Abstract Title Page
Not included in page count.

Title: Estimating statistical power when making adjustments for multiple tests

Authors and Affiliations: Kristin E. Porter, MDRC

Abstract Body
Limit 4 pages single-spaced.

Background / Context:
Description of prior research and its intellectual context.

In recent years, there has been increasing focus on the issue of multiple hypotheses testing in education evaluation studies. In these studies, researchers are typically interested in testing the effectiveness of an intervention on multiple outcomes, for multiple subgroups, at multiple points in time or across multiple treatment groups. When multiple hypotheses are tested, the probability of committing at least some Type I errors increases and more dramatically with a greater number of tests. Multiple testing procedures (MTPs) adjust p-values for statistical estimates upwards to counteract this problem. MTPs are being increasingly applied in impact evaluations in education. For example, the IES technical methods report, "Guidelines for Multiple Testing in Impact Evaluations," (Schochet, 2008) recommends multiple testing procedures as one of several strategies for dealing with the multiple hypotheses issue. In addition, the What Works Clearinghouse applies a particular procedure, the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) to statistically significant findings in studies under review that have estimated effects for multiple measures and/or groups (U.S. Department of Education, 2013).

However, an important consequence of making adjustments for multiplicity is a change in the statistical power for detecting true effects. It is typically argued that applying MTPs results in a *loss* of power, which can be substantial. The evidence that supports this claim focuses on "individual power," the power of each *individual* test among the multiple tests (Schochet, 2008). The extent of loss in individual power varies by circumstances particular to a given study, which may include one or more of the following: (1) number of tests, (2) the proportion of tests that are truly null, (3) the correlation between test statistics, (4) the specified probability of making a Type I error, and (5) the particular multiple testing procedure used to adjust p-values.

However, the individual power of specific hypothesis tests may not always be the most appropriate way to define power in impact evaluations with multiple tests. Just as we account for multiplicity with respect to Type I errors, we may want to account for multiplicity with respect to Type II errors (the inverse of power), as these two types of errors are inextricably linked. That is, just as we move from controlling Type I errors at the individual level to the set level (e.g. "family" of hypotheses) under multiplicity, we can also move from measuring power at the individual level to the set level. For example, perhaps in some cases it makes sense to consider the power to detect *at least one* true effect (as small as the specified minimum detectable effect sizes (MDES's)) across multiple outcomes or in other cases the power to detect *at least half* of all effects that exist or *all* effects that exist. The choice depends on the objectives of the study, or how success of the intervention is defined.

When ensuring that evaluation studies in education are sufficiently powered, typical current practice focuses on individual power and does not take the planned or actual use of multiple testing procedures into account. For determining power, sample size requirements or MDES for a single, non-adjusted test, the literature, resources and tools for helping researchers design education studies with adequate sample sizes are extensive (e.g. Dong (2013), Spybrook et al. (2011), Raudenbush et al. (2011), Hedges and Rhoads (2010), Bloom, Richburg-Hayes, and Black (2007)). However, no education or impact evaluation literature on estimating power/sample size/MDES while accounting for multiplicity adjustments was found by the

author. The IES guidelines for multiple testing (Schochet, 2008) state that "statistical power calculations for confirmatory analysis must account for multiplicity," but give no explanation for how to do this in the case that multiple testing procedures are used to adjust p-values.

Purpose / Objective / Research Question / Focus of Study:
Description of the focus of the research.

This paper provides critical alternatives to current practice in two ways. First, it presents alternatives to how power is typically defined in studies with multiple tests. The definition of power that is appropriate for a particular study can have substantial implications for the power (or the sample size requirement or the minimum detectable effects (MDE's)). With some alternative definitions of power, studies that focus on multiple outcomes may actually gain statistical power over a focus on a single outcome, even after making p-value adjustments. With other definitions, power losses can be substantial. Second, for multiple definitions of statistical power under multiplicity, this paper presents methods for estimating power while accounting for p-value adjustments using one of five common multiple testing procedures (MTPs) – Bonferroni (Dunn, 1959, 1961), Holm (Holm, 1979), single-step and step-down versions of Westfall-Young (Westfall and Young, 1995), and Benjamini-Hochberg (Benjamini and Hochberg, 1995) procedures.

The paper focuses on the scenario in which multiple tests are conducted due to an interest in effects on multiple primary outcomes. It also focuses on the simplest research design and analysis plan that education evaluations would typically use in practice: a multisite randomized trial with the randomization of individuals blocked by site - in which effects are estimated using a model with site-specific intercepts, assuming constant effects across blocks. However, the power estimation methods can easily be extended to other model assumptions as well as other study designs.

Setting: (Not applicable)
Description of the research location

Population / Participants / Subjects: (Not applicable)
Description of the participants in the study: who, how many, key features, or characteristics.

Intervention / Program / Practice: (Not applicable)
Description of the intervention, program, or practice, including details of administration and duration.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

Current practice for ensuring that impact evaluations in education have adequate statistical power does not take the use of multiplicity adjustments into account. This paper presents alternative ways to define power and presents methods for estimating power under multiplicity. All the methods for estimating power are easy to implement, fast, and can easily be extended to other multiple testing procedures not covered in this paper, as well as to other research designs.

Statistical, Measurement, or Econometric Model:

Description of the proposed new methods or novel applications of existing methods.

The methods for estimating power are motivated by the approach of using Monte-Carlo data simulation. With a simulation approach, an analyst would first specify a data-generating distribution that corresponds to her research design and model assumptions. For a blocked RCT in which we assume that effects are constant across all blocks, one would need to specify the number of blocks, the (harmonic) mean number of individuals within each block, the proportion assigned to the treatment group, the number of outcomes, the desired or predicted MDES for each outcome, the explanatory power of the block indicators and any baseline covariates (R^2), and the correlational structure of the residuals (which determines the correlational structure of the test statistics). In each of a large number of samples generated by this data generating distribution, the analyst could then estimate impacts on the multiple outcomes, use an MTP to adjust the resulting p-values for the multiple outcomes and record which p-values fall below a specified significance level, α. Individual power for detecting statistically significant effects on any given outcome when using a particular MTP is the proportion of samples in which the adjusted p-values corresponding to the outcome are less than, α (e.g. 0.05). Then, *d*-minimal power is the proportion of samples in which at least *d* of the adjusted p-values are less than α, and complete power is the proportion of samples in which all *non*-adjusted p-values are less than α. (With complete power, no adjustments to raw p-values are required because the probability that *all* tests will have a raw p-value less than α when any single test would have a raw p-value less than α just by chance, is less than the probability that any single test would have p-value less than α by chance (Koch, 1996; Westfall et al., 2011).

This simulation approach can be complex and extremely computationally intensive – particularly when using an MTP that relies on resampling (i.e. bootstrapping or permutation) to estimate the null distribution of test statistics or p-values, as is the case for the Westfall-Young MTPs.[*] The methods presented in this paper, however, avoid the need to generate any data or fit any regressions. Instead, building on the work of Band and Young (2005), the methods rely on the generation (i.e. simulation) of *test statistics* distributed under the null hypothesis, with the assumed correlation structure. These test statistics can then be shifted so that they then have the distribution under an alternative hypothesis that the effects are at least as large as specified MDES's. These test statistics under the alterative can also be converted to p-values. MTPs that can be carried out simply with raw p-values (e.g., Bonferroni, Holm, Benjamini-Hochberg) are straight-forward. For Westfall-Young, the null distribution of test statistics can be generated by simulation rather than by resampling or permutation. We implemented the methods in R.[†] On a 16 core/30 GB RAM machine, it takes, for example, just a few seconds to estimate all definitions of power for six outcomes and all MTPs other than Westfall-Young. For Westfall-Young, it takes 4.5 minutes.

Usefulness / Applicability of Method:

Demonstration of the usefulness of the proposed methods using hypothetical or real data.

[*] For example, there are three outcomes, 10,000 permutations, and 10,000 simulated samples, computing estimates of power for any given set of assumptions would involve fitting 3 * 10,000^2 regressions.

[†] We validated the power estimation methods by (1) assuming just a single block, replicating power estimates presented in Schochet (2008), which focus on Bonferroni, Holm and Benjamini-Hochberg adjustment procedures; (2) for our design of interest (a blocked RCT) and our assumed model (with constant impacts across all blocks and with school dummies included in the intercept) replicating power estimates provided in Power-Up! (Table RBD2-c) (Dong and Maynard, 2013), and replicating estimates obtained by Monte-Carlo simulation, as described above.

Overall, this paper should help applied researchers specify more accurate estimates of power and perhaps more appropriate estimates of power – for their particular study - than those that are currently used in education research.

Research Design: (Not applicable)
Description of the research design.

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

The power estimation methods were used to examine how statistical power changes when moving from testing for effects on a single outcome to testing for effects on more outcomes and how the changes in power depend on the MTP, the numbers of tests, R^2, the correlations between the test statistics, and the proportions of null hypotheses that are actually false. Assuming that there will be effects on all outcomes may be misleading, especially if researchers include outcomes on which they do not necessarily expect impact but which they include because they are policy relevant.

Findings / Results:
Description of the main findings with specific details.

Some of the key preliminary findings are: (1) Having sufficient power to detect effects on each outcome (individual power) is not the same as having sufficient power to detect effects on all outcomes (complete power). If researchers really care about detecting impacts on all outcomes under primary consideration, a focus on complete power may be warranted. This typically means that researchers have much less power – or need a much larger sample size or must settle for higher MDES's – than they anticipated if they powered their studies for individual power or ignored multiplicity altogether in the planning phase. (2) If researchers really only care about detecting impacts on at least one outcome or a small fraction of outcomes, then they often will have more power than they would have anticipated for individual power or even if they had planned to ignore multiplicity altogether. This could alternatively mean smaller sample sizes are required or smaller MDES's can be detected. (3) Across a wide a variety of scenarios, the Benjamini-Hochberg MTP typically results in the best power (with some exceptions but often for unlikely scenarios in practice). However, a trade-off in using the Benjamini-Hochberg MTP is that because it controls the false discovery rate, it is less strict about false positives. If one is worried about any false positives, the step-down version of the Westfall-Young MTP, which provides strong control of the family-wise error rate, will most likely produce the most power, especially when the test statistics are moderately or strongly correlated and the number of outcomes is large. (4) When planning a study that will test for effects on multiple outcomes, power estimates that assume effects on all outcomes can be overly optimistic if that is not the case, often substantially so. (5) Relative to power for a single outcome, changes in power that are the result of adjusting for multiplicity tend to be even greater with a higher R^2 in a blocked RCT.

Conclusions:
Description of conclusions, recommendations, and limitations based on findings.

Researchers should consider alternative definitions of power when it is appropriate for the objectives of their study, and they should estimate power that takes multiplicity adjustments into account. The methods for estimating power are straight-forward and fast.

Appendices
Not included in page count.

Appendix A. References
References are to be in APA version 6 format.

Bang, S.J. and Young, S.S. (2005). Sample size calculation for multiple testing in microarray data analysis. *Biostatistics* 6, 157–169.

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*(1), 289-300.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using Covariates to Improve Precision for Studies that Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59.

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. and 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*(1), 24-67. doi: 10.1080/19345747.2012.673143

Dudoit, S., Shaffer, JP, Bodrick, JC. (2003). Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science, 18*(1), 71-103.

Dunn, Olive Jean (1959). Estimation of the Medians for Dependent Variables. *Annals of Mathematical Statistics* 30(1): 192–197.

Dunn, Olive Jean (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association* 56(293): 52–64.

Hedges, L. V., & Rhoads, C. (2010). Statistical Power Analysis in Education Research. NCSER 2010-3006: National Center for Special Education Research. 400 Maryland Avenue SW, Washington, DC 20202. Tel: 800-437-0833; Fax: 202-401-0689; Web site: http://ies.ed.gov/ncser/.

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

Koch, G., Gansky, MS. (1996). Statistical Considerations for Multiplicity in Confirmatory Protocols. *Drug Information Journal, 30*, 523-533.

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X., Martinez, A., Bloom, H., & Hill, C. (2011). Optimal Design Plus Empirical Evidence (Version 3.0).

Schochet, P. Z. (2008). Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations. NCEE 2008-4018: National Center for Education Evaluation and Regional Assistance. Available from: ED Pubs. P.O. Box 1398, Jessup, MD 20794-1398. Tel: 877-433-7827; Web site: http://ies.ed.gov/ncee/pubs/.

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

Spybrook, J., Bloom, H. S., Congdon, R., Hill, C. J., Martinez, A., & Raudenbush, S. W. (Eds.). (2011).

U.S. Department of Education. (2013) *Institute of Education Sciences*. National Center for Education Evaluation and Regional Assistance: What Works Clearinghouse.

Westfall, P. H. and Young, S. S. (1993). Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. John Wiley, New York.

Westfall, P. H., Tobias, R. D., & Wolfinger, R. D. (2011). *Multiple Comparisons and Multiple Tests Using SAS, Second Edition*. Cary, N.C.: The SAS Institute.

Appendix B. Tables and Figures
Not included in page count.