Statistical Power in Evaluations That Investigate Effects on Multiple Outcomes

A Guide for Researchers

Kristin E. Porter

July 2016

mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

Acknowledgments

This working paper was supported by Grant R305D140024 to MDRC from the Institute of Education Sciences, U.S. Department of Education. The author thanks Sara Staszak, Charles Michalopoulos, Howard Bloom, and Richard Hendra for their assistance and review.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Annie E. Casey Foundation, Charles and Lynn Schusterman Family Foundation, The Edna McConnell Clark Foundation, Ford Foundation, The George Gund Foundation, Daniel and Corinne Goldman, The Harry and Jeanette Weinberg Foundation, Inc., The JBP Foundation, The Joyce Foundation, The Kresge Foundation, Laura and John Arnold Foundation, Sandler Foundation, and The Starr Foundation.

In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, Sandler Foundation, and The Stupski Family Fund, as well as other individual contributors.

The findings and conclusions in this report do not necessarily represent the official positions or policies of the funders.

For information about MDRC and copies of our publications, see our website: www.mdrc.org.

Abstract

In education research and in many other fields, researchers are often interested in testing the effectiveness of an intervention on multiple outcomes, for multiple subgroups, at multiple points in time, or across multiple treatment groups. The resulting multiplicity of statistical hypothesis tests can lead to spurious findings of effects. Multiple testing procedures (MTPs) are statistical procedures that counteract this problem by adjusting p-values for effect estimates upward. While MTPs are increasingly used in impact evaluations in education and other areas, an important consequence of their use is a change in statistical power that can be substantial. Unfortunately, researchers frequently ignore the power implications of MTPs when designing studies. Consequently, in some cases, sample sizes may be too small, and studies may be underpowered to detect effects as small as a desired size. In other cases, sample sizes may be larger than needed, or studies may be powered to detect smaller effects than anticipated.

Researchers typically worry that moving from one to multiple hypothesis tests and thus employing MTPs results in a *loss* of power. However, that need not always be the case. Power is indeed lost if one focuses on *individual* power: the probability of detecting an effect of a particular size or larger for each particular hypothesis test, given that the effect truly exists. However, in studies with multiplicity, alternative definitions of power exist that in some cases may be more appropriate. For example, when testing for effects on multiple outcomes, one might consider 1-minimal power: the probability of detecting effects of at least a particular size on at least one outcome. Similarly, one might consider ½-minimal power: the probability of detecting effects of at least a particular size on at least ½ of the outcomes. Also, one might consider complete power: the power to detect effects of at least a particular size on all outcomes. The choice of definition of power depends on the objectives of the study and on how the success of the intervention is defined. The choice of definition also affects the overall extent of power.

This paper presents methods for estimating statistical power, for multiple definitions of statistical power, when applying any of five common MTPs — Bonferroni, Holm, single-step and step-down versions of Westfall-Young, and Benjamini-Hochberg. The paper also presents empirical findings on how power is affected by the use of MTPs. To contain its scope, the paper focuses on multiplicity that results from estimating effects on multiple outcomes. The paper also focuses on the simplest research design and analysis plan that education studies typically use in practice: a multisite, randomized controlled trial (RCT) with the blocked randomization of individuals, in which effects are estimated using a model with block-specific intercepts and with the assumption of constant effects across blocks. However, the power estimation methods presented can easily be extended to other modeling assumptions and other study designs.

Contents

Appendix

List of Exhibits

Table

Figure

Figure

1. Introduction

In education research and in many other fields, researchers are often interested in testing the effectiveness of an intervention on multiple outcomes, for multiple subgroups, at multiple points in time, or across multiple treatment groups. The resulting multiplicity of statistical hypothesis tests can lead to spurious findings of effects. Multiple testing procedures (MTPs) are statistical procedures that counteract this problem by adjusting p-values for effect estimates upward. When not using an MTP, the probability of false positive findings increases, sometimes dramatically, with the number of tests. When using an MTP, this probability is reduced.

MTPs are increasingly used in impact evaluations in education. For example, the Institute for Education Sciences (IES), the primary research arm of the U.S. Department of Education, published a technical methods report on multiple testing that recommends MTPs as one of several strategies for dealing with the multiplicity problem (Schochet, 2008). In addition, IES's What Works Clearinghouse, which reviews and summarizes thousands of education studies, applies a particular MTP, the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) to studies' statistically significant findings when effects are estimated for multiple measures or groups (U.S. Department of Education, 2014).

However, an important consequence of MTPs is a change in statistical power that can be substantial. That is, the use of MTPs changes the probability of detecting effects when they truly exist, compared with the situation when the multiplicity problem is ignored. Unfortunately, while researchers are increasingly using MTPs, they frequently ignore the power implications of their use when designing studies. Consequently, in some cases sample sizes may be too small, and studies may be underpowered to detect effects as small as a desired size. In other cases, sample sizes may be larger than needed, or studies may be powered to detect smaller effects than anticipated.

Researchers typically worry that moving from one to multiple hypothesis tests and thus employing MTPs results in a *loss* of power. However, that need not always be the case. Power is indeed lost if one focuses on *individual* power — the probability of detecting an effect of a particular size or larger for each particular hypothesis test, given that the effect truly exists. However, in studies with multiplicity, alternative definitions of power exist and in some cases may be more appropriate (e.g., see Westfall, Tobias, & Wolfinger, 2011; Dudoit, Shaffer, & Bodrick, 2003; Chen, Luo, Liu, & Mehrotra, 2011; and Senn & Bretz, 2007). For example, when testing for effects on multiple outcomes, one might consider 1-minimal power: the probability of detecting effects of at least a particular size (which can vary by outcome) on at least one outcome. Similarly, one might consider ½-minimal power: the probability of detecting effects of at least a particular size on at least ½ of the outcomes. Also, one might consider complete power: the power to detect effects of at least a particular size on all outcomes. The choice of definition of power depends on the objectives of the study and on how the success of the intervention is defined. The choice of definition also affects the overall extent of power.

This paper fills an important gap in the existing literature on designing impact studies in education and social policy. The literature and tools on statistical power are extensive but do not take multiplicity into account (e.g., Dong & Maynard, 2013; Spybrook et al., 2011; Raudenbush et al., 2011; Hedges & Rhoads, 2010). Also, the literature on the multiple testing problem in these fields does not provide clear guidance for estimating power, nor does it explore power under alternative definitions (which do exist in literature in medicine and genomics).

This paper presents methods for estimating statistical power, for multiple definitions of statistical power, when applying any of five common MTPs — Bonferroni, Holm, single-step and step-down versions of Westfall-Young, and Benjamini-Hochberg. It also provides R code so that researchers can implement the power estimation methods in their studies. The paper also presents empirical findings on how power is affected by the use of MTPs. The extent to which studies are underpowered or overpowered varies with circumstances particular to those studies, including: the definition of power, the number of tests, the proportion of tests that are truly null, the correlation between tests, the $R^{2\prime}$s of baseline covariates, and the particular MTP used to adjust p-values. The paper explores all of these factors and discusses the implications for practice.

To contain the scope of the paper, it focuses on multiplicity that results from estimating effects on multiple outcomes.[1] The paper also focuses on the simplest research design and analysis plan that education studies typically use in practice: a multisite, randomized controlled trial (RCT) with the blocked randomization of individuals, in which effects are estimated using a model with block-specific intercepts and with the assumption of constant effects across blocks. However, as will be discussed at the end of the paper, the power estimation methods presented can easily be extended to other modeling assumptions and other study designs.

The remainder of the paper proceeds as follows: Section 2 provides an overview of multiple testing. It provides some intuition of the multiple testing problem, summarizes how MTPs address the multiple testing problem, and discusses features of the MTPs in this paper that affect power. Section 3 then gives a brief overview of a methodological approach for

[1] We note that there are different guidelines for *when* to adjust for multiple outcomes in education studies. For example, Schochet (2008) recommends organizing primary outcomes into domains, conducting tests on composite domain outcomes, and applying multiplicity corrections to composites across domains. The What Works Clearinghouse applies multiplicity corrections to findings within the same domain rather than across different domains. This paper would apply to either case. In this paper, the word "outcome" refers to either a single outcome or an outcome domain, and the paper focuses on any situation in which an analyst would apply adjustments to account for multiple outcomes.

estimating power and provides an example of how researchers can carry out power estimation under multiplicity. Section 4 presents empirical findings for a variety of realistic scenarios. Finally, Section 5 provides a summary of the empirical findings and recommendations for practice and next steps. A detailed description of the MTPs in this paper can be found in Appendix A. R code implementing the power estimation methodology can be found in Appendix B. Also, power comparisons with other sources that validate the accuracy of the power estimation methodology can be found in Appendix C.

2. Overview of Multiple Testing

2.1 The Multiple Testing Problem

This paper focuses on the frequentist framework of hypothesis testing, as it is currently the prevailing framework in education and social policy research. Under this framework, the treatment and control groups in an RCT are considered random samples from a defined population (assumed to be the same across all blocks under the assumed design). Following the Rubin-Neyman counterfactual framework (Neyman, 1923; Rubin, 1974, 2006), $Y0_i(m)$ is the $m^{th}$of M outcomes for individual i when not exposed to the treatment, and $Y1_i(m)$ is the $m^{th}$of M outcomes for individual i when exposed to treatment. Then the population average treatment on the m^{th} outcome, given by

$$\psi(m) = E(Y1_i(m)) - E(Y0_i(m)), \tag{1}$$

is considered to be fixed. Researchers often express the average treatment effect in standard deviation units — as an effect size. The effect size parameter for the m^{th} outcome is given by

$$ES(m) = \frac{\psi(m)}{\sigma(m)}, \tag{2}$$

where $\sigma(m)$ is the standard deviation of the m^{th} outcome.[2]

In the frequentist framework, one typically tests a null hypothesis of no effect, $H0(m){:}\ ES(m) = 0$, against an alternative hypothesis of $H1(m){:}\ ES(m) \neq 0$ for a two-sided test or $H1(m){:}\ ES(m) > 0$ or $H1(m){:}\ ES(m) < 0$ for a one-sided test. However, for the purposes of computing power, as discussed below, researchers must specify an alternative hypothesis of at least a particular effect size — that is, a minimum detectable effect size

[2] It is assumed here that the standard deviation is the same in both counterfactual settings.

(MDES).[3] A significance test, such as a two-sided or one-sided *t*-test, is then conducted, and one obtains a test statistic given by

$$t(m) = \frac{\widehat{ES}(m)}{\widehat{SE}\left(\widehat{ES}(m)\right)}, \tag{3}$$

from which a raw p-value is computed. Here, the term "raw" is used to distinguish this p-value from a p-value that has been adjusted for multiple hypothesis tests, as discussed below. The raw p-value is the probability of a test statistic being at least as extreme as the one observed, given that the null hypothesis is true. For a two-sided test, which is the focus of this paper going forward, the raw p-value for the m^{th} test is $p(m) = 2 * \Pr\{T(m) \geq |t(m)|\}$.[4] This expression means we use our knowledge of the sampling distribution of the *t*-statistic, and we identify where our observed test statistic falls in that distribution when it is centered around zero.

When testing a *single* hypothesis under this framework (such that $M = 1$), researchers typically specify an acceptable maximum probability of making a Type I error, α. A Type I error is the probability of erroneously rejecting the null hypothesis when it is true. The quantity α is also referred to as the significance level. If $\alpha = 0.05$, then the null hypothesis is rejected if the p-value is less than 0.05, and it is concluded that the intervention had an effect because there is less than a 5% chance that this finding is a false positive.

When one tests *multiple* hypotheses under this framework (such that $M > 1$) and one conducts a separate test for each of the hypotheses with $\alpha = 0.05$, there is a *greater* than 5% chance of a false positive finding in the study. If the multiple tests are independent, the probability that at least one of the M null hypothesis tests will be erroneously rejected is $1 - Pr$ (none of the null hypotheses will be erroneously rejected) $= 1 - (1 - \alpha)^M$. Therefore, if researchers estimate effects on two independent outcomes, the probability of at least one false positive finding is almost 10%. If researchers estimate effects on five independent outcomes, the probability of a false positive finding is 23%. This Type I error inflation for independent outcomes demonstrates the crux of the multiple testing problem. In practice, however, the multiple outcomes are at least somewhat correlated, which makes the test statistics correlated and reduces the extent of Type I error inflation. Nonetheless, any error inflation can still make it problematic to draw reliable conclusions about the existence of effects. As introduced above, to

[3] An MDES is defined as the smallest true effect size that a study can detect with statistical significance. For a discussion of minimum detectable effects (MDEs), which are expressed in outcomes' units, and MDESs, which are expressed in standard deviation units, see, for example, Bloom (1995), Schochet (2005), and Bloom (2006).

[4] For a one-sided test, depending on the direction of our alternative hypothesis, the raw p-value for the m^{th} test is computed as $p(m) = \Pr\{T(m) \geq t(m)\}$ or $p(m) = \Pr\{T(m) \leq t(m)\}$.

counteract the multiple testing problem, MTPs adjust p-values upward.[5] The sections that follow will describe how the MTPs do so.

Recall that the power of an individual hypothesis test is the probability of rejecting a false null hypothesis of at least a specified size. If raw p-values are adjusted upward, one is less likely to reject the null hypotheses that are true (meaning there is truly no effect of at least a specified size), which reduces the probability of Type I errors, or false positive findings. Reducing this probability is the goal of MTPs. But if raw p-values are adjusted upward, one is also less likely to reject the null hypotheses that are false (meaning there truly is an effect of at least a specified size). Therefore, all MTPs reduce *individual* power (the power of separate hypothesis tests for each outcome) compared with the situation when no multiplicity adjustments are made or the situation when there is only one hypothesis test.

MTPs also reduce all other definitions of power compared with the situation when no multiplicity adjustments are made — but not necessarily compared with the situation when there is only one hypothesis test. For example, 1-minimal power, the probability of detecting effects (of at least a specified size) on *at least one* outcome — after adjusting for multiplicity — is typically greater than the probability of detecting an effect of the same size on a single outcome. This increase may or may not occur with other definitions of power (e.g., the probability of detecting a third, half, or all false null hypotheses), which will be investigated and discussed in Section 4.

2.2 Using MTPs to Protect Against Spurious Impact Findings

The MTPs that are the focus of this paper fall into two different classes. The first class reframes Type I error as a rate across the entire set or "family" of multiple hypothesis tests. This rate is called the familywise error rate (FWER; Tukey, 1953). It is typically set to the same value as the probability of a Type I error for a single test, or to α. MTPs that control the FWER at 5% adjust p-values in a way that ensures that the probability of at least one Type I error across the entire set of hypothesis tests is no more than 5%. The MTPs introduced by Bonferroni (Dunn, 1959, 1961), Holm (1979), and Westfall and Young (1993) control the FWER.

The second class of MTPs takes an entirely different approach to the multiple testing problem. MTPs in this class control the false discovery rate (FDR). Introduced by Benjamini and Hochberg (1995), the FDR is the expected proportion of all rejected hypotheses that are erroneously rejected.

[5] Alternatively, MTPs can decrease the critical values for rejecting hypothesis tests. For ease of presentation, this paper focuses only on the approach of increasing p-values.

Table 1

Numbers of Hypothesis Types and Decisions

	Observed Decisions		
Unobserved Truths	Number not rejected	Number rejected	Total
Number of true null hypotheses	*A*	*B*	*M0*
Number of false null hypotheses	*C*	*D*	*M1*
Total	*M-R*	*R*	*M*

The two-by-two representation in Table 1 is often found in articles on multiple hypothesis testing. It helps to illustrate the difference between FWER and FDR. Let M be the total number of tests. Therefore, we have M unobserved truths: whether or not the null hypotheses are true or false. We also have M observed decisions: whether or not the null hypotheses were rejected, because the p-values were less than α. In Table 1, A, B, C, and D are four possible scenarios: the numbers of true or false hypotheses not rejected or rejected. $M0$ and $M1$ are the unobservable numbers of true null and false null hypotheses. R is the number of null hypotheses that were rejected, and $M - R$ is the number of null hypotheses that were not rejected.

In Table 1, B is the number of erroneously rejected null hypotheses, or the number of false positive findings. Therefore, the FWER is equivalent to $\Pr(B > 0)$, the probability of at least one false positive finding. Recall the examples above about Type I error inflation when testing for effects on independent outcomes in the case that α is set to 0.05 and no MTPs are applied. The Type I error was almost 10% when testing effects on two independent outcomes and 23% when testing effects on five independent outcomes. These Type I error rates both correspond to the FWER. The goal of MTPs that control the FWER is to bring these percentages back down to 5%.

Also in Table 1, the FDR is equal to $E(\frac{B}{R})$ but is defined to be 0 when $R = 0$, or when no hypotheses are rejected. As is frequently noted in the literature (e.g., Shaffer, 1995; Schochet, 2008), the FWER and FDR have different objectives. Control of the FWER protects researchers from *any* spurious findings and so may be preferred when even a single false positive could lead to the wrong conclusion about the effectiveness of an intervention. On the other hand, the FDR is more lenient with false positives. Researchers may be willing to accept a few false positives, B, when the total number of rejected hypotheses, R, is large. Note that under the complete null hypothesis that all M null hypotheses are null, the FDR is equal to the FWER, because when referring back to Table 1 we have $FWER = P(R > 0) = E\left(\frac{B}{R}\right) = FDR$. However, if any effects truly exist, then $FWER \geq FDR$. As a result, in the case where there is at least one false null hypothesis (at least one true effect at least as large as a specified MDES), an MTP that controls the FDR at 5% will have a Type I error rate that is greater than 5%.

Note that MTPs may provide either weak or strong control of the error rate they target. An MTP provides weak control of the FWER or the FDR at level α if the control can only be guaranteed when all nulls are true, or when the effects on all outcomes are zero. An MTP provides strong control of the FWER or FDR at level α if the control is guaranteed when some null hypotheses are true and some are false, or when there may be effects on at least some outcomes. Of course, strong control is preferred.[6]

2.3 Common MTPs in Education Research and Their Impact on Power

The five MTPs included in this paper were chosen because they are common in research in education and other social policy areas. An intuitive overview of each procedure, expressions defining the calculations involved, and references for more details, including proofs of the MTPs' properties, can be found in Appendix A. The goal of the discussion here is to briefly summarize the features of the MTPs that affect statistical power.

The first feature of an MTP that affects its statistical power is whether it controls the FWER or the FDR. Recall that the Bonferroni, the Holm, and both Westfall-Young MTPs control the FWER, while the Benjamini-Hochberg MTP controls the FDR. MTPs that control the FDR adjust p-values upward less than MTPs that control the FWER. Consequently, MTPs that control the FDR will typically have more power than FTPs that control the FWER. However, as discussed earlier, a disadvantage of MTPs that control the FDR is that they are more lenient with false positives than MTPs that control the FWER.

A second feature of an MTP that affects its statistical power is whether it is "single-step" or "stepwise." Single-step procedures adjust each p-value independently of the other p-values. For example, the Bonferroni MTP multiplies all raw p-values by M. Therefore, one p-value adjustment does not depend on other p-value adjustments, only on the number of tests. In contrast, stepwise procedures first order raw p-values (or test statistics), and then adjust according to the order of the tests. The adjustments depend on null hypotheses already rejected

[6] It is beyond the scope of this paper to provide technical details as to how the MTPs achieve strong or weak control, but proofs of these properties can be found in, for example, Ewens and Grant (2005) and Benjamini and Hochberg (1995).

in previous steps. For example, the Holm MTP — the stepwise counterpart to the Bonferroni MTP — orders raw p-values from smallest to largest. The procedure then multiplies the smallest p-value by M, the second smallest p-value by M-1, and so on, but also enforces that each adjusted p-value is greater than or equal to the previous adjusted p-value and that it is not greater than one. (For more details, see Appendix A.) Overall, stepwise MTPs allow for less adjustment than single-step MTPs in later steps, and therefore preserve more power. The Bonferroni and one of the Westfall-Young MTPs are single-step; the Holm and Benjamini-Hochberg MTPs and the other Westfall-Young MTP are stepwise. Note that stepwise procedures may be "step-up" or "step-down." Examples of both are included in the five MTPs studied in this paper, as described in Appendix A.

In the discussion that follows, the following shorthand is employed, which includes information on whether the MTPs are single-step or stepwise: BF-SS for Bonferroni (SS = single-step), HO-SD for Holm (SD = step-down), WY-SS and WY-SD for Westfall-Young single-step and step-down, and BH-SU for Benjamini-Hochberg (SU = step-up).

Finally, a third feature of an MTP that affects its statistical power is whether or not it takes into account the correlation of test statistics. The Bonferroni and Holm procedures strongly control the FWER when the multiple tests' statistics are correlated, but they adjust p-values more than is necessary in that case. The truth of this assertion can be seen if one considers the scenario in which all tests are perfectly correlated. Then one would not need to adjust p-values in order to control the FWER (because there would be essentially just one outcome), yet the p-values would be increased substantially, to an extent depending on M. Along with the Bonferroni and Holm MTPs, the Benjamin-Hochberg MTP also does not take correlations into account.[7]

In contrast, both of the Westfall-Young MTPs rely on the estimation of the joint distribution of test statistics when the "complete null hypothesis" (that there are not effects on any of the outcomes) is true. This joint distribution of the test statistics is estimated from the study's data. For example, permutations of the treatment indicator can be used to estimate impacts when the association between treatment status and the outcome is broken. Random permutations of the research group assignments are conducted a large number of times, resulting in a distribution of test statistics under the complete null. Because the actual data are used to generate this null distribution, correlations among the test statistics are captured. Then observed test statistics can be compared with the distribution of test statistics under the complete null

[7] The Benjamini-Hochberg procedure was originally shown to control the FDR for independent test statistics. However, Benjamini and Yekutieli (2001) showed that it also controls the FDR for true null hypotheses with "positive regression dependence." This condition is satisfied for most applications in practice.

Table 2

Summary of Features of MTPs

	Controls FWER or FDR	Single-Step or Stepwise	Accounts for Correlation Between Tests
Bonferroni (BF-SS)	FWER	Single-step	No
Holm (HO-SD)	FWER	Stepwise	No
Westfall-Young (WY-SS)	FWER	Single-step	Yes
Westfall-Young (WY-SD)	FWER	Stepwise	Yes
Benjamini-Hochberg (BH-SU)	FDR	Stepwise	No

hypothesis.[8] Again, for more details, see Appendix A. The main point is that by taking the correlations into account, one can make p-value adjustments that are not overly conservative, and thus better preserve power.

Table 2 summarizes the essential features of the MTPs. Empirical findings on how much these factors affect each definition of power are presented in Section 4.

3. Estimating Power in Studies of Impacts on Multiple Outcomes

This section of the paper summarizes a methodological approach for estimating power when investigating impacts on multiple outcomes and when using one of the MTPs presented above. It then provides an illustrative example of how researchers can use the estimation approach to guide the design of a study. It describes how to think about some of the needed assumptions, some of which are different from those needed to estimate the power of studies focused on a single outcome.

As noted above, the power estimation methodology described here focuses on studies in which multiplicity is due to having multiple outcomes. It also focuses on studies in which one is using the simplest research design and analysis plan that education studies typically use in practice: a randomized trial with the blocked randomization of individuals, in which effects are estimated using a model that has block-specific intercepts and that assumes constant effects across blocks.

[8] Instead of using test statistics, the Westfall-Young MTPs can alternatively compare raw p-values with the estimated joint null distribution of p-values.

3.1 Overview of Power Estimation Methods

For this RCT design and these assumptions of focus, the model for estimating impacts on the m^{th} of M outcomes is given by:

$$Y_i(m) = \psi(m)T_i + \sum_{j=1}^{J} \theta_j Block_{j_i} + \sum_{k=1}^{K(m)} \gamma_k(m) C_{k_i}(m) + r_i(m), \tag{4}$$

where, for individual i, $Y_i(m)$ is the m^{th} outcome; T_i is the treatment indicator; $Block_{j_i}$ is an indicator of whether individual i belongs to the j^{th} block; $C_{k_i}(m)$ is the k^{th} individual-level covariate; and $r_i(m)$ is the residual, normally distributed with mean zero and variance $\epsilon^2(m)$.[9] The coefficient $\psi(m)$ is the treatment effect on the m^{th} outcome, as defined in (1) using the counterfactual framework.

In this model, the standard error of the treatment effect estimate, $\hat{\psi}(m)$ is given by

$$SE\big(\hat{\psi}(m)\big) = \sqrt{\frac{\sigma^2(m)(1 - R^2(m))}{\bar{T}(1 - \bar{T})Jn_j}}, \tag{5}$$

where $\sigma^2(m)$ is the pooled outcome variance of the m^{th} outcome;[10] $R^2(m)$ is the proportion of the variance in the m^{th} outcome that is explained by the baseline covariates (including the block indicators); $\bar{T}$ is the proportion of the sample within each block that is assigned to the treatment group; J is the number of blocks and n_j is the number of individuals within each block (Bloom, 2006).

When expressing the estimated treatment effect as an effect size, as defined in the previous section, the standard error of the effect size estimate is given by

$$\begin{aligned} SE\left(\widehat{ES}(m)\right) &= SE(\frac{\hat{\psi}(m)}{\sigma(m)}) \\ &= \sqrt{\frac{1 - R^2(m)}{\bar{T}(1 - \bar{T})Jn_j}}. \end{aligned} \tag{6}$$

[9] The assumption of normally distributed residuals is not needed to estimate impacts or implement the MTPs.

[10] Here it is assumed that the variance of the outcome is the same in both the treatment and control groups.

For convenience, let $Q(m) \equiv SE\left(\widehat{ES}(m)\right)$. To estimate $Q(m)$, known values are inserted for $\bar{T}, J$, and n_j, and all other parameters in (6) are replaced by sample estimates. Then, when testing the m^{th} null hypothesis, $ES(m) = 0$, the test statistic for a t-test is given by

$$t(m) = \frac{\widehat{ES}(m)}{\hat{Q}(m)}. \tag{7}$$

When the null is true, $t(m)$ has a t-distribution with mean zero and degrees of freedom df. For our assumed model in (4), $df(m) = Jn_j - g^*(m) - 1$, where $g^*(m)$ is the total number of baseline covariates included in the model for the m^{th} outcome, including the block indicators such that $g^*(m) = K(m) + J$.

As mentioned above, in evaluations, researchers typically design studies so that they will have sufficient statistical power to detect, with a p-value less than α, at least the smallest effect that would be meaningful for the program under study. This is the MDES when focusing on standard deviation units, as is the case here. If the m^{th} hypothesis is false such that $|ES(m)|$ is greater than or equal to a *specific* MDES, then $t(m)$ has a t-distribution with mean $MDES(m)/Q(m)$, and again degrees of freedom df.

When $M > 1$, one can define a set of M null hypotheses and M alternative hypotheses. The set of null hypotheses is $ES(m) = 0$ for all m. This set defines the complete null hypothesis (referred to as $\boldsymbol{H0}$) that there are not effects on any of the outcomes. The set of two-sided alternative hypotheses focused on minimum detectable effects (referred to as $\boldsymbol{H1}$), is $|ES(m)| \geq MDES(m)$ for $m = 1, \ldots, M$, where the MDES may vary for each outcome.

Under the complete null hypothesis, $\boldsymbol{H0}$, the set of test statistics for all M hypothesis tests, which can be written collectively as $\boldsymbol{t0}$, have a multivariate t-distribution with means of zero, degrees of freedom equal to the vector $\boldsymbol{df}$, and correlation matrix $\boldsymbol{\rho}$. Under the set of specific alternative hypotheses, $\boldsymbol{H1}$, the set of test statistics, which can be written collectively as $\boldsymbol{t1}$, have the same multivariate t-distribution — except that the means are equal to the vector $\boldsymbol{MDES/Q}$.

Thus, the following are the essential insights for estimating power when adjusting for multiple hypothesis tests due to estimating effects on multiple outcomes:

1. When one assumes a correlational structure for the test statistics, the *joint null distribution* of the test statistics for the M tests is known.
2. When one specifies an MDES for each outcome and when one can identify $Q(m) \equiv SE(\hat{\psi}(m))$ for each outcome, as we have above, the *joint alternative distribution* of the test statistics for the M tests is also known.

3. Therefore, the test statistics $\boldsymbol{t0}$ and $\boldsymbol{t1}$ can be generated (i.e., simulated) with statistical software. That is, one can generate a large number of test statistics under $\boldsymbol{H0}$ and under $\boldsymbol{H1}$, *as if* the study had been conducted a large number of times. For example, one may simulate test statistics that correspond to results from 10K draws from the assumed population. Doing so results in a matrix of 10K rows and M columns for both $\boldsymbol{t0}$ and $\boldsymbol{t1}$. Additionally, $\boldsymbol{t0}$ and $\boldsymbol{t1}$ can be converted to 10K x M matrices of p-values, $\boldsymbol{p0}$ and $\boldsymbol{p1}$.

Once $\boldsymbol{t0}$ and $\boldsymbol{t1}$, as well as $\boldsymbol{p0}$ and $\boldsymbol{p1}$, have been generated, any of the MTPs can be implemented in order to obtain a 10K x M matrix of adjusted p-values.

For example, since each row of $\boldsymbol{p1}$ contains, for a single sample, the raw p-values that one could obtain for M effect estimates when there are true effects equal to the MDESs specified under $\boldsymbol{H1}$, these p-values can be easily adjusted using the Bonferroni, Holm, or Benjamini-Hochberg MTPs. Recall from Section 2 that for these MTPs, only the raw p-values are needed to make the adjustments. The adjustments are repeated in every row of the matrix, or for all 10K samples from the assumed population, resulting in a new matrix of p-values corresponding to any given MTP: $\widetilde{\boldsymbol{p}}^{BF-SS}$, $\widetilde{\boldsymbol{p}}^{HO-SD}$, or $\widetilde{\boldsymbol{p}}^{BH-SU}$.

It is more complicated to obtain p-values adjusted by the Westfall-Young single-step and step-down MTPs. As described in Section 2, in this MTP, observed test statistics (or p-values) can be compared with the distribution of test statistics (or p-values) under the complete null hypothesis. In the implementation for this paper, test statistics were used. Therefore, both $\boldsymbol{t0}$ and $\boldsymbol{t1}$ are used to obtain adjusted p-values. That is, to adjust p-values for one data sample, one row of $\boldsymbol{t1}$ is compared with all rows in $\boldsymbol{t0}$.

For each MTP, the resulting 10K x M adjusted p-values can then be compared with a specified value of α and null hypothesis rejections can be recorded. Doing so results in a 10K x M matrix of hypothesis rejection indicators from which all definitions of power can be computed:

- Individual power for the m^{th} outcome is the proportion of the 10K rows in which the m^{th} null hypothesis was rejected (the mean of the m^{th} column of indicators).
- d-minimal power is the proportion of the 10K rows in which at least d of the M null hypotheses were rejected.
- Complete power is the proportion of the 10K rows in which all of the null hypotheses were rejected based on the *raw* p-values rather than adjusted p-values. The reason that complete power is based on raw p-values is that the probability of all tests having a raw p-value less than α when the null hy-

pothesis is true is less than the probability that any single test would have a p-value less than α by chance (Koch & Gansky, 1996; Westfall et al., 2011).

In effect, the power estimation approach laid out above relies on simulation, but rather than (first) simulating a large number of datasets, (second) carrying out impact analyses on each simulated dataset, and (third) adjusting the resulting p-values from each analysis, the approach skips to the third step, saving lots of effort and computing time.[11]

Note that this approach of simulating test statistics builds on work by Bang, Young, and George (2005), who use simulated test statistics to identify critical values based on the distribution of the maximum test statistics. Their approach produces the same estimates as the approach described here for the single-step Westfall-Young MTP. Chen et al. (2011) derived explicit formulas for *d*-minimal powers of stepwise procedures and for complete power of single-step procedures, but only for up to three tests. The approach presented here is more generally applicable, as it can be used for all MTPs, for any number of tests, and for all definitions of power discussed in the present paper.

To check that the power estimates obtained from the methodological approach just described are correct, three validation analyses were conducted. First, for the design of interest (a blocked RCT) and the assumed model (with constant effects across all blocks and with block dummies included in the intercept), estimates of individual power for a *single hypothesis test* were compared with those computed in PowerUp! (Dong & Maynard, 2013, Table RBD2-c). The comparisons, which match closely, can be found in Appendix C, Table C.1. Second, *assuming a single block*, individual power estimates after adjusting with the Bonferroni, Holm, and Benjamini-Hochberg MTPs were compared with power-estimation results in Schochet (2008). Power estimates for Westfall-Young MTPs are not found in this paper. Results of these comparisons, which also match closely, can be found in Table C.2. For the third validation exercise, a selection of results obtained from the methodology described above — for all definitions of power examined in this paper — were compared with power estimates obtained from Monte Carlo data simulations. In these simulations, 2,000 samples were generated according to the assumed study design and model. In each data sample, M regression models specified as in (4) were fit, and M effect estimates and corresponding raw p-values were computed and adjusted. Then each definition of power was computed the same way as described above. Table C.3 shows comparisons between power estimates obtained with these data simulations and results obtained with the approach above, which skips straight to the simulation

[11] When the power estimation methodology is coded in R (as shown in Appendix B), all power estimates for all MTPs other than the Westfall-Young MTPs take less than one minute. Power estimates for Westfall-Young MTPs take a few minutes — depending on the number of samples, processing power, and degree of parallelization available.

of test statistics. Again, the comparisons are extremely close. Together, the three validation exercises demonstrate the accuracy of the methodology proposed in this paper.

3.2 An Example of Estimating Power When Adjusting for Multiple Tests

Suppose researchers are designing a multisite trial in which they plan to investigate the effects of an education intervention on three confirmatory outcomes — assessments in three different subject areas. Based on prior research, they assume that the correlation between all pairs of these outcomes is 0.5. They plan to use the model specified in (4) to estimate effects, and they will have a baseline measure of each assessment, each with an R^2 of 0.4. Because the sites (the blocks) will also explain variation in the outcomes, they assume an overall R^2 of 0.5 for all three impact models. Additionally, they plan to have 20 sites in the study, with 50 individuals per site, and 50% of the individuals at each site will be randomly assigned to the treatment group. To counteract the multiplicity problem, they plan to use the Holm correction to control the FWER at 5%. The researchers expect the intervention to have an effect on all three outcomes with an effect size of at least 0.125 standard deviations. They want to be able to detect effects as small as this size; therefore the desired MDES for each outcome is 0.125. While they expect effects on all outcomes, after discussing their study with stakeholders the researchers realize that policymakers would consider the intervention to be a success if it raises test scores by at least 0.125 standard deviations in *at least one* subject area. Therefore, the researchers define power in their study as the probability of rejecting at least one of an assumed three false null hypotheses (1-minimal power).

If the researchers ignore the fact that they will make adjustments for multiplicity, they would estimate that the study has individual power of 80% for each outcome, given their assumptions. However, they want to know the power of their study when they take their actual analysis plan into account. First, they generate $\boldsymbol{t1}$. Therefore, they simulate a 10K-row x 3-column matrix of test statistics following a multivariate t-distribution with correlation matrix

$$\begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}, \text{and means equal to}$$

$$\frac{MDES(m)}{Q(m)} = \frac{MDES(m)}{\sqrt{\frac{1-R^2(m)}{\bar{T}(1-\bar{T})Jn_j}}}$$
$$= \frac{0.125}{\sqrt{\frac{1-(0.5)^2}{0.5(0.5)(20)(50)}}} \qquad (8)$$
$$= 2.3$$

for all m, and $df(m) = Jn_j - g^*(m) - 1 = 20(50) - 21 - 1$ = 1,978 for all m. They then convert each test statistic in their 10K x 3 matrix to a p-value. The resulting matrix of p-values $(\boldsymbol{p1})$ is a simulation of raw, or unadjusted, p-values that would be obtained by estimating impacts 10K times (in 10K samples from the target population). Next, the researchers adjust the three p-values in each of the 10K rows, following the Holm procedure, as described in the previous section. Finally, since they focus on 1-minimal power, their statistical power is the proportion of the 10K rows in which *at least one* of three p-values is less than 0.05.

They find that 1-minimal power — the probability of detecting at least one true effect with effect size 0.125 or greater — is 87% if such effects actually exist on all three outcomes. That is, if there are impacts of a magnitude at least as large as a 0.125 effect size on all three outcomes, they have an 87% chance of a statistically significant effect estimate for at least one of them. Their power is better than the typical 80% standard. With 80% 1-minimal power, their MDES is smaller than 0.125; it is 0.114. Alternatively, they can include 17 sites with 50 individuals instead of 20 sites with 50 individuals to achieve at least 80% power for an MDES of 0.125 (1-minimal power is 82% in this case). Also, while 1-minimal power is sufficient, they may want to be able tell their stakeholders that with their original MDES and sample size specifications, the probability of detecting impacts on at least two of the three outcomes is 73% but that the probability of detecting impacts on all three outcomes is 61%.

3.3 Notes About the Assumptions

Before embarking on power calculations, the researchers in the example above had to decide on the number of outcomes for which they would adjust for multiplicity, the MTP they would use to make those adjustments, and the definition of power that best fit with the objective of their study. They also made a set of assumptions for each outcome that corresponded to those they would have made if they had only had one outcome. That is, they assumed the number of blocks; the number of individuals within blocks;[12] the proportion of individuals assigned to the treatment group; the explanatory power of baseline covariates, including block indicators (R^2); and an MDES. In the above example, the researchers assumed the same R^2 and the same MDES for all outcomes. However, these two may often vary by outcome in practice.

In addition, the researchers must make some new types of assumptions that only come into play when estimating power that accounts for multiplicity adjustments. First, they must assume the correlations between the test statistics. These M pairwise correlations are equal to the M pairwise correlations between the residuals in the M impact models. If there are no covariates in the impact models or if the $R^{2\prime}s$ of the covariates are equivalent in all impact models, then the correlations between the test statistics are equal to the correlations between the outcomes. However, having different $R^{2\prime}s$ across the impact models reduces the correlations between the residuals and therefore between test statistics.[13] Models of outcomes that are highly correlated are more likely to have residuals that are highly correlated because baseline covariates will tend to have similar $R^{2\prime}s$. The gaps between the correlations between outcomes and the correlations between residuals — and therefore the test statistics — may be wider for moderately or weakly correlated outcomes. In any case, the upper bounds of correlations between the test statistics are the correlations between the outcomes.

The second new assumption that must be considered when estimating power that takes multiplicity adjustments into account is the proportion of outcomes on which there are truly impacts of at least the size of the researchers' desired MDESs, or, equivalently, the number of truly false null hypotheses. There is one scenario in which this assumption does not matter, which is the scenario when one focuses on individual power and uses a single-step MTP. In this case, when adjusting a p-value for a single test, the information from other tests is disregarded. For all other scenarios, however, this assumption can be an important one.

Researchers may be inclined to assume that there will be effects on all outcomes, as hypotheses of effects probably drive the selection of outcomes in the first place. And when estimating power for a single hypothesis test, power is only defined when a true effect exists. However, as will be shown in the next section, if the researchers are incorrect and there turn out not to be effects on all outcomes, the probability of detecting the effects that do actually exist can be diminished, sometimes substantially.

[12] When the number of individuals per block is not the same within each block, then n_j is assumed to be the harmonic mean of the numbers of individuals per block (Bloom, 2006).

[13] For example, one of the multiple outcomes may have a baseline covariate with a high R^2 while another may have a baseline covariate with a smaller R^2. Also, block dummies may explain more variation in some outcomes than in others.

It is important to point out that under the assumption that there are not truly effects on every outcome under study, the definitions of the *d*-minimal powers (e.g., 1-minimal power, 1/3-minimal power, etc.) and of complete power become fuzzy. For example, 1/3-minimal power is defined as the probability of detecting effects (of a specified size or larger) on at least 1/3 of the total outcomes (*M*), regardless of the number of outcomes with actual effects. That is, 1/3-power is *not* defined as the probability of detecting effects among the *M* outcomes on which the effects truly exist. Therefore, while power is *technically* defined based on false nulls, the definition is loosened here and includes the probability of erroneous rejections of false nulls (which are controlled to occur at no more than 5% for those MTPs that control the FWER). This fuzziness of definition is needed because the researcher would only ever define power based on the total number of tests. Moreover, if the d-minimal powers are defined only based on truly false nulls, then their levels could *increase* when the proportion of false nulls decreases. Complete power has the same issue. If there are truly only effects on two of the three outcomes, then complete power is not the probability of rejecting just two false null hypotheses. In this case, complete power is undefined.

4. Empirical Findings on How Various Factors Affect Power

This section uses the power estimation approach in Section 3 to investigate how power varies with the many factors that affect it in studies that adjust for multiplicity due to testing for effects on multiple outcomes. Sticking with the example of a blocked RCT with 20 blocks of 50 individuals, in which half are assigned to the treatment group, in which the targeted MDES is 0.125 for all outcomes on which there are effects, and in which effects will be estimated with the model in (4), the following factors are varied as described below:

- *The number of outcomes.* This number is equivalent to the number of hypothesis tests, and is specified to be 3, 6, 9, or 12.
- *The definition of power.* The following definitions are considered: individual power (for each individual outcome, the probability of detecting a true effect as large as the specified MDESs); 1-minimal power, 1/3-minimal power, and 2/3-minimal power (across all outcomes with true effects as large as the specified MDESs, the probability of detecting at least 1, 1/3, and 2/3 respectively); and complete power (the probability of detecting effects as large as the specified MDESs for all outcomes)
- *The MTP used.* Each of the five MTPs discussed in Section 2 is explored.
- *The correlations between the test statistics*.

- *The explanatory power of the covariates (R^2's).* It is well known that higher R^2's are associated with more power. The point of varying the R^2's here is to investigate how they affect the *relative* power when comparing the different MTPs with each other and with the situation when no adjustments are made. The benchmark R^2's are 0.5 for all outcomes, and they are lowered to 0.1 for comparison. The R^2's are assumed to be the same for all outcomes; therefore the correlations between the test statistics equal the correlations between the outcomes.

- *The proportion of outcomes on which there are truly impacts at least as large as the specified MDESs.* This proportion is of course unknown to researchers, but as discussed above, it is an assumption that needs to be considered.

4.1 Findings for Individual Power

Figure 1 presents estimates of individual power for 20 blocks of 50 individuals, assuming an MDES of 0.125 and an R^2 of 0.5 for all outcomes. With this set of assumptions, individual power for a single hypothesis test (or for the situation when no multiplicity adjustments are made) is 80%. Plot (a) in the figure presents estimates when the correlation between all pairs of outcomes is low, 0.2, and plot (b) in the figure presents estimates when this correlation is high, 0.8.

Along the top *X*-axis in both plots, the number of outcomes is varied (3, 6, 9, or 12) and along the bottom *X*-axis, the MTP's are varied within each number of outcomes. The shadings of the dots (as explained in the legend at the bottom of the page) indicate the proportion of the outcomes on which there are truly effects. Within each column, the darkest-shaded dot indicates individual power when there are truly effects on all three outcomes, the medium-shaded dot indicates individual power when there are truly effects on 2/3 of the outcomes, and the lightest-shaded dot indicates individual power when there are truly effects on just 1/3 of the outcomes. Note that for the single-step MTPs there is just one dot, because as discussed earlier, the proportion of outcomes with true effects does not affect power when using single-step MTPs.

Figure 1, plot (a) shows that compared with individual power when conducting just one hypothesis test (80%), after adjusting for multiplicity individual power can be — but is not necessarily — substantially lower. As expected, the extent of power loss depends on the number of outcomes and the MTP used. For stepwise MTPs, the extent of power loss also depends on the proportion of outcomes with true effects at least as large as 0.125 standard deviations. However, even if one were to assume that only 1/3 of the outcomes truly have effects, the stepwise MTPs still improve upon their single-step counterparts. This improvement can be seen by comparing HO-SD with BF-SS and WY-SD with WY-SS.

As expected, Benjamini-Hochberg (BH-SU), which controls the FDR, results in the least power loss compared with the situation when no adjustments are made. This MTP's power advantage over the other MTPs that control the FWER is more pronounced when there are more hypothesis tests. With as many as 12 hypothesis tests, the individual power is 75% in the case that there are truly effects on all outcomes. While power drops off considerably when there are truly effects on just 2/3 or 1/3 of the outcomes, the power that remains after adjusting with BH-SU is substantially greater than the power that remains after adjusting with any of the other MTPs.

A lesson here is that when there are a large number of hypothesis tests, BH-SU is greatly preferred for preserving individual power. With this many hypothesis tests, using BH-SU, and thereby controlling the FDR, may also make sense — with as many as 12 tests, researchers may be willing to tolerate an increased likelihood of a false positive finding because BH-SU is designed to produce false positive findings only along with many true positive findings. On the other hand, with a small number of tests, BH-SU may not make sense even though it results in the best power, because an erroneous rejection could alter the conclusions about an intervention's effectiveness.

Of the MTPs that control the FWER, the stepwise procedures (HO-SS and WY-SS) perform almost equivalently when the correlation between the test statistics is low (0.2), as in plot (a). When the test statistics are highly correlated (0.8), as shown in plot (b), WY-SD results in more power than HO-SS. In addition, when test statistics are highly correlated, WY-SD produces a level of individual power that is much closer to BH-SU, compared with the situation when test statistics are modestly correlated. In sum, to limit the probability of a false positive finding across a set of tests and to maximize individual power, the WY-SD MTP, which takes the correlation of test statistics into account, may be worth the added computational complexity when the correlation between tests is large. However, HO-SD, which is much simpler and which can be directly computed from raw p-values, is also a good choice for controlling the FWER when the correlation between test statistics is not high or the number of tests is small.

Figure 2 presents the same plots as Figure 1 but in these plots, the R^2 for all outcomes is lowered from 0.5 to 0.1, while all other assumptions remain the same. In this case, power for a single hypothesis test is lowered from 80% to 67%, as seen by the dashed horizontal line in the plots. The main lesson of the plots in Figure 2 is that regardless of the MTP used, a lower R^2 increases the power losses relative to the situation when only one hypothesis test is conducted or when no adjustments are made. This increased power loss can be seen in the greater distances between the dots and the dashed horizontal lines in Figure 2 compared with Figure 1.

4.2 Findings for 1-Minimal, 1/3-Minimal, and 2/3-Minimal Power

Figures 3 and 4 present estimates of 1-minimal power: the probability of detecting at least one true effect at least as large as the specified MDESs. The plots in these figures are similar to those already presented except that now along the top *X*-axis, the correlation between test statistics is varied, from 0 to 0.9. In Figure 3 the number of tests is held constant at three, and in Figure 4 the number of tests is held constant at six. All other assumptions are the same as in earlier plots. The benchmark power level obtained when testing just one hypothesis is again 80%.

Figure 3 demonstrates that with three uncorrelated false nulls, the probability of rejecting at least one of them is substantially greater than the benchmark power level. As the correlation increases, this probability declines but still remains at or above the benchmark of 80%, regardless of the MTP used, unless the correlation is as high as 0.9 and an MTP other than one of the WY options is used. When just two out of three of the null hypotheses are actually false (meaning there are true effect sizes of at least 0.125), as seen by the medium-shaded dots, the probability of rejecting at least one null (of three, not two, as discussed earlier) is higher than the 80% benchmark when the correlation is 0.5 or less. It is only when just one of the three null hypotheses is actually false that there is a substantial loss of power compared with the benchmark.

A comparison of Figure 4 with Figure 3 shows that, regardless of the proportion of null hypotheses that are truly false and regardless of the MTP used, 1-minimal power improves with more tests. With six tests, even when just 1/3 of them are actually false, 1-minimal power is not far from the 80% benchmark. This result does not imply that researchers should test for effects on a large number of outcomes to improve their chances of finding impacts. Rather, researchers should focus on the primary outcomes among which at least one needs to have a statistically significant finding in order for there to be policy implications.

Both Figures 3 and 4 also show that the choice of MTP matters much less when focusing on 1-minimal power. All MTPs result in similar power levels when the test statistics have a low or moderate correlation. When test statistics are highly correlated, the Westfall-Young MTPs are preferred, and the simpler single-step version is sufficient.

Figure 5 focuses on 1/3-minimal power while holding the number of tests fixed at six, and Figure 6 focuses on 2/3-minimal power while holding the number of tests fixed at six. Recall that 1/3-minimal power (or 2/3-minimal power) is the probability of detecting effects of a specified size or larger on at least 1/3 (or 2/3) of the total number of outcomes (*M*), regardless of the number of outcomes with actual effects. With 1/3-minimal power, the trends are similar to those observed for 1-minimal power. However, the proportion of outcomes with true effects matters more and the choice of MTPs matters more. There can still be improvements over the benchmark when correlations are low and effects exist on all outcomes. With 2/3-minimal power, the story is quite different. Figure 6 shows that if researchers need to detect effects on at least four of six outcomes after adjusting for multiplicity, then the probability of detecting those effects is substantially less than 80% for most correlations and MTPs.

4.3 Findings for Complete Power

Figure 7 presents results for complete power — the probability of statistically significant effect estimates of impacts for all outcomes on which there are truly effects. Recall from earlier that when focusing on complete power, p-values are not adjusted. Therefore, Figure 7 does not have different results for different MTPs. The *X*-axis in Figure 7 is the correlation between the test statistics. For each correlation, the figure shows the probability of rejecting all of two, three, four, five, or six null tests. As shown in the legend, the darkest dot is for two tests and the lightest dot is for six tests.

The primary lesson of Figure 7 is that if researchers follow current standard practice and only estimate power for a single hypothesis test (so that their assumed power is 80%) and if the success of the intervention under study requires evidence of effects on all of multiple tests, then their study is probably substantially underpowered. The extent to which the study is underpowered depends on the number of hypothesis tests and the correlation between the tests. Take for example the study assumptions in the plot and a correlation of 0.5 between all pairs of test statistics. If researchers need to detect effects on three of three outcomes, and effects truly exist on all three, then the probability of detecting all three effects is 60%. In order to increase this probability to 80%, they would need to increase the number of blocks from 20 of 50 individuals to 28 of 50 individuals. Otherwise, they would have to be able to assume MDESs on all outcomes of 0.148 instead of 0.125.

5. Discussion

This section summarizes the empirical findings on how various factors affect statistical power when adjusting for multiplicity due to estimating effects on multiple outcomes in a blocked randomized trial. It then provides some recommendations for practice and concludes with next steps.

5.1 Summary of Findings

With Respect to Number of Outcomes

When researchers are considering the number of outcomes across which they will make multiplicity adjustments, the implications depend on (1) which definition of power makes sense for their study and (2) which MTP they use. If the researchers are focusing on individual power, then having more outcomes will lead to a decrease in power. This decrease may not be very substantial with the Benjamini-Hochberg MTP, which controls the FDR, but power drops off much more dramatically with all other MTPs when additional outcomes are added. If researchers are focusing on complete power (the power to detect effects at least as large as the MDESs on all outcomes), then having more outcomes also leads to a loss of power. In this case, the amount of power lost depends on the correlation between the tests. The same is true to a lesser extent for power to detect a majority of effects (e.g., 2/3-minimal power). If researchers are focusing on 1-minimal power, the probability of detecting at least one effect increases with the number of outcomes.

With Respect to Correlations Between Test Statistics

The correlations between test statistics have nontrivial implications for all types of power. These correlations, which are the pairwise correlations of the residuals in the individual regression models, have an upper bound of the pairwise correlations between the outcomes and will be lower when the baseline covariates in the models have different $R^{2\prime}$s. For individual power-of-multiple-hypothesis tests, the loss of power compared with the situation when there is just one hypothesis test is greater with higher correlations between test statistics. Higher correlations between tests also mean that the Westfall-Young MTPs, which take dependencies in the data into account, are worth implementing to maximize power when controlling the FWER. The step-down version in particular maximizes power the most. Next, 1-minimal power and 1/3-minimal power are maximized with independent tests and typically decrease with higher correlations between tests — except when the proportion of nulls that are false is small. For 2/3-minimal power, the impact of the correlation varies with the MTP used and the proportion of nulls that are false. Finally, complete power improves substantially with higher correlations between test statistics.

With Respect to the Proportion of Outcomes with True Effects

Strong hypotheses of effects probably influence researchers' selection of outcomes. It may therefore seem unnecessary to assume true effects on only a subset of outcomes. However, the empirical findings in the section above show that if researchers make a mistake and there are *not* truly effects on *all* outcomes, there can be substantial consequences for detecting those effects that actually do exist.

With Respect to the $R^{2\prime}s$ of Baseline Covariates

Finally, while it is well known that higher $R^{2\prime}$s are associated with greater power, it tends also to be the case that higher $R^{2\prime}$s provide some protection against power losses from multiplicity adjustments (compared with power when estimating effects on one outcome).

Higher $R^{2}{}'$s may also diminish the power gains of 1-minimal and 1/3-minimal power, due to a ceiling effect.

5.2 Recommendations for Practice

The following recommendations for practice are based on the findings in this paper:

1. **Prespecify all hypothesis tests and prespecify a plan for making multiplicity adjustments.**

This paper has demonstrated that if one plans to use MTPs to adjust for multiple tests, the change in statistical power can be substantial. Therefore, it seems essential to plan ahead and take the consequences of the intended adjustments into account when designing one's study. Otherwise, in some cases, sample sizes may be too small, and studies may be underpowered to detect effects as small as a desired size. In other cases, sample sizes may be larger than needed, or studies may be powered to detect smaller effects than anticipated.

2. **Think about the definition of success for the intervention under study and choose a corresponding definition of statistical power.**

The prevailing default in education studies — individual power — may or may not be the most appropriate type of power. In some cases, it may provide misleading estimates of the probability that researchers will be able to find sufficient evidence that an intervention was successful. If the researchers' goal is to find statistically significant estimates of effects on all primary outcomes of interest, then even after taking multiplicity adjustments into account, estimates of individual power can grossly understate the actual power required — complete power. On the other hand, if the researchers' goal is to find statistically significant estimates of effects on at least one or on a small proportion of outcomes, then their power may be much better than anticipated. They may be able to get away with a smaller sample size, or they may be able to detect smaller MDESs.

The choice of power definition may not be a simple one. First, it may not be easy to define the success of an intervention. Even when it is easy, aligning the definition of success with a definition of power may not always be. For example, even if a program would be considered successful should an effect of a specified size be found for at least one outcome, researchers may still want sufficient individual power because they want to know the probability of detecting effects on each particular outcome.

It may be best for researchers to estimate and share power estimates for multiple power definitions. For example, consider the case in which a sample size is fixed. The probability of detecting statistically significant effects (at least as large as specified MDESs) may be unacceptably low. While complete power may be a goal in this case, it may be valuable for researchers to also be able to say that it is still tenable to achieve a high probability of detecting effects on at least half of the outcomes.

3. **Consider whether it is more appropriate to control the FWER or the FDR.**

Even though the Benjamini-Hochberg MTP, which controls the FDR, generally results in the most power, it may not necessarily be the best MTP to use. An MTP that controls the FDR is more lenient with false positives. Researchers may tolerate a few false positives when testing for effects on a large number of outcomes. However, when investigating effects on a small number of outcomes, a single false positive is more likely to lead to the wrong conclusion about an intervention's effectiveness. Therefore, with a small number of outcomes, controlling the FWER is likely to be preferable.

If researchers determine that it makes sense to control the FDR, they should use the Benjamini-Hochberg MTP. When controlling the FWER, the Westfall-Young step-down MTP generally results in the most power. However, if there will be a low or moderate correlation between outcomes or if the study will use a 1-minimal definition of power, the Holm MTP or the single-step Westfall-Young MTP may suffice.

4. **Consider the possibility that there may not be impacts on all outcomes.**

For the reasons summarized in Section 5.1, it is important to incorporate this possibility when estimating power.

5. **Take all of the above into account in the design phase of a study to estimate power, sample size requirements, or MDESs.**

Working through recommendations (1) to (4) is not a linear process. Each affects the others. For example, using a 1-minimal definition of power will allow researchers to consider more outcomes without any power loss, whereas other definitions of power may mean that they want to be very parsimonious in selecting their primary outcomes. Also, the Benjamini-Hochberg MTP may be preferable for a large number of outcomes, but a 1-minimal definition of power may mean that the Benjamini-Hochberg MTP is too dangerous, as the elevated chance of a false positive finding may not be tolerable when success rests on just one statistically significant effect.

5.3 Next Steps

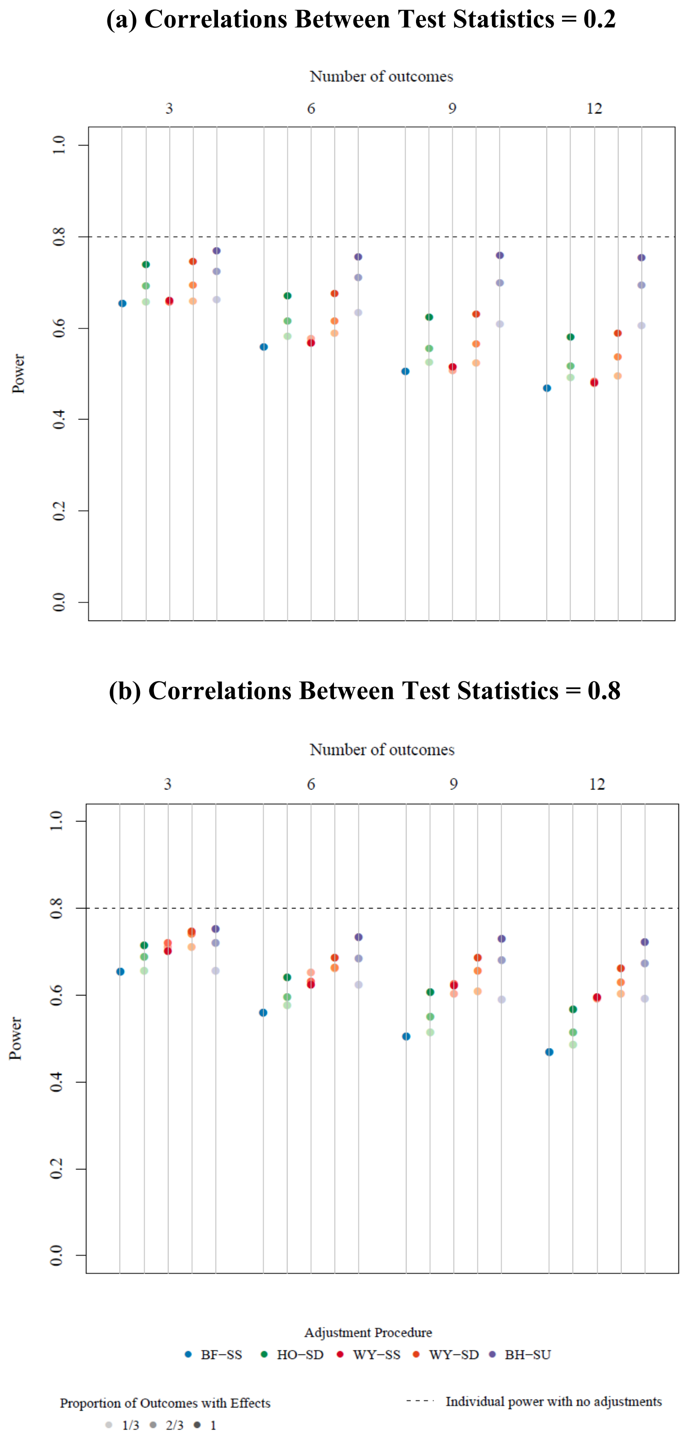
This paper focused on a blocked RCT in which effects are estimated using a model with block-specific intercepts and with the assumption of constant effects across blocks. Extensions to other analysis assumptions and designs should be straightforward. They would simply involve defining $Q(m) \equiv SE\left(\widehat{ES}(m)\right)$, which is a function of the standard error of the effect estimator in the regression model used. Then, once we know $Q(m)$ and an assumption for the correlations between test statistics, we can generate those test statistics and use them to empirically estimate all definitions of power for all MTPs.

This paper also focused on studies investigating effects on multiple outcomes. A next step for this research is to extend the methodology to estimate power when multiplicity adjustments are needed due to estimating effects on multiple subgroups, at multiple points in time, or across multiple treatment groups.

Finally, the R code that implements the power estimation method in Section 3 (see Appendix B) only allows a user to estimate power for a specified sample size and for specified MDESs. Another next step will be to develop code that allows users to enter a desired level of power and then return either a sample size or MDESs.

Figure 1

Individual Power, by Number of Outcomes, Adjustment Procedure, Proportion of Outcomes with Effects, and Pairwise Correlations Between Test Statistics: 20 Sites of 50 Individuals Each, R^2 = *0.5*, and MDES = 0.125 for All Outcomes on Which There Are Effects

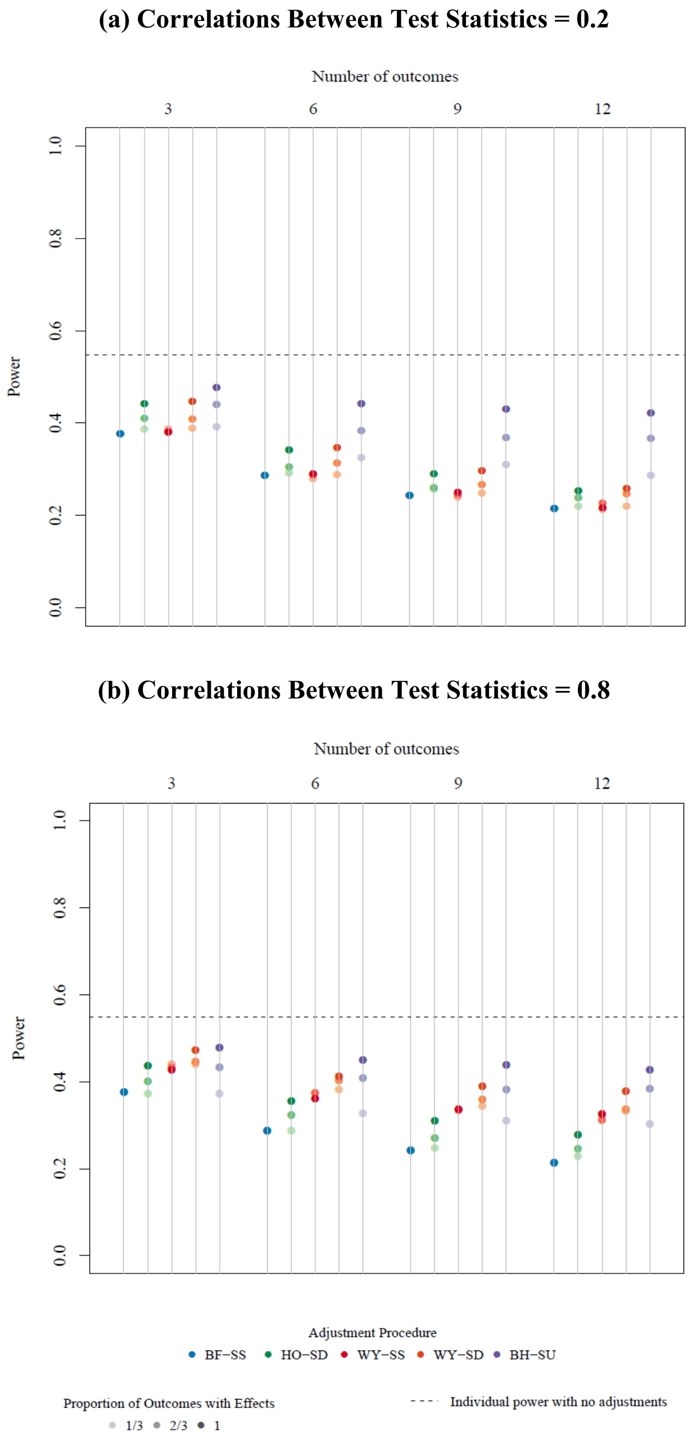


(a) Correlations Between Test Statistics = 0.2

(b) Correlations Between Test Statistics = 0.8

Figure 2

Individual Power, by Number of Outcomes, Adjustment Procedure, Proportion of Outcomes with Effects, and Pairwise Correlations Between Test Statistics: 20 Sites of 50 Individuals Each, R^2 = *0.1*, and MDES = 0.125 for All Outcomes on Which There Are Effects



(a) Correlations Between Test Statistics = 0.2

(b) Correlations Between Test Statistics = 0.8

Figure 3

1-Minimal Power, by Adjustment Procedure, Proportion of Outcomes with Effects, and Pairwise Correlations Between Test Statistics: _Three Outcomes_, 20 Sites of 50 Individuals Each, R^2 = 0.5, and MDES = 0.125 for All Outcomes on Which There Are Effects

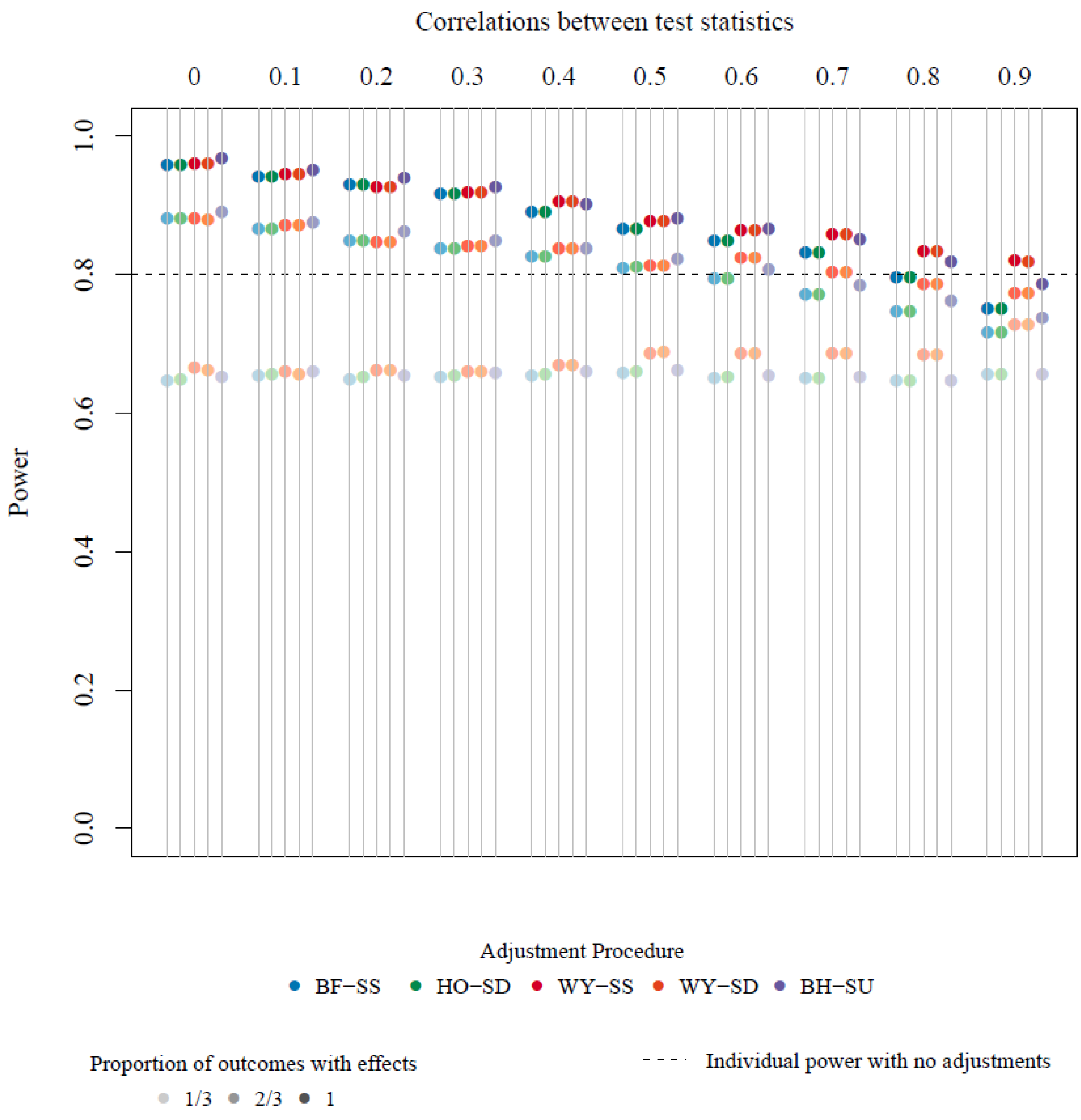


Figure 4

1-Minimal Power, by Adjustment Procedure, Proportion of Outcomes with Effects, and Pairwise Correlations Between Test Statistics: _Six Outcomes_, 20 Sites of 50 Individuals Each, $R^2 = 0.5$, and MDES = 0.125 for All Outcomes on Which There Are Effects

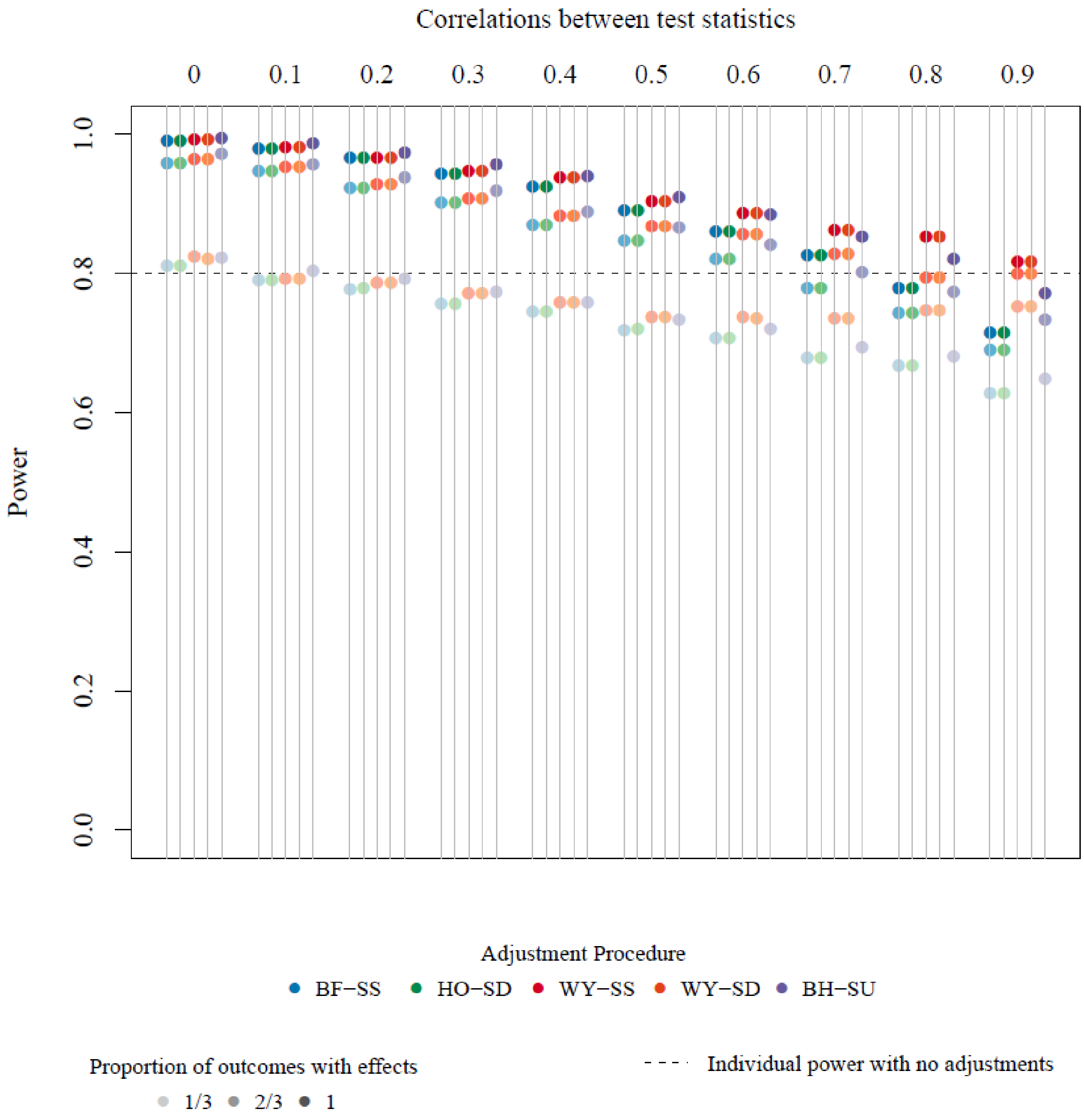


Figure 5

***1/3-Minimal Power*, by Adjustment Procedure, Proportion of Outcomes with Effects, and Pairwise Correlations Between Test Statistics: *Six Outcomes*, 20 Sites of 50 Individuals Each, R^2 = 0.5, and MDES = 0.125 for All Outcomes on Which There Are Effects**

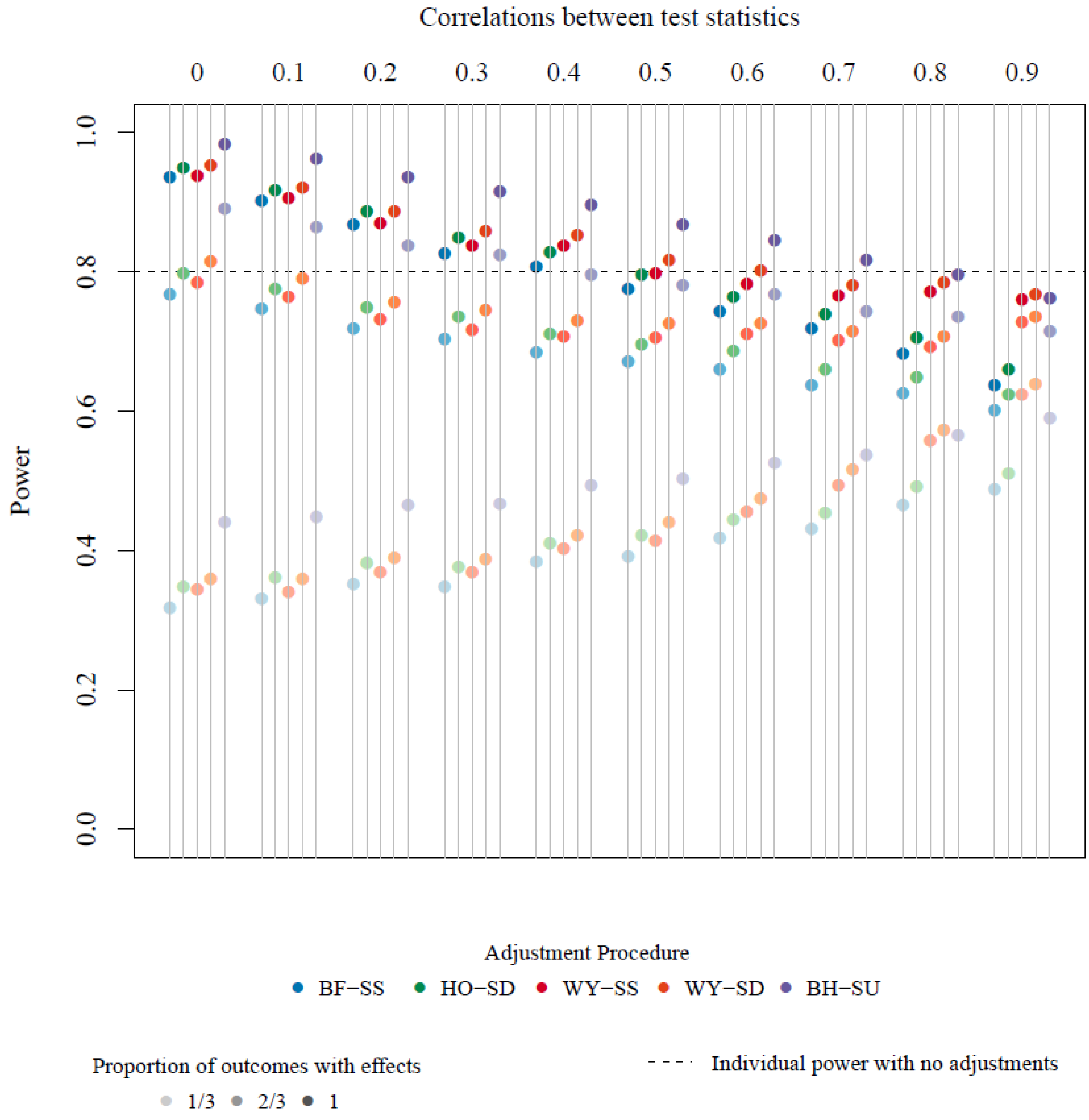


Figure 6

***2/3-Minimal Power,* by Adjustment Procedure, Proportion of Outcomes with Effects, and Pairwise Correlations Between Test Statistics: *Six Outcomes*, 20 Sites of 50 Individuals Each, R^2 = 0.5, and MDES = 0.125 for All Outcomes on Which There Are Effects**

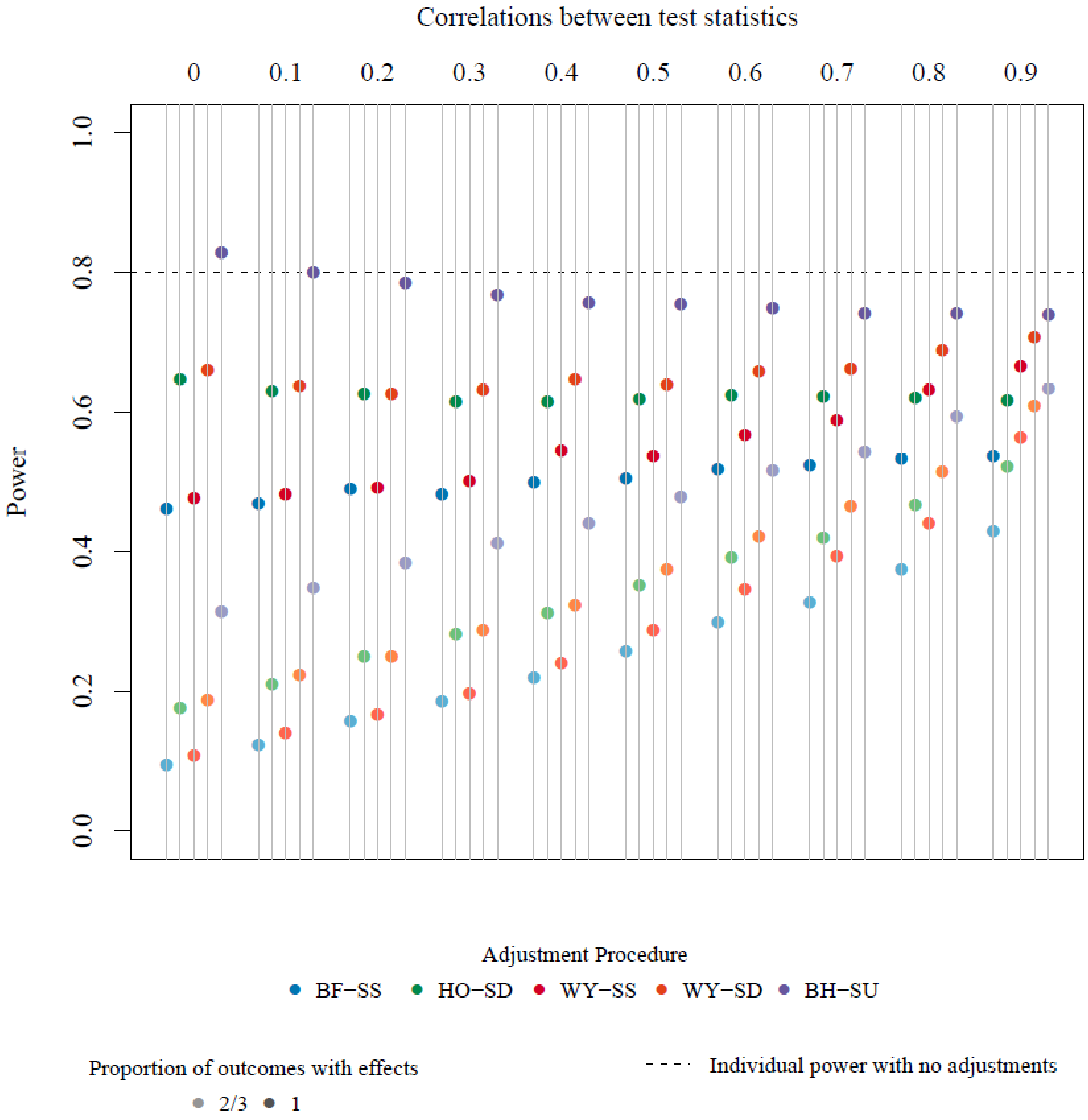
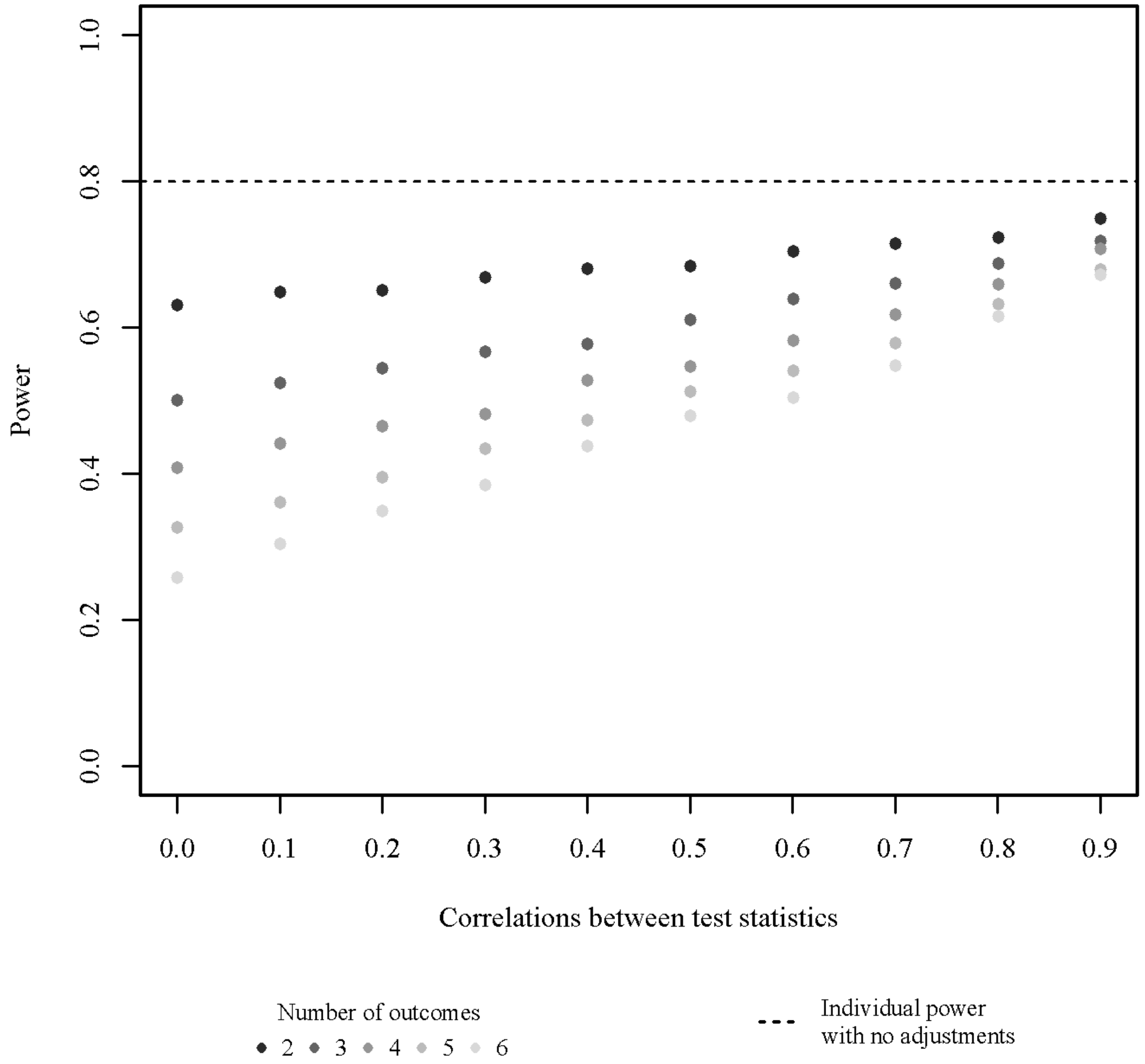


Figure 7

Complete Power, by Number of Outcomes and Pairwise Correlations Between Test Statistics: 20 Sites of 50 Individuals Each, $R^2 = 0.5$, and MDES = 0.125 for All Outcomes on Which There Are Effects



Appendix A

Descriptions of the Bonferroni, Holm, Westfall-Young Single-Step, Westfall-Young Step-Down, and Benjamini-Hochberg Multiple Testing Procedures (MTPs)

Bonferroni-Class MTPs That Control the Familywise Error Rate (FWER)

The *Bonferroni* MTP (Dunn, 1959, 1961) provides strong control of the FWER at level α by setting the significance level for each of M individual tests to α/M. However, for ease of presentation, researchers typically hold the significance level at the FWER level and multiply raw p-values by M (but truncating them at 1). Clearly, with this MTP, adjustments increase proportional to the number of hypothesis tests. Because the Bonferroni procedure adjusts the p-value for each test independently, it is referred to as a "single-step" procedure. The shorthand *BF-SS* is used to refer to the Bonferroni MTP in the remainder of this appendix, where "*SS"* is a reminder that it is a single-step MTP. The Bonferroni-adjusted p-value can be written as

$$\tilde{p}_j^{BF-SS} = \min\{Mp_j, 1\} \text{ for } j = 1, \ldots, M, \tag{A.1}$$

where p_j is the raw (unadjusted) p-value for the j^{th} test of M tests.

The *Holm* procedure (Holm, 1979) improves on the Bonferroni procedure with a stepwise approach. This procedure orders raw p-values from smallest to largest, such that $p_{r1} \leq p_{r2} \leq \cdots \leq p_{rM}$. The corresponding null hypotheses are then $H_{r1}, H_{r2}, \ldots, H_{rM}$. (The r subscripts here indicate that the p-values are ranked by their value.) The Holm procedure then rejects H_{rj} if and only if $p_{rj} \leq \alpha/(M - j + 1)$ for $j = 1, \ldots, M$. To frame the procedure in terms of p-value adjustments, as we did for Bonferroni, we multiply the smallest p-value by M, the second smallest p-value by *M-1*, and so on, but we also enforce that each adjusted p-value is greater than or equal to the previous adjusted p-value and that it is not greater than one. Using the shorthand *HO-SD* for Holm, where "SD" notes it is a step-down procedure, the adjusted p-value for this MTP is given by

$$\begin{aligned} \tilde{p}_{rj}^{HO-SD} &= \min\{Mp_{rj}, 1\} \text{ for } j = 1 \\ &= \min\{\max\{\tilde{p}_{r(j-1)}^{HO-SD}, (M - j + 1)p_{rj}\}, 1\} \text{ for } j = 2, \ldots, M. \end{aligned} \tag{A.2}$$

Note that all hypotheses rejected by Bonferroni will also be rejected by Holm, but it is possible that the Holm MTP will reject additional hypotheses.[1] In general, stepwise procedures improve on single-step methods by allowing the possible rejection of less significant hypotheses in subsequent steps, depending on the null hypotheses already rejected in previous steps. Therefore, step-down MTPs can result in greater power than single-step MTPs.

[1] The Hochberg (1988) procedure is another stepwise enhancement of the Bonferroni procedure, but is a "step-up" procedure rather than step-down. It is not included in our comparisons.

Resampling-Based MTPs That Control the FWER

Westfall and Young (1993) introduced both a single-step and a step-down procedure to control the FWER, both which take the correlations between test statistics into account. These procedures rely on estimating the joint distribution of test statistics when the "complete null hypothesis" is true. The complete null hypothesis ($\boldsymbol{H0}$) is that $ES(m) = 0$ for all m, or, that there are no impacts on any of the M outcomes. The distribution of test statistics can be estimated by resampling one's analysis dataset — using either permutation or bootstrapping. With permutation, for example, one randomly scrambles the treatment assignment indicator in the data, thereby breaking the relationship between treatment assignment and outcomes.[2] Impacts are estimated with the scrambled treatment indicator and test statistics are computed. After repeating this process many times, like 10K, one has 10K test statistics for each of *M* outcomes that were computed with data in which no effects exist. Therefore, the resulting 10K x *M* matrix of test statistics provides an approximation of the test statistics one would obtain for 10K data samples under $\boldsymbol{H0}$.[3] Because other than the treatment indicator, all other variables remain the same in the resampled data, the correlations between the test statistics across the *M* hypothesis tests are preserved.

When using the single-step version of the Westfall Young MTP (denoted by *WY-SS*), the adjusted p-value for the m^{th} hypothesis test is the proportion of samples in which the maximum random variable test statistic (in absolute value) is greater than or equal to the actual test statistic from the original, unpermuted data (in absolute value) for the m^{th} hypothesis test. Thus, the Westfall-Young single step p-values are given by

[2] When data are blocked, as they are in the research design analyzed in this paper, the permutation can be carried out separately for each block.

[3] The single-step and step-down Westfall Young MTPs always provide at least weak control of the FWER. In order for these procedures to provide strong control of the FWER, they require the assumption of subset pivotality (Ge, Dudoit, & Speed, 2003). The distribution of the unadjusted test statistics or p-values is said to have subset pivotality if for any subset of null hypotheses, the joint distribution of the test statistics or of the p-values for the subset is identical to the distribution under the complete null. A consequence of this assumption is that the resampling of test statistics or p-values can be done under the complete null hypothesis rather than under the unknown partial hypothesis (Ge et al., 2003).

$$\tilde{p}_m^{WY-SS} = \Pr(max_{1 \le l \le M}|T_l| \ge |t_m| \,|H0^C), ^4 \quad \text{(A.3)}$$

where T_l is the random-variable test statistic for the l^{th} test.

The step-down version of the Westfall Young MTP (*WY-SD*) is more complicated. To carry out this version, one first orders the test statistics for each test such that $t_{s1} \ge t_{s2} \ge \cdots \ge t_{sM}$. Then one compares $max\{|T_{s_1}|, \ldots, |T_{s_m}|\}$ with $|t_{s_1}|$, $max\{|T_{s_2}|, \ldots, |T_{s_m}|\}$ to $|t_{s_2}|$ and so on, until one compares $max\{|T_{s_m}|\}$ with $|t_{s_m}|$. Again monotonicity of adjusted p-values is enforced. Using notation to present what was just described, Westfall-Young step-down adjusted p-values are given by

$$\tilde{p}_{s_i}^{WY-SD} = max_{k=1,\ldots,i}\{\Pr\left(max_{l=k,\ldots,m}|T_{s_l}| \ge \left|t_{s_k}\right| \left|H0^C\right)\right\}. ^5 \quad \text{(A.4)}$$

An MTP That Controls the False Discovery Rate (FDR)

The *Benjamini-Hochberg* MTP (Benjamini & Hochberg, 1995) is a "step-up" procedure that provides strong control of the FDR when the test statistics are independent or exhibit "positive regression dependency,"[6] which means that no pairs of test statistics are negatively correlated (Benjamini & Yekutieli, 2001). In this procedure, raw p-values of the M tests are ranked from smallest to largest, as above, such that $p_{r1} \le p_{r2} \le .. \le p_{rM}$. Then k is defined as the maxi-

[4] Alternatively, one can convert test statistics to p-values to obtain the distribution of p-values under $\boldsymbol{H0}$. In this case, the adjusted p-values are given by $\tilde{p}_j^{WY-SS} = \Pr\left(min_{1 \le l \le M}P_l \le p_j \middle| H0^C\right)$. When the test statistics are identically distributed, as is the case in this paper, estimating the null distribution of p-values and estimating the null distribution of test statistics produce the same adjusted p-values. However, when the test statistics are not identically distributed (e.g., they have different degrees of freedom), not all of the tests contribute equally to the p-values when the maximum-test-statistics version of the MTP is used, which can lead to unbalanced adjustments (Beran, 1988; Westfall & Young, 1993). When the minimum p-value version is used and permutation is used to estimate the complete null distribution of p-values, the adjusted p-values can be sensitive to the number of permutations when the number of tests is large, and they tend to be more conservative than the maxT-adjusted p-values (Ge et al., 2003).

[5] If one instead obtains the distribution of p-values under $H0^C$, then one orders the raw p-values such that $p_{r1} \ge p_{r2} \ge \cdots \ge p_{rM}$, then in this case, the adjusted p-values are given by
$\tilde{p}_{r_i}^{WY-SD} = max_{k=1,\ldots,i}\{\Pr\left(min_{l=k,\ldots,m}P_{r_l} \ge p_{r_k} \left|H0^C\right)\right\}$.

[6] Benjamini and Yekutieli (2001) point out that the condition for positive dependency is general enough for most problems of practical interest. For other forms of dependency, they provide a modification of the original procedure. However, this modification comes with a substantial loss of power. Several studies using simulated data that violate Benjamini-Hochberg's assumptions have shown that in practice, it still works quite well (e.g., Groppe, Urbach, & Kutas, 2011). Because adjustments for other forms of dependency are typically not needed, the What Works Clearinghouse uses the "original Benjamini-Hochberg procedure rather than its more conservative modified version as the default approach to correcting for multiple comparisons" (U.S. Department of Education, 2014).

mum j for which $p_{rj} \leq \frac{j}{M}\alpha$, and all null hypotheses H_{0_j} for $j = 1, \dots, k$ are rejected. Adjusted Benjamini-Hochberg p-values are therefore given by

$$\tilde{p}_{rj}^{BH-SU} = min_{k=1,\dots,j}\left\{\min(\frac{m}{k}p_{r_k}, 1)\right\}. \tag{A.5}$$

Here, the notation BH-SU is used to indicate that this is a step-up MTP. The BH-SU MTP originally assumed independent test statistics (Benjamini & Hochberg, 1995), but Benjamini and Yekutieli (2001) show that it can typically be applied when the test statistics have positive dependency. As the What Works Clearinghouse guidelines point out, the condition for positive dependency makes BH-SU highly applicable in practice, and while a modification to the BH-SU MTP can be made for other forms of dependency, it is typically unnecessary and is more conservative (U.S. Department of Education, 2014). The What Works Clearinghouse uses the Benjaimin-Hochberg MTP whenever it makes corrections for multiplicity (U.S. Department of Education, 2014).

Appendix B

R Code for Implementing Power Estimation Methodology

The function **comp.rawt.SS** is needed to implement the Westfall-Young single-step multiple testing procedure (MTP). It operates on one row of null test statistics. It compares $max_{1\leq l\leq M}T_l$ with $|t_m|$ for all m.

The parameters for this function are defined as follows:

- abs.Zs.H0.1row = 1 row of tests statistics under the complete null.
- abs.Zs.H1.1samp = raw test statistics for 1 sample.

```
comp.rawt.SS <- function(abs.Zs.H0.1row, abs.Zs.H1.1samp) {
 M<-length(abs.Zs.H0.1row)
 maxt <- rep(NA, M)
 for (m in 1:M) {maxt[m] <- max(abs.Zs.H0.1row) > abs.Zs.H1.1samp[m]}
 return(as.integer(maxt))
}
```

The function **adjust.allsamps.WYSS** is also needed to implement the Westfall-Young single-step MTP. It carries out the above function for all rows in the matrix of test statistics under the complete null and does so for all samples of raw test statistics under the alternative hypothesis.

The parameters for this function are defined as follows:

- snum = the number of samples for which test statistics under the alternative hypothesis are compared with the distribution (matrix) of test statistics under the complete null.
- abs.Zs.H0 = a matrix of tests statistics under the complete null.
- abs.Zs.H1 = a matrix of raw test statistics under the alternative.

```
adjust.allsamps.WYSS<-function(snum,abs.Zs.H0,abs.Zs.H1) {
 adjp.WY<-matrix(NA,snum,ncol(abs.Zs.H0))
 doWY<-for (s in 1:snum) {
  ind.B<-t(apply(abs.Zs.H0, 1, comp.rawt.SS, abs.Zs.H1.1samp=abs.Zs.H1[s,]))
  adjp.WY[s,]<-colMeans(ind.B)
 }
 return(adjp.WY)
}
```

The function **comp.rawt.SD** is needed to implement the Westfall-Young step-down MTP. It operates on each row of null test statistics. It first orders the test statistics such that

$t_{s1} \geq t_{s2} \geq \cdots \geq t_{sM}$, and then compares $max|T_{s_1}|, \dots, |T_{s_m}|$ with $|t_{s_1}|$, $max|T_{s_2}|, \dots, |T_{s_m}|$ with $|t_{s_2}|$ and so on, until one compares $max|T_{s_m}|$ with t_{s_m}.

The parameters for this function are defined as follows:

- abs.Zs.H0.1row = 1 row of tests statistics under the complete null.
- abs.Zs.H1.1samp = raw test statistics for 1 sample.
- oo = the ordering of the raw test statistics for 1 sample.

```
comp.rawt.SD <- function(abs.Zs.H0.1row, abs.Zs.H1.1samp, oo) {
 M<-length(abs.Zs.H0.1row)
 maxt <- rep(NA, M)
 nullt.oo<-abs.Zs.H0.1row[oo]
 rawt.oo<-abs.Zs.H1.1samp[oo]
 maxt[1] <- max(nullt.oo) > rawt.oo[1]
 for (h in 2:M) {maxt[h] <- max(nullt.oo[-(1:(h-1))]) > rawt.oo[h]}
 return(as.integer(maxt))
}
```

The function **adjust.allsamps.WYSD** is also needed to implement the Westfall-Young step-down MTP. It carries out the above function for all rows in the matrix of test statistics under the complete null and does so for all samples of raw test statistics under the alternative hypothesis.

The parameters for this function are defined as follows:

- snum = the number of samples for which test statistics under the alternative hypothesis are compared with the distribution (matrix) of test statistics under the complete null.
- abs.Zs.H0 = a matrix of test statistics under the complete null.
- abs.Zs.H1 = a matrix of raw test statistics under the alternative.
- order.matrix = a matrix in which each row corresponds to the order of the test statistics for 1 sample.

```
adjust.allsamps.WYSD<-function(snum,abs.Zs.H0,abs.Zs.H1,order.matrix) {
 cl <- makeCluster(ncl)
 registerDoParallel(cl)
 clusterExport(cl=cl, list("comp.rawt.SD"))
 M<-ncol(abs.Zs.H0)
 adjp.WY<-matrix(NA,snum,M)

 doWY <- foreach(s=1:snum, .combine=rbind) %dopar% {
   ind.B<-t(apply(abs.Zs.H0, 1, comp.rawt.SD, abs.Zs.H1.1samp=abs.Zs.H1[s,],
oo=order.matrix[s,]))
   pi.p.m <- colMeans(ind.B)
 # enforcing monotonicity
   adjp.minp <- numeric(M)
   adjp.minp[1] <- pi.p.m[1]
   for (h in 2:M) {adjp.minp[h] <- max(pi.p.m[h], adjp.minp[h-1])}
   adjp.WY[s,] <- adjp.minp[order.matrix[s,]]
 }
 return(doWY)
 stopCluster(cl)
}
```

The function **t.mean.H1** computes the means of the test statistics under the joint alternative hypothesis. Recall that $t(m)$ has a t-distribution with mean $MDES(m)/Q(m)$, where

$$Q(m) = \sqrt{\frac{(1-R^2(m))}{p(1-p)Jn_j}}.$$

The parameters for this function are defined as follows:

- MDES = a vector of length M corresponding to the minimum detectable effect sizes (MDESs) for the M outcomes.
- J = the number of blocks.
- n.j = the harmonic mean of the number of units per block.
- R2 = a vector of length M corresponding to the R^2's for the M outcomes.
- Tbar = the proportion of units assigned to treatment within each block.

```
t.mean.H1<-function(MDES,J,n.j,R2,Tbar) { MDES * sqrt(Tbar*(1-Tbar)*J*n.j) / sqrt(1-R2) }
```

Finally, the function **est.power.mult** estimates power for all definitions and for all MTPs in this paper. The parameters for this function are defined as follows:

- M = the number of hypothesis tests (number of outcomes).

- MDES = a vector of length M corresponding to the MDESs for the M outcomes.
- Tbar = the proportion of units assigned to treatment within each block.
- alpha = the familywise error rate (FWER)
- J = the number of blocks.
- n.j = the harmonic mean of the number of units per block.
- numcovar = the number of baseline covariates, not including block dummies.
- R2 = a vector of length M corresponding to the R^2's for the M outcomes.
- sigma = the MxM correlation matrix for the test statistics.
- tnum = the number of test statistics to generate, and also the number of permutations for WY. It has a default of 10,000.
- snum = the number of samples used to estimate WY. It has a default of 1,000.
- ncl = the number of clusters to use for parallel processing. It has a default of 2.

```
est.power.mult<-
function(M,MDES,Tbar,alpha,J,n.j,numcovar,R2,sigma,tnum=10000,snum=1000,ncl=2) {

 require(MASS)
 require(mvtnorm)
 require(multtest)
 require(doParallel)

 # getting the mean of the test statistics under joint alternative hypothesis
 shift.t<-t.mean.H1(MDES,J,n.j,R2,Tbar)

 # creating a matrix of the means repeated in every row
 shift.t.mat<-t(matrix(rep(shift.t,tnum),M,tnum))

 # computing the degrees of freedom
 t.df<- J*n.j - J - numcovar - 1
```

```
# generating test statistics under joint null and alternative hypotheses, and getting absolute values
Zs.H0<-rmvt(tnum, sigma = sigma, df = t.df, delta = rep(0,M),type = c("shifted", "Kshirsagar"))
Zs.H1 <- Zs.H0 + shift.t.mat
abs.Zs.H0 <- abs(Zs.H0)
abs.Zs.H1 <- abs(Zs.H1)

# converting test statistics to p-values
pvals.H0<-2*pt(-abs(Zs.H0),df=t.df)
pvals.H1<-2*pt(-abs(Zs.H1),df=t.df)

# using mt.rawp2adjp function in multtest package to adust p-values for BF-SS, HO-SD and BH-SU
adjp<-apply(pvals.H1,1,mt.rawp2adjp,proc=c("Bonferroni","Holm","BH"),alpha=alpha)

# grabbing p-values from information returned in object adjp
grab.pval<-function(...,proc) {return(...$adjp[order(...$index),proc])}
rawp<-do.call(rbind,lapply(adjp,grab.pval,proc="rawp"))
adjp.BF<-do.call(rbind,lapply(adjp,grab.pval,proc="Bonferroni"))
adjp.HO<-do.call(rbind,lapply(adjp,grab.pval,proc="Holm"))
adjp.BH<-do.call(rbind,lapply(adjp,grab.pval,proc="BH"))

# each row of oo.all is the order of each row of absolute value test statistics
order.matrix<-t(apply(abs.Zs.H1,1,order,decreasing=TRUE))

# using functions above to adjust p-values with WY-SS and WY-SD
adjp.SS<-adjust.allsamps.WYSS(snum,abs.Zs.H0,abs.Zs.H1)
adjp.WY<-adjust.allsamps.WYSD(snum,abs.Zs.H0,abs.Zs.H1,order.matrix)

# creating list of raw p-values and all adjusted p-values
adjp.all<-list(rawp,adjp.BF,adjp.HO,adjp.BH,adjp.SS,adjp.WY)

# for each matrix in adjp.all, for all entries, determining if null is rejected
reject<-function(x) {as.matrix(1*(x<alpha))}
reject.all<-lapply(adjp.all,reject)

# determining how many rejections among tests that are truly false
gt.alpha<-function(x) {apply(as.matrix(x[,MDES>0]),1,sum)}
```

```
 gt.allpha.all<-lapply(reject.all,gt.alpha)

 # computing individual power
  power.ind.fun<-function(x) {apply(x,2,mean)}
  power.ind.all<-lapply(reject.all,power.ind.fun)
  power.ind.all.mat<-do.call(rbind,power.ind.all)

 # d-mininmal powers powers for all MTPs (including complete power when m=M)
 power.min.fun <- function(x,M) {
   power.min<-numeric(M)
   cnt<-0
   for (m in 1:M) {
     power.min[m]<-mean(x>cnt)
     cnt<-cnt+1
   }
   return(power.min)
   }
 power.min<-lapply(gt.allpha.all,power.min.fun,M=M)
 power.min.mat<-do.call(rbind,power.min)

 # for complete power, grab results for raw p-values
 power.cmp<-rep(power.min.mat[1,M],length(power.min))

 # put everything together and label
 all.power.results<-cbind(power.ind.all.mat,power.min.mat[,-M],power.cmp)
 mean.ind.power <- apply(as.matrix(all.power.results[,1:M][,MDES>0]),1,mean)
 all.power.results<-cbind(mean.ind.power,all.power.results)
 colnames(all.power.results)<-c("avg indiv",paste0("indiv",1:M),paste0("min",1:(M-
1)),"complete")
 rownames(all.power.results)<-c("rawp","BF","HO","BH","WY-SS","WY-SD")
 return(all.power.results)
}
```

Here is an example.

```
ncl<-2
M<-3
sigma<-matrix(rep(0.5,M*M),nrow=M,ncol=M)
diag(sigma)<-1
MDES<-c(rep(0.125,M))
```

```
est.power.mult(MDES=MDES,M=M,Tbar=0.5,alpha=0.05,J=20,n.j=50,numcovar=1,R2=0.5,sigma=sigma,tnum=10000,snum=3,ncl=ncl)
```

```
##             avg indiv indiv1 indiv2 indiv3   min1   min2 complete
## rawp  0.7971000 0.7976 0.7985 0.7952 0.9463 0.8370    0.608
## BF    0.6577333 0.6572 0.6587 0.6573 0.8699 0.6842    0.608
## HO    0.7304333 0.7312 0.7325 0.7276 0.8699 0.7346    0.608
## BH    0.7601667 0.7615 0.7619 0.7571 0.8836 0.7889    0.608
## WY-SS 0.6450000 0.6340 0.6600 0.6410 0.8640 0.6620    0.608
## WY-SD 0.7200000 0.7180 0.7210 0.7210 0.8640 0.7210    0.608
```

Appendix C

Validation Results

Table C.1

Comparing Power Estimates with Those Obtained by PowerUp!: One Outcome with MDES = 0.125, 20 Blocks of Size 100, and Varying R^2

	PowerUp!	Paper
Power	Estimates	Estimates
Level 1 R^2		
0	0.287	0.285
0.2	0.346	0.344
0.8	0.878	0.877

SOURCE: PowerUp! estimates were generated using PowerUp! (Dong, 2013, Table RBD2-c).

Table C.2

Comparing Power Estimates Obtained Using Power Estimation Methodology in Section 3 with Power Estimates in Schochet (2008) for One Site with 2,000 Individuals, Correlations 0, MDES = 0.125, and Varying Numbers of Tests and Proportions of Tests That Are False

	Schochet (2008)				Method in Section 3			
	MTP Used				MTP Used			
Tests and true null hypotheses	None	BF	HO	BH	None	BF	HO	BH
80% of tests with true null								
Number of tests								
5	0.80	0.59	0.59	0.55	0.80	0.59	0.59	0.58
10	0.80	0.50	0.50	0.55	0.80	0.49	0.50	0.55
20	0.80	0.41	0.42	0.55	0.80	0.41	0.42	0.52
50% of tests with true null								
Number of tests								
5	0.80	0.59	0.61	0.67	0.80	0.59	0.61	0.66
10	0.80	0.50	0.53	0.67	0.80	0.49	0.53	0.67
20	0.80	0.41	0.44	0.67	0.80	0.41	0.44	0.66
20% of tests with true null								
Number of tests								
5	0.80	0.59	0.66	0.74	0.80	0.59	0.66	0.74
10	0.80	0.50	0.57	0.74	0.80	0.49	0.56	0.74
20	0.80	0.41	0.47	0.74	0.80	0.41	0.46	0.74

SOURCE: Schochet (2008, Table B.4).

Table C.3

Comparing Power Estimates Obtained Using Power Estimation Methodology in Section 3 with Power Estimates Obtained by Monte Carlo Simulation for Six Tests Each with MDES = 0.125, 20 Sites Each with 100 Individuals, and 2,000 Samples Each with 10,000 Permutations

	Individual		**1-minimal**		**1/3-minimal**		**2/3-minimal**		**Complete**	
MTP Used and Correlations	Section 3 Method	Simulation	Section 3 Method	Simulation	Section 3 Method	Simulation	Section 3 Method	Simulation	Section 3 Method	Simulation
No adjustment										
Correlations between tests										
0	0.798	0.796	1.000	1.000	0.999	0.998	0.899	0.899	0.260	0.256
0.2	0.798	0.800	0.998	0.999	0.985	0.987	0.849	0.855	0.349	0.341
0.5	0.798	0.795	0.982	0.982	0.946	0.944	0.809	0.805	0.471	0.487
0.8	0.798	0.796	0.934	0.932	0.889	0.889	0.792	0.783	0.613	0.615
Bonferroni										
Correlations between tests										
0	0.561	0.556	0.992	0.995	0.939	0.932	0.468	0.450	0.260	0.256
0.2	0.561	0.563	0.966	0.968	0.868	0.873	0.483	0.492	0.349	0.341
0.5	0.561	0.557	0.896	0.882	0.775	0.775	0.505	0.509	0.471	0.487
0.8	0.561	0.558	0.780	0.773	0.684	0.684	0.527	0.527	0.613	0.615
Holm										
Correlations between tests										
0	0.679	0.674	0.992	0.995	0.952	0.944	0.651	0.642	0.260	0.256
0.2	0.672	0.675	0.966	0.968	0.888	0.888	0.627	0.643	0.349	0.341
0.5	0.663	0.662	0.896	0.882	0.797	0.797	0.619	0.617	0.471	0.487
0.8	0.652	0.647	0.780	0.773	0.706	0.705	0.620	0.614	0.613	0.615

(continued)

Table C.3 (continued)

	Individual		1-minimal		1/3-minimal		2/3-minimal		Complete	
MTP Used and	Section 3		Section 3		Section 3		Section 3		Section 3	
Correlations	Method	Simulation	Method	Simulation	Method	Simulation	Method	Simulation	Method	Simulation
Benjamini-Hochberg										
Correlations between tests										
0	0.769	0.767	0.996	0.998	0.984	0.984	0.833	0.828	0.260	0.256
0.2	0.758	0.763	0.975	0.978	0.941	0.949	0.783	0.796	0.349	0.341
0.5	0.745	0.741	0.913	0.902	0.869	0.866	0.752	0.745	0.471	0.487
0.8	0.739	0.731	0.816	0.805	0.792	0.779	0.741	0.736	0.613	0.615
Westfall-Young										
Correlations between tests										
0	0.684	0.676	0.992	0.995	0.953	0.945	0.667	0.646	0.260	0.256
0.2	0.670	0.680	0.960	0.971	0.892	0.892	0.643	0.649	0.349	0.341
0.5	0.674	0.679	0.905	0.894	0.820	0.821	0.632	0.635	0.471	0.487
0.8	0.687	0.704	0.832	0.837	0.759	0.768	0.657	0.680	0.613	0.615

References

Bang, H., Jung, S., & George, S. L. (2005). Sample size calculations for simulation-based multiple-testing procedures. *Journal of Biopharmaceutical Statistics, 15*, 957-967.

Benjamini, Y., & 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.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29*, 1165-1188.

Beran, R. (1988). Pre-pivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association, 83*, 679-686.

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review, 19*, 547-556. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ514281

Bloom, H. S. (2006). *The core analytics of randomized experiments for social research. MDRC working papers on research methodology*. Retrieved from http://www.mdrc.org/sites/default/files/full_533.pdf

Chen, J., Luo, J., Liu, K., & Mehrotra, D. (2011). On power and sample size computation for multiple testing procedures. *Computational Statistics and Data Analysis, 55*, 110-122.

Dong, N., & Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*, 24-67. doi:10.1080/19345747.2012.673143

Dudoit, S., Shaffer, J. P., & Bodrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science, 18*, 71-103.

Dunn, O. J. (1959). Estimation of the medians for dependent variables. *Annals of Mathematical Statistics, 30*, 192-197. doi:10.1214/aoms/1177706374

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*(293), 52-64. doi:10.1080/01621459.1961.10482090

Ewens, W., & Grant, G. (2005). *Statistical methods in bioinformatics: An introduction*. New York: Springer.

Ge, Y., Dudoit, S., & Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test, 12*, 1-77.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology, 48*, 1711-1725.

Hedges, L. V., & Rhoads, C. (2010). *Statistical power analysis in education research. NCSER 2010-3006*. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED509387

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*(2), 65-70.

Koch, G. G., & Gansky, M. S. (1996). Statistical considerations for multiplicity in confirmatory protocols. *Drug Information Journal, 30*, 523-533.

Neyman, J. (1923). Statistical problems in agricultural experiments. *Journal of the Royal Statistical Society, 2*, 107-180.

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X., Martinez, A., Bloom, H., & Hill, C. (2011). *Optimal design plus empirical evidence (version 3.0)*. New York: William T. Grant Foundation. Retrieved from http://wtgrantfoundation.org/resource/optimal-design-with-empirical-information-od

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688-701. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ118470

Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. *Statistical Science, 21*, 299-309.

Schochet, P. Z. (2005). *Statistical power for random assignment evaluations of education programs*. Princeton, NJ: Mathematica Policy Research, Inc. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED489855

Schochet, P. Z. (2008). *Guidelines for multiple testing in impact evaluations of educational interventions. Final report*. Princeton, NJ: Mathematica Policy Research, Inc. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED502199

Senn, S., & Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics, 6*, 161-170. doi:10.1002/pst.301

Shaffer, J.P. (1995). Multiple hypothesis testing. *Annual Review of Psychology, 46*, 561-584. doi:10.1146/annurev.ps.46.020195.003021

Spybrook, J., Bloom, H. S., Congdon, R., Hill, C. J., Martinez, A., & Raudenbush, S. W. (2011). *Optimal design plus empirical evidence: Documentation for the "optimal design" software version 3.0.* New York: William T. Grant Foundation. Retrieved from http://wtgrantfoundation.org/resource/optimal-design-with-empirical-information-od

Tukey, J.W. (1953). The problem of multiple comparisons. Mimeographed notes. Princeton, NJ: Princeton University.

U.S. Department of Education. (2014). *What Works Clearinghouse procedures and standards handbook version 3.0*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. Retrieved from: http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

Westfall, P. H., Tobias, R. D., & Wolfinger, R. D. (2011). *Multiple comparisons and multiple tests using SAS, second edition*. Cary, NC: The SAS Institute.

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

About MDRC

MDRC is a nonprofit, nonpartisan social and education policy research organization dedicated to learning what works to improve the well-being of low-income people. Through its research and the active communication of its findings, MDRC seeks to enhance the effectiveness of social and education policies and programs.

Founded in 1974 and located in New York City and Oakland, California, MDRC is best known for mounting rigorous, large-scale, real-world tests of new and existing policies and programs. Its projects are a mix of demonstrations (field tests of promising new program approaches) and evaluations of ongoing government and community initiatives. MDRC's staff bring an unusual combination of research and organizational experience to their work, providing expertise on the latest in qualitative and quantitative methods and on program design, development, implementation, and management. MDRC seeks to learn not just whether a program is effective but also how and why the program's effects occur. In addition, it tries to place each project's findings in the broader context of related research — in order to build knowledge about what works across the social and education policy fields. MDRC's findings, lessons, and best practices are proactively shared with a broad audience in the policy and practitioner community as well as with the general public and the media.

Over the years, MDRC has brought its unique approach to an ever-growing range of policy areas and target populations. Once known primarily for evaluations of state welfare-to-work programs, today MDRC is also studying public school reforms, employment programs for ex-offenders and people with disabilities, and programs to help low-income students succeed in college. MDRC's projects are organized into five areas:

- Promoting Family Well-Being and Children's Development
- Improving Public Education
- Raising Academic Achievement and Persistence in College
- Supporting Low-Wage Workers and Communities
- Overcoming Barriers to Employment

Working in almost every state, all of the nation's largest cities, and Canada and the United Kingdom, MDRC conducts its projects in partnership with national, state, and local governments, public school systems, community organizations, and numerous private philanthropies.