



Statistical Power in Evaluations That Investigate Effects on Multiple Outcomes: A Guide for Researchers

Kristin E. Porter

To cite this article: Kristin E. Porter (2017): Statistical Power in Evaluations That Investigate Effects on Multiple Outcomes: A Guide for Researchers, Journal of Research on Educational Effectiveness, DOI: [10.1080/19345747.2017.1342887](https://doi.org/10.1080/19345747.2017.1342887)

To link to this article: <http://dx.doi.org/10.1080/19345747.2017.1342887>

 View supplementary material [↗](#)

 Accepted author version posted online: 19 Jun 2017.

 Submit your article to this journal [↗](#)

 Article views: 1

 View related articles [↗](#)

 View Crossmark data [↗](#)

**Statistical Power in Evaluations That Investigate Effects on Multiple Outcomes:
A Guide for Researchers**

Kristin E. Porter

MDRC, Oakland, California, USA

CONTACT Kristin E. Porter kristin.porter@mdrc.org MDRC, 475 14th Street, Oakland, CA 94612,
USA

Abstract

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. This paper presents methods for estimating statistical power, for multiple definitions of statistical power and presents empirical findings on how power is affected by the use of MTPs.

Keywords

statistical power, multiple hypothesis testing, multiple testing procedures

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 constrained to an acceptable level, regardless of the number of tests.

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). The use of MTPs is also gaining more attention due to recent press about a so-called reproducibility, or replication, crisis in many areas of research, for which the testing of multiple hypotheses plays a role.¹

¹ For example, Christensen & Miguel (2016) provide evidence of replication problems in economics research and discuss the contributing factors, which include multiple hypotheses testing.

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; Bretz, Hothorn, & Westfall, 2011; Dudoit, Shaffer, & Boldrick, 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 $\frac{1}{2}$ -minimal power: the probability of detecting effects of at least a particular size on at least $\frac{1}{2}$ 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. It also affects the overall extent of power.

Education researchers are likely most inclined to focus on individual power because it matches current practice and because they often are interested in effects on each of multiple outcomes (or for each of multiple subgroups, points in time or treatment groups). However, other definitions of power (discussed often in medical research literature) may be important to consider in some cases, either in place of or in addition to individual power. Imagine, for example, a pilot or “proof on concept” phase of testing a new education program. Perhaps this program would be deemed successful enough for modifications and replication if it has a statistically significant impact (of at least a specified size) on at least one of a few primary outcomes. In this case, the researchers may want to focus on 1-minimal power. On the other hand, consider a randomized control trial (RCT) that will be used to determine whether an expensive education program will scaled up. Funders of the program may require evidence of statistically significant impacts (of at least a specified size) on all primary outcomes of interest in order to scale up the program. In this case, complete power would be important to estimate.

This paper does not advocate that decisions about programs should directly hinge on p-values. Rather, it reflects that decisions often do in practice, and it suggests that a focus on individual power may be insufficient for the objectives of some studies. At the same time, even if individual power is a preferred focus, it can be useful to compute and present other power definitions as well. An underpowered study, with respect to individual power may prove to have a high probability of detecting at least one true impact. A sufficiently powered study with respect to individual power may have a low probability to detect all true impacts. This is valuable information to share.

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.

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 '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.² The paper also focuses on a single, simple research design and analysis plan that education studies often use in practice: a multisite, 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

²Note 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.

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, beginning with a motivating example. The section 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 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 in the online supplemental materials. R code implementing the power estimation methodology can be found in Appendix B in the online supplemental materials. Also, power comparisons with other sources that validate the accuracy of the power estimation methodology can be found in Appendix C in the online supplemental materials.

2. Overview of Multiple Testing

2.1 A Motivating Example from Education

This section begins with a realistic example to which the paper will refer throughout in order to illustrate various concepts. Suppose that researchers are designing a multisite, blocked trial in which they plan to investigate the effects of a mentoring program on three confirmatory outcomes related to social and emotional development --- measures of social competence, emotional

competence and self-regulation.³ For each outcome, an impact with an effect size of at least 0.125 standard deviations would be policy relevant.

In their study, each school is a site -- and a block for randomization. They successfully recruit 20 schools, and within each school they randomly assign 50 students to the program and control groups in equal proportion.⁴ That is, one half of the students in each school are assigned to the mentoring program and one half are assigned to “business as usual.” The researchers estimate the statistical power to detect effects on each outcome given their sample size, desired effect size of 0.125 and assumptions about their estimation model. They plan to estimate effects using a model with block-specific intercepts (for each of the sites or schools) and with the assumption of constant effects across blocks (sites or schools). By including baseline measures of the outcome and school intercepts in their model, they assume that the explanatory power of baseline measures, the R^2 , is 0.5 for all three impact models. The researchers also specify a statistical significance level (discussed further below) of 0.05. They estimate that their statistical power, the same for each outcome in this example, is 80%.

The researchers realize, however, that they have a multiple testing problem, and they want to address the problem and understand the consequences for statistical power before finalizing their study design and analysis plan. They want to focus on individual power because stakeholders will

³ Because these outcomes are within a single domain of social and emotional development, it is assumed that the What Works Clearinghouse would apply multiplicity corrections to the findings.

⁴ Note that multisite RCTs tend to be at least this large. For example, Weiss et. al. (forthcoming), which summarizes data from 15 multisite RCTs of educational and training programs, report that these RCTs, the number of sites ranges from 9 to 300, the average number of individuals per site ranges from 11 to hundred, and the total number of individuals ranges from 3000 to 100,000.

want to understand the effectiveness of the intervention with respect to each of the three primary outcomes. However, they also know that the funder of the program would consider the mentoring program a success and will likely continue supporting the program if it leads to improvements on at least one of the outcomes. Therefore, they also want to estimate 1-minimal power. The remainder of this paper helps the researchers understand choices for addressing their multiple testing problem and the consequences for statistical power, for all possible definitions.

2.2 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.⁵ In the above motivating example, $M = 3$. Then the population average treatment on the m^{th} outcome, given by

$$\psi(m) = E(Y1_i(m)) - E(Y0_i(m)), \quad (1)$$

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

⁵ While this paper focuses on multiplicity of treatment effects estimated in RCTs, the lessons also apply to other analyses that rely on statistical significance.

$$ES(m) = \frac{\psi(m)}{\sigma_Y(m)}, \quad (2)$$

where $\sigma_Y(m)$ is the standard deviation of the m^{th} outcome.⁶

In the frequentist framework, one typically tests a null hypothesis of no effect, $H_0(m): ES(m) = 0$, against an alternative hypothesis of $H_1(m): ES(m) \neq 0$ for a two-sided test or $H_1(m): ES(m) > 0$ or $H_1(m): ES(m) < 0$ for a one-sided test. For the purposes of computing power researchers specify an alternative hypothesis of at least a particular effect size (ES). In the above example, the researchers specified an ES of 0.125. A significance test, such as a two-sided or one-sided t -test, is then conducted, and one obtains a test statistic given by

$$t(m) = \frac{ES(m)}{SE(ES(m))}, \quad (3)$$

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

$$p(m) = 2 * \Pr\{T(m) \geq |t(m)|\} \quad p(m) = \Pr\{T(m) \geq t(m)\} \quad p(m) = \Pr\{T(m) \leq t(m)\} \quad \text{This}$$

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.

⁶It is assumed here that the standard deviation is the same in both counterfactual settings.

When testing a *single* hypothesis under this framework (such effects are being assessed on just one outcome, so 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 $1 - Pr(\text{none of the null hypotheses will be erroneously rejected}) = 1 - (1 - \alpha)^M$. Therefore, in the above motivating example in which the researchers are estimating effects on three outcomes, if these outcomes are assumed independent, the probability of at least one false positive finding is 14%. If the researchers were instead estimating effects on five independent outcomes, the probability of at least one 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 counteract the multiple testing problem, MTPs adjust p-values upward.⁷ 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.

⁷Alternatively, 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.

So far, we see that if the researchers in the motivating example ignore their multiplicity of hypothesis tests, the probability that they will have at least one false positive finding is greater than 5% and possibly (although unlikely) as high as 14% if the three test statistics are independent of one another. (This will be discussed further below.) Therefore, the researchers realize they need to adjust their p-values upwards. They are now concerned about the implications for their statistical power. The next two sections explain some choices of MTPs for p-value adjustment and the implications for all definitions of power.

2.3 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.

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. M_0 and M_1 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 $E\left(\frac{B}{R}\right)$ 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 effect size), 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.⁸

2.4 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

⁸It 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).

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 MTPs 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 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.⁹

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

⁹The 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.

the distribution of test statistics under the complete null hypothesis.¹⁰ 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.

How should researchers in the hypothetical example take MTPs into account when designing their study? Since a single false positive finding could lead to the wrong conclusion about the mentoring program's effectiveness, they decide that controlling the FWER is preferable to controlling the FDR since the FDR is more lenient with false positives. Among the MTPs discussed that control the FWER, they think they want to choose the one that preserves the most power. They expect the test statistics associated with the impact estimates for their three outcomes to be correlated. In fact, based on prior research, they assume that the correlation between all pairs of their outcomes is 0.5. Therefore, they think it might be worthwhile to use the Westfall-Young MTP, which takes these correlations into account. But what are the power implications for their study? And are they correct that the Westfall-Young MTP is optimal and worth the extra trouble given that it can be more complicated to implement than other MTPs? In particular, they wonder if the Holm MTP, which is much simpler to implement but preserves more power than Bonferroni since it is a stepwise procedure, may suffice. The next section describes how they can estimate their statistical power for all MTPs.

¹⁰Instead of using test statistics, the Westfall-Young MTPs can alternatively compare raw p-values with the estimated joint null distribution of p-values.

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 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.

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:

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

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 $\sigma^2(m)$.¹¹

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

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

where $\sigma_Y^2(m)$ is the pooled outcome variance of the m^{th} outcome;¹² $R^2(m)$ 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

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

¹¹ The assumption of normally distributed residuals is not needed to estimate impacts. Without normality, all the MTPs except Westfall-Young control the FWER or FDR asymptotically. The Westfall-Young MTP guarantees strong control of the FWER under non-normality (Westfall & Troendle, 2008).

¹²Here it is assumed that the variance of the outcome is the same in both the treatment and control groups.

For convenience, let $Q(m) \equiv SE(ES(m))$. 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

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

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 effect size 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* effect size (ES), then $t(m)$ has a t -distribution with mean $ES(m)/Q(m)$, 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 H_0) that there are not effects on any of the outcomes. The set of two-sided alternative

hypotheses focused on a specified effect size (referred to as $H1$), is $|ES(m)| \geq ES(m)$ for $m=1, \dots, M$, where the MDES may vary for each outcome.

Under the complete null hypothesis, $H0$, the set of test statistics for all M hypothesis tests, which can be written collectively as $t0$, have a multivariate t -distribution with means of zero, degrees of freedom equal to the vector df , and correlation matrix ρ . Under the set of specific alternative hypotheses, $H1$, the set of test statistics, which can be written collectively as $t1$, have the same multivariate t -distribution --- except that the means are equal to the vector ES/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 ES 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 $t0$ and $t1$ can be generated (i.e., simulated) with statistical software. That is, one can generate a large number of test statistics under $H0$ and under $H1$, 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 $t0$ and

$t1$. Additionally, $t0$ and $t1$ can be converted to $10K \times M$ matrices of p-values, $p0$ and $p1$.

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

For example, since each row of $p1$ 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 ESs specified under $H1$, 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:

$$\tilde{p}^{BF-SS}, \tilde{p}^{HO-SD}, \text{ or } \tilde{p}^{BH-SU}.$$

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 $t0$ and $t1$ are used to obtain adjusted p-values. That is, to adjust p-values for one data sample, one row of $t1$ is compared with all rows in $t0$.

For each MTP, the resulting $10K \times M$ 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 $10K \times M$ 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).¹³
- d -minimal power is the proportion of the 10K rows in which at least d of the M null hypotheses were rejected.¹⁴
- 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.¹⁵
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 hypothesis 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).¹⁶

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

¹³ Individual power may also be referred to as “marginal power” (e.g., Senn & Bretz, 2007). One can also focus on “average power,” the mean individual power of all false null hypotheses (e.g., Dudoit, Shaffer, & Boldrick, 2003; Bretz, Hothorn & Westfall, 2011).

¹⁴ Note that others refer to 1-minimal power simply as “minimal power” (e.g., Maurer & Lellein, 1988; Chen, Luo, Liu, & Mehrotra, 2011; Westfall, Tobias, & Wolfinger, 2011), “disjunctive power” (e.g., Bretz, Hothorn, & Westfall, 2011), or “any pair” power (Ramsey, 1978). Chen, Luo, Liu, & Mehrotra (2011) use the terminology of “r-power” for what is referred to here as d-minimal power for $d > 1$.

¹⁵ Complete power has also been referred to as “conjunctive power” (Bretz, Hothorn, & Westfall, 2011) and “all pairs power (Ramsey, 1978).”

¹⁶ Complete power does not in itself require unadjusted tests. The above approach for not adjusting tests assumes that all tests must to be statistically significant in order to claim impacts on all outcomes.

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.¹⁷

Note that this approach of simulating test statistics builds on work by Bang, Young, & 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

¹⁷When the power estimation methodology is coded in R (as shown in online 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.

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.¹⁸ 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 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 Estimating Power in the Motivating Example

Recall that in the motivating example, researchers are planning a multisite trial to investigate the effects of a mentoring program on three confirmatory outcomes --- measures of social competence, emotional competence and self-regulation. They have recruited 20 schools (the sites, or blocks) and randomly assigned 50 students within each school to either the program or control group (50% to each group). They plan to use the model specified in (4) to estimate effects, and assume an R^2 of 0.5 for all three impact models. Based on prior research, they assume that the correlation between all pairs of their outcome measures and that the correlation between all pairs of test statistics is 0.5. (Further discussion of this assumptions is provided below).

¹⁸ For a discussion of using simulation for power calculations, see Westfall et. al. (2008) and Arnold et. al. (2011). Also, see Westfall et. al. (2011) for a discussion of the SAS macro %simpower, which uses Monte Carlo simulation to estimate statistical power when testing the equivalency of a series of unadjusted means when adjusting p-values using a MTPs not discussed in this paper.

Their desired ES for each outcome is at least 0.125. They are interested in estimating individual power - because they want to understand impacts on each particular outcome -- as well as 1-minimal power because the funder would consider the mentoring program a success if there is an impact on at least one outcome.

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. To illustrate how the power computation works, this example focuses on estimation of 1-minimal power when using the Westfall-Young MTP. Further calculations are presented in Section 4, which will provide information about all the power estimation goals of the researchers in the motivating example.

First, the researchers generate $\mathbf{t1}$. Therefore, they simulate a 10K-row x 3-column matrix of test statistics following a multivariate t -distribution with correlation matrix

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

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

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 ($\mathbf{p1}$) 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 Westfall-Young 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 88% 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 88% chance of a statistically significant effect estimate for at least one of them. This is better than the typical 80% standard. With 1-minimal power set at 80%, the researchers' minimum detectable effect size (MDES) - the smallest true effect size that their study can detect with statistical significance¹⁹ - is smaller than the ES of 0.125; it is 0.111. Alternatively, they can include 16 sites with 50 individuals instead of 20 sites with 50 individuals to achieve at least 80% power for an MDES of 0.125.

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

¹⁹ For a discussion of minimum detectable effects (MDEs), which are expressed in outcomes' units, and MDEs, which are expressed in standard deviation units, see, for example, Bloom (1995), Schochet (2005), and Bloom (2006).

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;²⁰ the proportion of individuals assigned to the treatment group; the explanatory power of baseline covariates, including block indicators (R^2); and an ES. In the above example, the researchers assumed the same R^2 and the same ES 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 '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 's across the impact models reduces the correlations between the residuals and therefore between test statistics.²¹ 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 '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

²⁰When 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).

²¹For 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.

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 ESs, 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.

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 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 ES 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 ESs); 1-minimal power, 1/3-minimal power, and 2/3-

minimal power (across all outcomes with true effects as large as the specified ESs, 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 ESs 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 ESs.* 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 ES 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. (This is seen by the lighter shaded points of each color corresponding to an MTP.) 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. (This is seen with the darkest purple point under 12 tests in plot (a).) While power drops off considerably when there are truly effects on just 2/3 or 1/3 of the outcomes (the medium and light purple points), 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 (seen by comparing the medium shaded purple point to medium shaded points of colors for other stepwise MTPs, and to the only point for non-stepwise MTPs, and making similar comparisons with the light shaded points).

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-SD and WY-SD) 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 (as shown in plot (a)).

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 ESs. 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. This is seen in the darkest shaded points in the first set of vertical lines for all MTPs. As the correlation increases (moving from left to right across the upper X axis), 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 points, 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, as seen by the light-shaded points, that there is a substantial loss of power compared with the benchmark.

A comparison of Figure 4 (focused on six tests) with Figure 3 (focused on three tests) 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. As shown in Figure 4, 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. This corresponds to the assumptions in our motivating example. If the researchers in this example needed to detect effects on all 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 ESs on all outcomes of 0.148 instead of 0.125.

4.4 Implications for the Motivating Example

Recall that the researchers in the motivating example wanted to ensure that they have sufficient 1-minimal power. They also wanted to maximize individual power. They were leaning towards using the Westfall-Young MTP because it takes the correlation between their test statistics (assumed to be 0.5 between all pairs) into account, but they wondered if the Holm MTP, which is far simpler to implement would result in sufficient power.

If they use the Westfall-Young MTP, the probability to detect at least one effect with statistical significance at the 0.05 level is 88%, as was shown in Section 3.2 above. This assumes that there are effects of at least 0.125 standard deviations on all three outcomes. If there are actually only effects on two of the three outcomes, the probability to detect at least one of them is 82%, and if there is actually only an effect on one of the outcomes, the probability of detecting it is 66%. If they use the

Holm MTP instead, their levels of minimal power are almost the same as when using the Westfall-Young MTP. The levels are 87%, 81% and 66%, depending on the number of true effects.

If they use the Westfall-Young MTP, their individual power -- the probability of detecting a statistically significant effect size of at least 0.125 on each of their outcomes (which in the motivating example is the same for all outcomes) is 76%, 71% or 66%, respectively, depending on whether there are truly effects on all three, just two or just one of the outcomes. If they use the Holm MTP instead, their individual power drops somewhat - to 73%, 68% or 66%, respectively.

The implications for the researchers' final design (the number of sites they recruit or perhaps an adjustment to the number of primary outcomes of interest) and for their analysis plan (which MTP to use) depend on (1) their level of confidence in whether there will actually be effects on their primary outcomes of interest; and (2) how much they weigh a focus on individual power against a focus on 1-minimal power. (For illustration purposes, this is assuming they have complete confidence in their assumptions about the correlation between test statistics and the R^2 's). If there are indeed impacts on all three outcomes, they have a high probability of detecting effects on at least one of them, satisfying the funder's priorities. At the same time, while their power to detect effects on each particular outcome is less than the 80% norm, it is still pretty good (at 76% with Westfall-Young and 73% with Holm). They will be better off with Westfall-Young, but not by much. To be most conservative, assuming an impact on just one of their three outcomes, the researchers may want to increase their sample to include 27 sites. In this case, they have an 80% chance of detecting the single effect of at least 0.125 standard deviations (and 1-minimal and individual power are identical). If

they cannot add more sites than 20, they would have to settle for a minimum detectable effect size of 0.145 instead of 0.125 to achieve 80% power (using either the Westfall-Young MTP or Holm MTP).

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 general 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, which is standard practice in education, 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 ESs 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 '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 's of Baseline Covariates

Finally, while it is well known that higher R^2 's are associated with greater power, it tends also to be the case that higher R^2 '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 ESs.

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 ESs) 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 $Q(m) \equiv SE(ES(m))$, 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 ESs. 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.

Funding

This work was supported by the Institute of Education Sciences, U.S. Department of Education [Grant number R305D140024].

References

- Arnold, B. F., Hogan, D. R., Colford J. M. Jr. & Hubbard, A. E. (2011). Simulation methods to estimate design power: an overview for applied research. *BMC Medical Research Methodology*, *11*:94. <http://www.biomedcentral.com/1471-2288/11/94>
- 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

- Bretz, F., Hothorn, T., and Westfall, P. (2011). *Multiple comparisons using R*. Boca Raton, FL: Chapman & Hall/CRC, Taylor & Francis Group.
- 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.
- Christensen, G. S., & Miguel, E. (2016). “Transparency, Reproducibility, and the Credibility of Economics Research”, NBER Working Paper #22989.
- 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., & Boldrick, 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*. NCSE 2010-3006. Retrieved from [http://www.eric.ed.gov/ERICWebPortal/detail?accno = ED509387](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.
- Maurer, W. and Mellein, B. (1988). "On new multiple test procedures based on independent p-values and the assessment of their power," in *Multiple Hypotheses Testing*, eds. P. Bauer, G. Hommel, and E. Sonnemann, Heidelberg: Springer, pp. 48-66.
- Neyman, J. (1923). Statistical problems in agricultural experiments. *Journal of the Royal Statistical Society*, *2*, 107-180.
- Ramsey, P. H. (1978), Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association*, *73*, 479-487.

- 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

Weiss, M. J., Bloom, H. S., Verbitsky Savitz, N., Gupta, H. Vigil, A., and Cullinan, D. (forthcoming). “How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence from Existing Multisite Randomized Control Trials.” *Journal of Research on Educational Effectiveness.*

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.

Westfall P. H., Tsai K., Ogenstad S., Tomoiaga A., Moseley S., and Lu Y. (2008), Clinical Trials Simulation: A Statistical Approach. *Journal of Biopharmaceutical Statistics* 18, 611-630.

Westfall, P. H. & Troendle, J. F., (2008). Multiple Testing with Minimal Assumptions.

Biometrical Journal, 50:745-755.

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	M_0
Number of false null hypotheses	C	D	M_1
Total	$M-R$	R	M

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

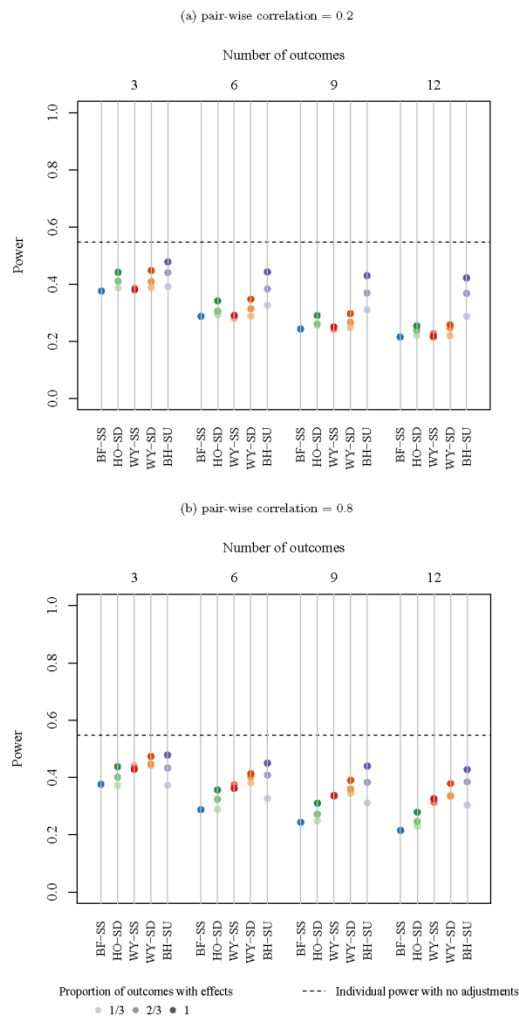


Figure 2. *Individual Power*, by Number of Outcomes, Adjustment Procedure, Proportion of Outcomes with Effects, and Pairwise Correlations Between Test Statistics: 20 Sites of 50 Individuals Each, $R^2 = 0.1$, and Effect Size = 0.125 for All Outcomes on Which There Are Effects.

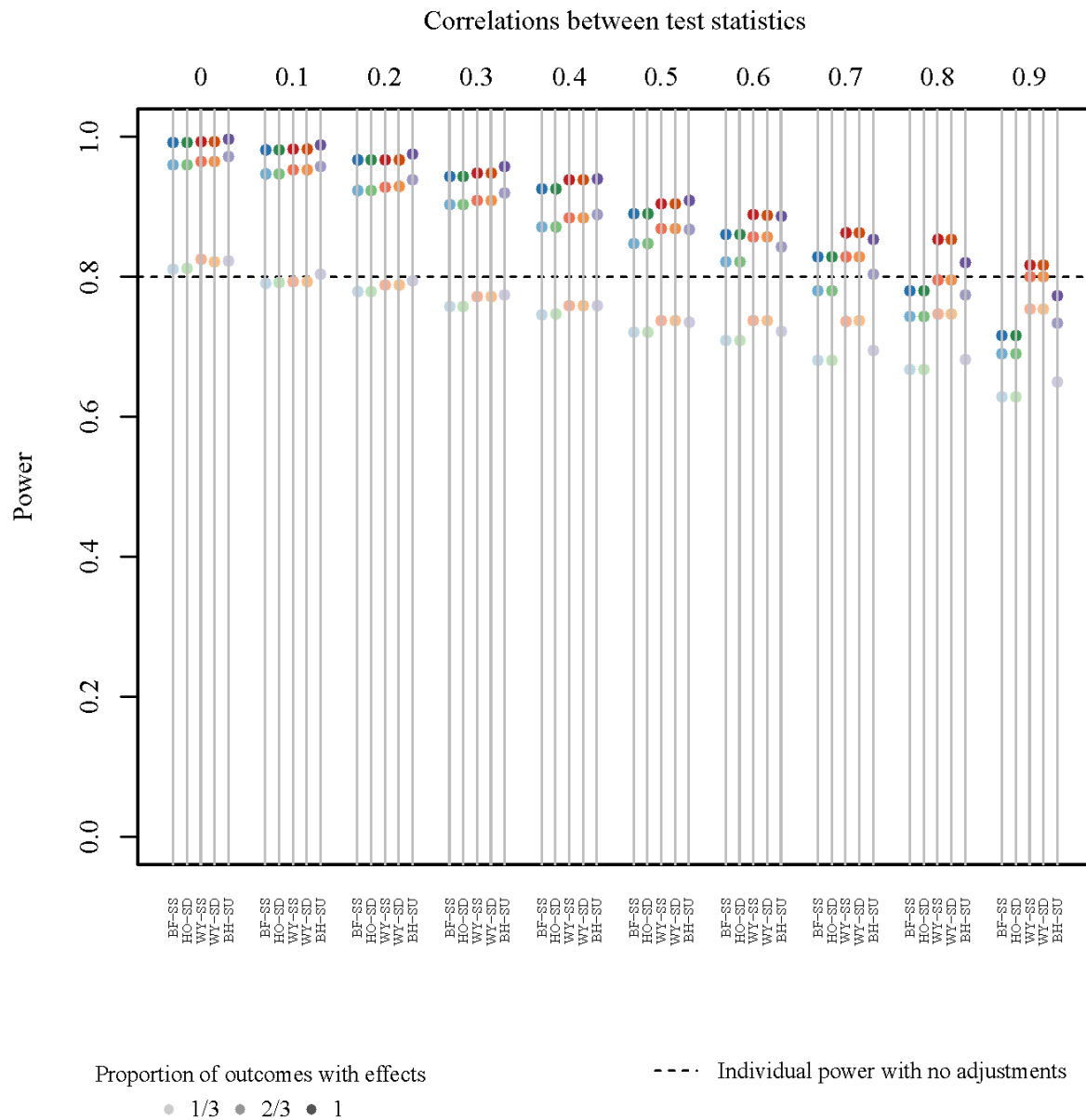


Figure 4. *1-Minimal Power*, by Adjustment Procedure, Proportion of Outcomes with Effects, and Pairwise Correlations Between Test Statistics: *Six Outcomes*, 20 Sites of 50 Individuals Each, $R^2 = 0.5$, and Effect Size = 0.125 for All Outcomes on Which There Are Effects.

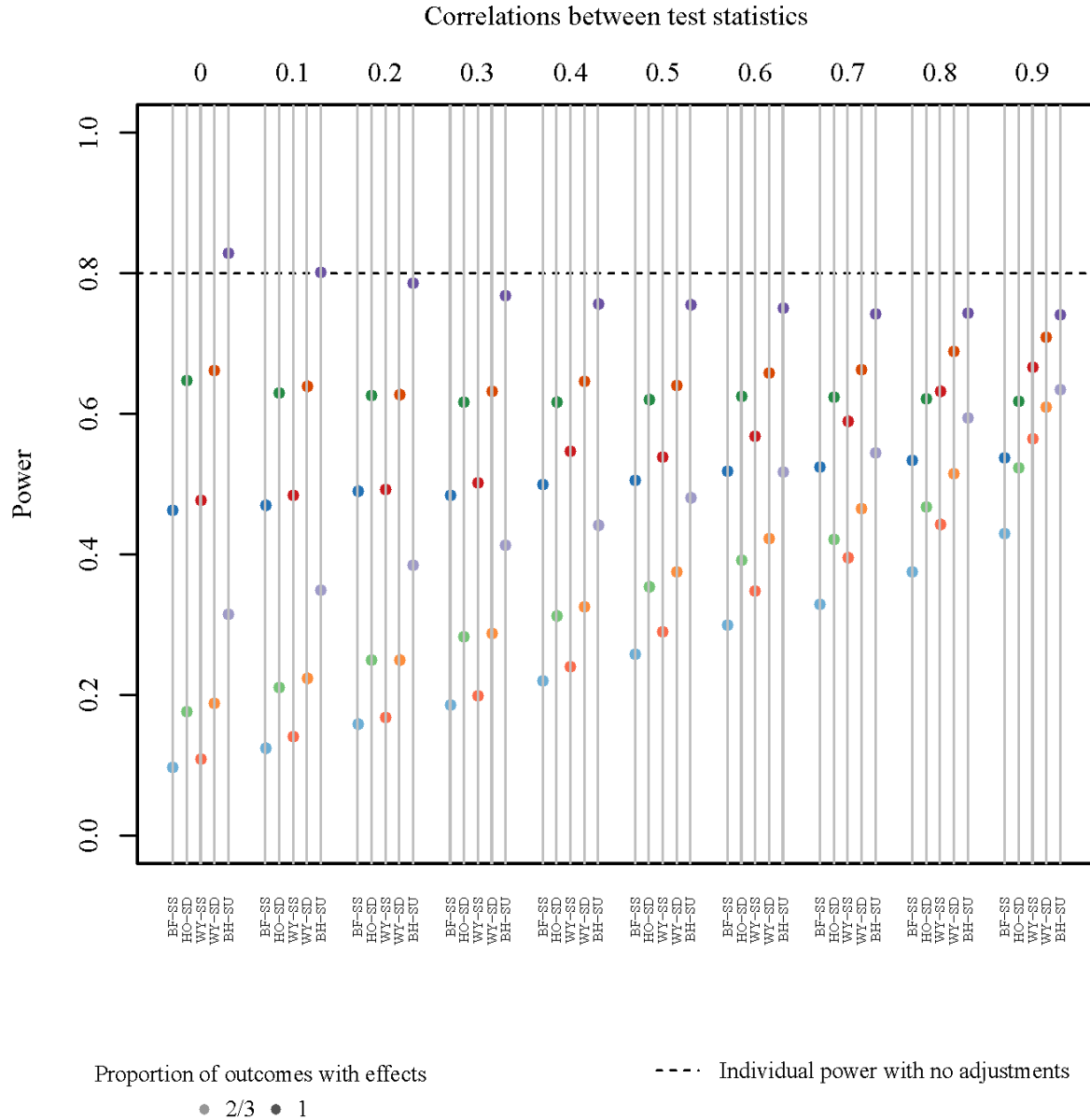


Figure 6. *2/3-Minimal Power*, by Adjustment Procedure, Proportion of Outcomes with Effects, and Pairwise Correlations Between Test Statistics: *Six Outcomes*, 20 Sites of 50 Individuals Each, $R^2 = 0.5$, and Effect Size = 0.125 for All Outcomes on Which There Are Effects.

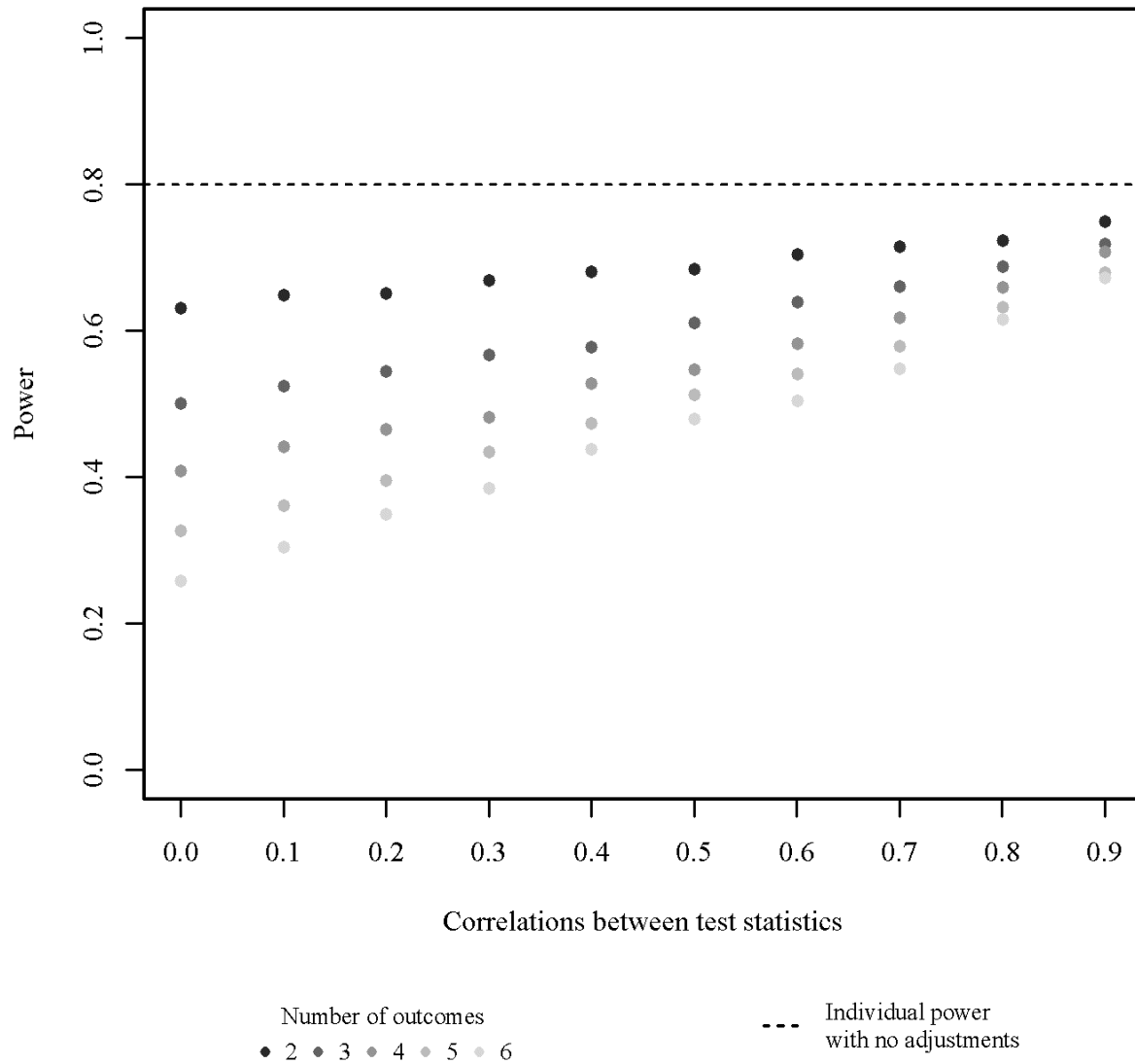


Figure 7. *Complete Power*, by Number of Outcomes and Pairwise Correlations Between Test Statistics: 20 Sites of 50 Individuals Each, $R^2 = 0.5$, and Effect Size = 0.125 for All Outcomes on Which There Are Effects.