Title: Estimating Statistical Power When Making Adjustments for Multiple Tests - Design and Software Extensions

Authors: Kristin E. Porter, Deni Chen and Zarni Htet

Background

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

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

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

A New Framework

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 (Chen, Luo, Liu, & Mehrotra, 2011; Dudoit, Shaffer, & Bodrick, 2003; Senn & Bretz, 2007; Westfall, Tobias, & Wolfinger, 2011). 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 vary by outcome) on at least one outcome. Similarly, one might consider ½-minimal power: the probability of detecting effects of at least a particular size on at least ½ of the outcomes.

Also, one might consider complete power: the power to detect effects of at least a particular size on all outcomes. The choice of definition of power depends on the objectives of the study and on how the success of the intervention is defined. The choice of definition also affects the overall extent of power.

Recent Progress

A recent paper by the proposed Principal Investigator, which is the product of an IES Early Career, Statistical and Research Methodology in Education grant (R305D140024), 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 (Porter, 2016). It also presents empirical findings on how power is affected by multiple factors under multiplicity: the definition of power, the number of tests, the proportion of tests that are truly null, the correlation between tests, the R2′s of baseline covariates, and the particular MTP used to adjust p-values.

The recent paper began to fill a gap in the existing literature on statistical power in education studies, which does not take multiplicity into account (Dong & Maynard, 2013; Hedges & Rhoads, 2010; Raudenbush et al., 2011; Spybrook et al., 2011). However, to contain the scope, Porter (2016) focuses only on multiplicity that results from estimating effects on multiple outcomes. The paper also focuses only on the simplest research design and analysis plan that education studies typically use in practice: a multisite, randomized controlled trial (RCT) with the blocked randomization of individuals, in which effects are estimated using a model with block-specific intercepts and with the assumption of constant effects across blocks.
The narrow focus of the paper allowed for the development, implementation and validation of a computationally efficient methodology for estimating power under multiplicity.

Next Steps Addressed in Current Paper

However, to maximize the value and widespread adoption of the methodology by applied researchers in education and other fields, it is important to: 1. extend the methodology and recommendations for practice to other modeling assumptions (e.g., random effects), other study designs (e.g. cluster RCT’s and blocked cluster RCT’s) and perhaps other types of multiplicity (e.g., due to multiple subgroups); 2. publish open-source software that is readily available and easy to use for applied researchers; and 3. provide a variety of ways to increase awareness of the issue and of the readily accessible solution (e.g. an interactive web application and a webinar available on MDRC’s website, as well as typical conference presentations and a “guide for researchers”). The paper presented in this poster will accomplish all three.

References

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). Annals of Mathematical Statistics, 30(1): 192–197.

Dunn, Olive Jean (1961). 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.

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.

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.