0.999	0.999
Perfect correlation	0.966	0.966	0.993	0.993	0.993	0.989	0.989	0.999	0.999	0.999

better. The randomization-based procedure generally outperforms all four of the other procedures. In our simulations, CACE tests with randomization-based multiple comparisons adjustments have up to 3.3 times the power of traditional Bonferroni ITT tests, when treatment effects are difficult to detect. The relative power gain is less pronounced when treatment effects are larger, though gains are still apparent in the absolute scale. In a particular experimental setting, the magnitude of the power gain from the combined analysis method depends on the compliance rate, the magnitude of the treatment effect, the α level, and the correlation of the multiple test statistics.

5. The National Job Training Partnership Act Study

Title II of the United States Job Training Partnership Act (JTPA) of 1982

funded employment training programs for economically disadvantaged residents (Bloom et al. (1997); Abadie, Angrist and Imbens (2002)). To evaluate the effectiveness of those training programs, the National JTPA Study conducted a randomized experiment through 16 local administration areas involving a total of around 20,000 participants who applied for JTPA services from November 1987 to September 1989 (W.E. Upjohn Institute for Employment Research (2013)). Treatment group participants were eligible to receive JTPA services, while control group participants were ineligible to receive JTPA services for 18 months. Not every participant assigned to the treatment group actually enrolled and received JTPA services.

5.1. The data

Monthly employment outcomes were recorded for 30 months after assignment through follow-up surveys and administrative records from state unemployment insurance agencies. Researchers were interested in measuring JTPA effects across three time periods representing various stages of training and employment: months 1–6 (after assignment), the period when most JTPA enrollees were in the program; months 7–18, approximately the first post-program year; and months 19–30, approximately the second post-program year (Bloom et al. (1997)).

Bloom et al. (1997)'s original JTPA report evaluates effects on average income but does not explicitly address the large portion of zero-income participants. Although the report describes effects by subperiod as well as by various participant subgroups, it fails to mention or employ any multiple comparisons adjustments. Here we focus on JTPA's effects on employment status and use gender as our only background covariate; this facilitates standard, non-controversial modeling choices (see Section 5.2) and allows us to highlight our methodological contributions rather than discuss the sensitivity of our results to various, possibly complicated modeling decisions. Our methods can be extended to evaluate effects on other outcome variables, such as income and wages, provided that we outline a reasonable imputation model (Zhang, Rubin and Mealli (2009)).

We would like to evaluate whether JTPA had an effect on employment status for any of the three time periods. Because employment characteristics often differ by gender, we examined JTPA effects for the three time periods by gender, for a total of six gender-time groups. For illustrative purposes, we restricted our study population to adults who had obtained a high school or GED diploma (7,445, or 66.4%, of the 11,204 total adults in the original JTPA study) and assumed com-

plete randomization (with an approximate 2 : 1 treatment-to-control assignment ratio) of the participants, ignoring the local administration structure because of the limitations of the available data.

Of the 5,009 participants assigned to the treatment group, 3,316 (66.2%) subsequently received JTPA training. Although the study protocol barred participants assigned to the control group from receiving JTPA services for 18 months, 41 (1.7%) of 2,436 adults in the control group did in fact receive services within that time frame. To create a simpler setting with true one-sided non-compliance, we discarded these 41 participants (0.6% of the 7,445 total adults in our study) with the belief that their inclusion would have a negligible influence on the resulting inference.

Given two genders and three time periods, we have six complier-focused estimands in total, each one representing the difference in employment proportions within a particular gender-time group when receiving versus not receiving JTPA services. Two summaries of the observed data are provided in Figure 2 and Table 7. Figure 2 shows observed employment proportions across the six gender-time groups by observed compliance status. Within every group, observed compliers are employed at a higher rate than observed never-takers. Participants with unobserved compliance statuses are a mixture of compliers and never-takers, and tend to be employed at a rate in between the rates for observed compliers and observed never-takers.

Table 7 displays observed employment proportions across the gender-time groups according to both treatment assignment and treatment receipt, with the corresponding compliance compositions. We see that participants who received JTPA services, all of whom are compliers, tend to be employed at a higher rate than participants who were merely assigned to the treatment group (a mixture of compliers and never-takers), corroborating the findings in Figure 2 and suggesting that CACE statistics may lead to more significant estimated effects. In addition, we observe that participants who did not receive JTPA services — including any participants assigned to control as well as the never-takers assigned to JTPA — are employed at a lower rate than just the participants assigned to control. This inequality is intuitive because the observed never-takers are shown in Figure 2 to be employed at a lower rate than the assigned control group.

5.2. Imputation model for CACE

To test the null hypothesis of zero effects using the CACE statistic specified in Section 2, we need to specify an imputation model for the missing compliance

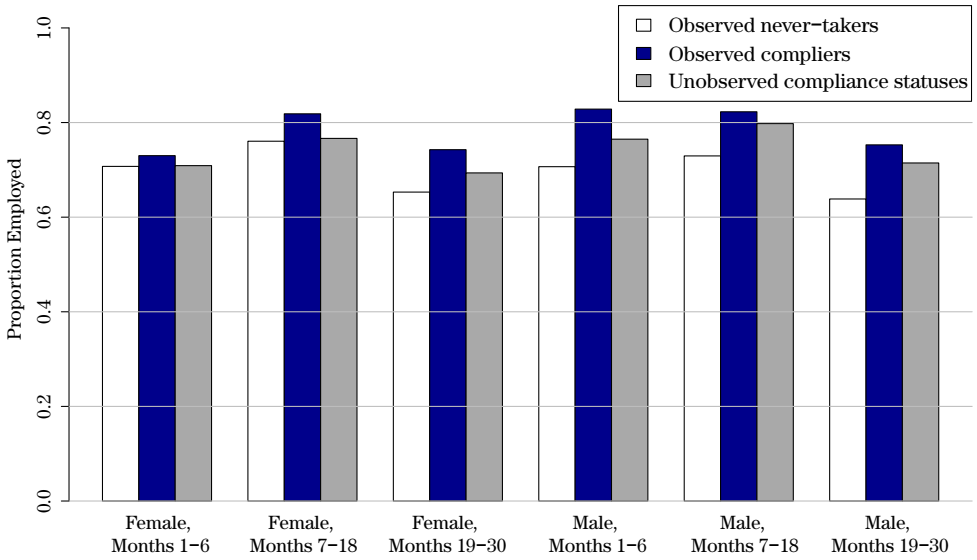

Figure 2. Observed employment proportions for JTPA participants by compliance status across the six gender-time groups.

Table 7. Observed employment proportions across the six gender-time groups according to both assignment to and receipt of JTPA services.

	Observed Employment Proportions	
	Assigned Control $C_i \in \{c, nt\}, Z_i = 0$	Assigned Treatment $C_i \in \{c, nt\}, Z_i = 1$
Female, Months 1–6	0.709	0.723
Female, Months 7–18	0.767	0.800
Female, Months 19–30	0.694	0.714
Male, Months 1–6	0.765	0.785
Male, Months 7–18	0.798	0.789
Male, Months 19–30	0.715	0.712
	Received Control $(C_i \in \{c, nt\}, Z_i = 0)$ or $(C_i = nt, Z_i = 1)$	Received Treatment $C_i = c, Z_i = 1$
Female, Months 1–6	0.708	0.730
Female, Months 7–18	0.764	0.818
Female, Months 19–30	0.677	0.743
Male, Months 1–6	0.740	0.828
Male, Months 7–18	0.769	0.823
Male, Months 19–30	0.683	0.753

statuses. Let X_i and $\boldsymbol{Y_i}$ denote the gender and the length-3 vector of employment outcomes (across the three time periods) of participant i. The three elements of $\boldsymbol{Y_i}$ are binary, so there are $2^3 = 8$ possible values of $\boldsymbol{Y_i}$; we model $\boldsymbol{Y}$ as a Multi-

nomial random variable with eight categories. Let ω_c be the super-population proportion of compliers, and let $\boldsymbol{\eta} = (\eta_{fc}, \eta_{fn}, \eta_{mc}, \eta_{mn})$ be the parameters that govern the outcome distributions of female compliers, female never-takers, male compliers, and male never-takers, respectively. Under the null hypothesis, these are the only four outcome distributions because we disregard treatment assignment. We placed a conjugate Beta(1,1) prior on ω_c and independent conjugate Dirichlet($\mathbf{1}$) priors on the four η parameters, where $\mathbf{1}$ is a length-8 vector of 1's.

Conditional on $\boldsymbol{\eta}$ and a participant's gender and compliance status, the natural outcome distribution under the null hypothesis is:

$$\boldsymbol{Y_i}^{\text{obs}} | X_i = x, C_i = q, \boldsymbol{\eta} \sim \text{Multinomial}(1, \eta_{xq}).$$

Note that we do not assume that the three employment outcomes are independent; this model is fully non-parametric for the joint distribution of the three outcomes. The posterior distributions of ω_c and $\boldsymbol{\eta}$ are informed by the outcomes of the participants with observed compliance statuses, i.e., those assigned to active treatment, and remain Beta and Dirichlet, respectively. For each gender x and compliance status q, write the Multinomial probability vector as $\eta_{xq} = (\pi_{xq1}, \ldots, \pi_{xq7}, 1 - \pi_{xq1} - \ldots - \pi_{xq7})$. Let

$$g_{xq}(\boldsymbol{y}; \eta_{xq}) = \pi_{xq1}^{I\{\boldsymbol{y}=(0,0,0)\}} \pi_{xq2}^{I\{\boldsymbol{y}=(0,0,1)\}} \ldots (1 - \pi_{xq1} - \ldots - \pi_{xq7})^{I\{\boldsymbol{y}=(1,1,1)\}}$$

denote the probability of outcome $\boldsymbol{y}$ for participants of gender x and compliance status q. Then, given a posterior draw of $(\omega_c, \boldsymbol{\eta})$, the missing compliance statuses were imputed probabilistically according to Bayes' rule:

$$P(C_i = c | \boldsymbol{Y_i^{\text{obs}}}, X_i{=}x, Z_i{=}0, \omega_c, \boldsymbol{\eta}) = \frac{\omega_c g_{xc}(\boldsymbol{Y_i^{\text{obs}}}; \eta_{xc})}{\omega_c g_{xc}(\boldsymbol{Y_i^{\text{obs}}}; \eta_{xc}) + (1{-}\omega_c) g_{xn}(\boldsymbol{Y_i^{\text{obs}}}; \eta_{xn})}. \tag{5.1}$$

5.3. Results and analysis

The observed values of the ITT and CACE statistics — i.e., the estimated effects of JTPA assignment and of receipt, respectively — are shown in the second column of Table 8. As we expect, the observed CACE values have larger magnitudes; the estimated ITT effects are diluted toward zero by the never-takers, who do not receive any treatment benefit. Because $\text{ITT} = \omega_c * \text{CACE} + (1 - \omega_c) * 0$, the estimated ITT effects are diluted by a proportion equal to one minus the compliance rate. Due to the random treatment assignment, we expect the overall compliance rate to be approximately equal to the compliance rate observed in the treatment group (66.2%).

Using randomization tests and the methods described in Section 4, we ob-

Table 8. Observed values, nominal p-values, and Bonferroni- and randomization-adjusted p-values for the six JTPA gender-time groups. Nominal p-values are obtained through randomization tests using 10,000 randomizations.

			Adjusted p-values	
ITT	Estimated Effect	Nominal p-value	Bonferroni	Randomization
Female, Months 1–6	0.014	0.351	1.000	0.895
Female, Months 7–18	0.020	0.199	1.000	0.685
Female, Months 19–30	0.033	0.014	0.085	0.077
Male, Months 1–6	−0.008	0.582	1.000	0.991
Male, Months 7–18	0.020	0.175	1.000	0.636
Male, Months 19–30	−0.003	0.874	1.000	1.000
			Adjusted p-values	
CACE	Estimated Effect	Nominal p-value	Bonferroni	Randomization
Female, Months 1–6	0.020	0.130	0.778	0.302
Female, Months 7–18	0.034	0.009	0.055	0.026
Female, Months 19–30	0.049	0.0002	0.001	0.001
Male, Months 1–6	−0.010	0.462	1.000	0.804
Male, Months 7–18	0.028	0.028	0.169	0.076
Male, Months 19–30	−0.001	0.967	1.000	1.000

tained one set of nominal ITT p-values and a second set of nominal CACE p-values, listed in the third column of Table 8. Each set contains six p-values, one for each gender-time group. We also applied Bonferroni and randomization adjustments to both sets of nominal p-values, resulting in four total sets of adjusted p-values, listed in the rightmost columns of Table 8.

The nominal ITT p-value for the "Female, Months 19–30" group indicates statistical significance at the $\alpha = .05$ level. However, after adjusting for multiple comparisons, neither the Bonferroni- nor randomization-adjusted ITT p-values for this group meets the .05 threshold. Across the six gender-time groups, the randomization-adjusted p-values tend to be smaller than their Bonferroni-adjusted counterparts; the adjusted p-values are tempered less when controlling the FWER via the statistics' joint randomization distribution because of the correlations among the six nominal p-values.

Overall, the CACE p-values are smaller — more sensitive to complier-only effects — than the ITT p-values. In particular, the CACE p-values for the "Female, Months 7–18" and "Female, Months 19–30" groups indicate a much greater level of significance for the estimated effects of JTPA on employment. Applying randomization-based instead of Bonferroni adjustments to the CACE p-values further increases the indicated significance of these estimated effects. The small randomization-adjusted CACE p-values for these groups suggest that either an

event has occurred that is *a priori* rare under the sharp null hypothesis of zero effects, or the sharp null hypothesis is not true — *receipt* of JTPA services did have an effect on the employment statuses of females with high school or GED diplomas in their first and second post-program years. The corresponding ITT p-values, although smallest among the six groups, are larger and do not have sufficient power to detect an effect on employment status for any of the gender-time groups.

This increase in power is general. We observed similar p-value trends when comparing our methods to ITT and Bonferroni analyses on JTPA data without the high school/GED diploma restriction as well as on other JTPA subgroups analyzed in Bloom et al. (1997).

6. Conclusion

We have detailed a randomization-based procedure for analyzing experimental data in the presence of both non-compliance and multiple testing that is more powerful than traditional ITT and Bonferroni analyses. As shown through simulations and analyses of the National JTPA Study data, a combined randomization-based procedure can be doubly advantageous, offering gains in power from both perspectives.

The ITT tests for the JTPA Study suggest that the training program had no real effects in increasing employment for either gender at any time point. The Bonferroni-adjusted CACE tests suggest that JTPA only increased employment for females in the long term (months 19–30). From a policy perspective, this initiative may be deemed too costly based on the time delay, as well as the fact that all five other subgroups had insignificant effects. Once we look at the randomization-adjusted CACE tests though, we conclude that JTPA actually had a positive effect on employment for females as soon as they finish the training program, and that the effect sustained into the longer term. Thus, it seems reasonable for policymakers to fund similar job training programs targeted for women.

Westfall and Young (1989) assumed Binomial data that facilitated closed-form calculations of nominal p-values, which were then adjusted using a permutation test. Here we propose fully randomization-based p-values ?- we exploit the randomization test to calculate *both* nominal and adjusted p-values. In addition, Westfall and Young (1989) described the adjusted p-values as "permutation-style," not explicitly motivated by the assignment mechanism in a randomized

experiment. In its exploration of non-compliance, Rubin (1998) required the randomization test to follow the randomized assignment mechanism actually used in the original experiment, an approach we advocate.

A number of other multiple comparisons procedures aim to address the false discovery rate (FDR) (Benjamini and Hochberg (1995)), rather than the FWER. These two error metrics are conceptually different; the choice of metric should be decided by the researcher depending on the field and specific research setting and goals. FDR is often preferred in settings with a large number of tests, such as genetic studies, in which finding one true genetic link may outweigh finding a few spurious links. In such cases, attempting to make exactly zero type I errors can be quite restrictive. On the other hand, FWER is often used in social science and pharmaceutical settings, in which governmental and regulatory agencies place the onus on the researcher to show that the treatment provides a beneficial effect. In these cases, the number of tests tends to be smaller, and type I errors can be extremely costly in terms of dollars to taxpayers and health risks to patients. For these reasons, we focused our discussion on the FWER.

Acknowledgment

The first author was supported by the U.S. Department of Defense through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. We also thank Don Rubin for the interesting and thought-provoking conversations we have had about this work.

Appendix

A. Marginal Distributions for Simulations

A.1. Non-compliance

For unit $i = 1, \ldots, N$, the control potential outcomes for compliers and never-takers have the marginal distributions

$$Y_i(0)|C_i = c \sim \text{Multinomial}(.45, .45, .10); \tag{A.1}$$

$$Y_i(0)|C_i = nt \sim \text{Multinomial}(.02, .02, .96). \tag{A.2}$$

Under the null hypothesis, $Y_i(1)$ has the same marginal distribution as $Y_i(0)$ regardless of compliance status. Under the alternative hypothesis, the complier treatment potential outcomes follow:

$$Y_i(1)|C_i = c \sim \text{Multinomial}(.80, .10, .10), \tag{A.3}$$

while the never-taker treatment potential outcomes follow (A.2).

A.2. Multiple testing

For unit $i = 1, \ldots, N$ and outcome $j = 1, 2, 3$, the marginal distributions of the control potential outcomes are

$$Y_{ij}(0) \sim \text{Multinomial}(.45, .45, .10). \tag{A.4}$$

Under the null hypothesis, $Y_{ij}(1)$ has the same marginal distribution as $Y_{ij}(0)$. Under the alternative hypotheses, the marginal distributions of the treatment potential outcomes are

$$Y_{ij}(1) \sim \text{Multinomial}(.50, .45, .05). \tag{A.5}$$

A.3. Non-compliance and multiple testing

Under the null hypothesis, the potential outcomes follow the marginal distributions at (A.1) and (A.2) in Appendix A.1. $Y_i(1)$ has the same marginal distribution as $Y_i(0)$ regardless of compliance status.

Under alternative hypothesis 1, the complier potential outcomes marginally follow

$$\begin{aligned} Y_i(0)|C_i = c &\sim \text{Multinomial}(.45, .45, .10); \\ Y_i(1)|C_i = c &\sim \text{Multinomial}(.80, .10, .10). \end{aligned} \tag{A.6}$$

Under alternative hypothesis 2, the complier potential outcomes marginally follow

$$\begin{aligned} Y_i(0)|C_i = c &\sim \text{Multinomial}(.30, .60, .10); \\ Y_i(1)|C_i = c &\sim \text{Multinomial}(.80, .10, .10). \end{aligned} \tag{A.7}$$

Under alternative hypothesis 3, the complier potential outcomes marginally follow

$$\begin{aligned} Y_i(0)|C_i = c &\sim \text{Multinomial}(.25, .55, .20); \\ Y_i(1)|C_i = c &\sim \text{Multinomial}(.80, .10, .10). \end{aligned} \tag{A.8}$$

B. Correlation Structure Generation

To simulate correlation structures among multiple outcomes, we used the following processes utilizing the marginal distributions described in Appendix A. For units $i = 1, \ldots, N$ and treatment assignment $z = 0, 1$,

- Zero correlation: all $Y_{ij}(z)$ $(j = 1, 2, 3)$ were drawn independently according

to their marginal distributions.

- Partial correlation: $Y_{i1}(z)$ was drawn according to its marginal distribution. With probability 1/2, $Y_{i2}(z)$ was set equal to the drawn value of $Y_{i1}(z)$; otherwise, $Y_{i2}(z)$ was drawn independently according to its marginal distribution. $Y_{i3}(z)$ was set equal to $Y_{i1}(z)$ with probability 1/3, set equal to $Y_{i2}(z)$ with probability 1/3, or drawn independently according to its marginal distribution.
- Perfect correlation: $Y_{i1}(z)$ was drawn according to its marginal distribution. Then both $Y_{i2}(z)$ and $Y_{i3}(z)$ were set equal to the drawn value of $Y_{i1}(z)$.

References

Abadie, A., Angrist, J. and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*. **70**, 91–117.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*. **91**, 444–455.

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

Bloom, H. S., Orr, L. L., Bell, S. H., Cave, G., Doolittle, F., Lin, W. and Bos, J. M. (1997). The benefits and costs of JTPA Title II-A programs: Key findings from the National Job Training Partnership Act Study. *Journal of Human Resources*. **32**, 549–576.

Brown, C. C. and Fears, T. R. (1981). Exact significance levels for multiple binomial testing with application to carcinogenicity screens. *Biometrics*. **37**, 763–774.

Cabin, R. J. and Mitchell, R. J. (2000). To Bonferroni or not to Bonferroni: when and how are the questions. *Bulletin of the Ecological Society of America*. **81**, 246–248.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. 1st ed. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935), *The Design of Experiments*. Oxford: Oliver & Boyd.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*. **58**, 21–29.

Good, P. I. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses*, vol. 3. Springer.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. **75**, 800–802.

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

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*. **75**, 383–386.

Imbens, G. W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*. **25**, 305–327.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*. **22**, 1142–1160.

Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology*. **15**, 1044–1045.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*. **5**, 465–472, translated by Dabrowska, DM and Speed, TP (1990).

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal*. **316**, 1236–1238.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*. **1**, 43–46.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. **66**, 688–701.

Rubin, D. B. (1980). Comment on randomization analysis of experimental data: the Fisher randomization test. *Journal of the American Statistical Association*. **75**, 591–593.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*. **81**, 961–962.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc.

Rubin, D. B. (1998). More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in Medicine*. **17**, 371–385.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*. **100**, 322–331.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*. **82**, 528–540.

W.E. Upjohn Institute for Employment Research (2013). The National JTPA Study. Public Use Data and Data Summary.

Westfall, P. H. and Young, S. S. (1989). p value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association*. **84**, 780–786.

Zhang, J. L., Rubin, D. B. and Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*. **104**, 166–176.

Department of Statistics, Harvard University, 1 Oxford St., Cambridge, MA 02138, USA

E-mail: joseph.j.lee@post.harvard.edu

Department of Statistics, University of Florence, Viale Morgagni, 59, 50134 Florence, Italy

E-mail: l.forastiere@disia.unifi.it

Harvard Graduate School of Education, 13 Appian Way, Cambridge, MA 02138, USA

E-mail: luke_miratrix@gse.harvard.edu

Department of Statistics, Harvard University, 1 Oxford St., Cambridge, MA 02138, USA

E-mail: pillai@fas.harvard.edu

(Received March 2016; accepted June 2016)