mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

REFLECTIONS ON METHODOLOGY

NOVEMBER 2017

Estimating Statistical Power When Using Multiple Testing Procedures

By Kristin E. Porter

This post is one in a series highlighting MDRC's methodological work. Contributors discuss the refinement and practical use of research methods being employed across our organization.

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 increase the likelihood of spurious findings: that is, finding statistically significant effects that do not in fact exist. Multiple testing procedures (MTPs) are statistical procedures that counteract this problem by adjusting p-values for effect estimates upward. Without the use of an MTP, the probability of false positive findings increases, sometimes dramatically, with the number of tests.

Yet the use of an MTP can result in a substantial change in statistical power, greatly reducing the probability of detecting effects when they do exist. Thus, there is a trade-off between a lower probability of detecting a true effect when adjusting and a higher probability of a false positive when not adjusting. Unfortunately, when designing studies, researchers using MTPs frequently ignore their power implications. In some cases sample sizes may be too small and studies may be underpowered to detect the desired size of an effect. In other cases, sample sizes may be larger than needed and studies may be powered to detect smaller effects than anticipated.

But the use of an MTP need not always mean a loss of power. *Individual* power is lost — the probability of detecting an effect of a particular size or larger for each hypothesis test. However, in studies with multiplicity, alternative definitions of power exist and in some cases may be more appropriate.[1] 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. Or one might consider *½-minimal* power: the probability of detecting effects of at least a particular size on at least half the outcomes. Finally, one might consider *complete* power: the power to detect effects of at least a particular size on all outcomes. How to define 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 study's overall level of power.

At MDRC, with two consecutive Statistical and Research Methodology in Education grants from the Institute of Education Sciences (IES) (R305D140024 and R305D170030), we have been developing methods for estimating statistical power for multiple definitions of power when applying any of five common MTPs — Bonferroni (Dunn 1959, 1961), Holm (1979), single-step and step-down versions of Westfall-Young (1993), and Benjamini-Hochberg (1995). Our methods are described in detail in an article and a working paper, focusing on multiplicity that results from estimating effects when using a simple and common research design: a multisite, randomized controlled trial in which individuals are randomized in blocks and effects are estimated using a model with block-specific intercepts and the assumption of constant effects across blocks.

[1] Chen, Luo, Liu, and Mehrotra (2011); Dudoit, Shaffer, and Boldrick (2003); Senn and Bretz (2007); Westfall, Tobias, and Wolfinger (2011).

NEW YORK
16 East 34th Street
New York, NY 10016
Tel: 212 532 3200

OAKLAND
475 14th Street
Suite 750
Oakland, CA 94612
Tel: 510 663 6372

WASHINGTON, DC
1990 M Street, NW
Suite 340
Washington, DC 20036

LOS ANGELES
11965 Venice Boulevard
Suite 402
Los Angeles, CA 90066

www.mdrc.org

We used these methods to produce empirical findings about how various factors in studies with multiplicity affect overall statistical power. The results point to several recommendations for practice, outlined below. In addition, with IES support, MDRC is exploring other applications for these methods and recommendations. We are in the process of creating and publishing user-friendly, open-source software for applied researchers and developing an interactive web application to allow users to create plots of power, minimum detectable effect sizes, or sample size requirements for studies with multiplicity.

RECOMMENDATIONS FOR PRACTICE

Prespecify all hypothesis tests and prespecify a plan for making multiplicity adjustments. Researchers who plan to use an MTP should account for its effects on statistical power when determining the study's sample size.

Think about the definition of success for the intervention under study and choose a corresponding definition of statistical power. The prevailing default in many studies — individual power — may or may not be the most appropriate type of power. 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 a small proportion of outcomes, 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 effects.

In some cases, it may be best for researchers to estimate and share power estimates for multiple power definitions. For example, even if a program would be considered successful should an effect of a specified size be found for at least one outcome (requiring 1-minimal power), researchers may still want to know the probability of detecting effects on each particular outcome (requiring individual power). Or consider the case in which a sample size is fixed and the probability of detecting statistically significant effects on all outcomes (complete power) is unacceptably low. It still may be valuable for researchers to be able to achieve a high probability of detecting effects on at least half the outcomes (½-minimal power).

Do not necessarily choose the MTP that results in the most power. Most MTPs control the family-wise error rate, or the probability of at least one false positive. The Benjamini-Hochberg MTP, which generally results in the most power, controls the false discovery rate, which is the expected proportion of rejected null hypotheses that are wrongly rejected. An MTP that controls the false discovery rate is more lenient with false positives than one that controls the family-wise error rate. Researchers may tolerate a few false positives when testing for effects on a large number of outcomes, but when the number of outcomes is small, a single false positive is more likely to lead to the wrong conclusion about an intervention's effectiveness. In that case, controlling the family-wise error rate is likely to be preferable, by using one of the Bonferroni, Holm, or Westfall-Young MTPs. Among these, the Westfall-Young step-down procedure generally results in the most power. However, if a low or moderate correlation between outcomes is expected, or if the study will use a 1-minimal definition of power, the simpler Holm MTP or the single-step Westfall-Young MTP may suffice.

Consider the possibility that there may not be impacts on all outcomes. 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 defined only when a true effect exists. However, in stepwise MTPs (Benjamini-

Hochberg and one version of Westfall-Young), the p-value adjustments are made in a series where each depends on the one before. If some outcomes show no effects, the probability of detecting the effects that do actually exist can be diminished, sometimes substantially (see Porter 2017).

Take all of the above into account in the design phase of a study when estimating power, sample size requirements, or minimum detectable effect sizes. Working through these recommendations 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 may mean that they want to be parsimonious in selecting their primary outcomes. As with all facets of study design, it is crucial to understand the trade-offs involved.

REFERENCES

Bang, Heejung, Sin-Ho Jung, and Stephen L. George. 2005. "Sample Size Calculations for Simulation-Based Multiple-Testing Procedures." *Journal of Biopharmaceutical Statistics* 15, 6: 957-967.

Benjamini, Yoav, and Yosef Hochberg. 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.

Chen, Jie, Jianfeng Luo, Kenneth Liu, and Devan V. Mehrotra. 2011. "On Power and Sample Size Computation for Multiple Testing Procedures." *Computational Statistics and Data Analysis* 55, 1: 110-122.

Dudoit, Sandrine, Juliet Popper Shaffer, and Jennifer C. Boldrick. 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. doi:10.1214/aoms/1177706374

Dunn, Olive Jean. 1961. "Multiple Comparisons Among Means." *Journal of the American Statistical Association* 56, 293: 52-64. doi:10.1080/01621459.1961.10482090

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6, 2: 65-70.

Porter, Kristin E. 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

Senn, Stephen, and Frank Bretz. 2007. "Power and Sample Size When Multiple Endpoints Are Considered." *Pharmaceutical Statistics* 6, 3: 161-170. doi:10.1002/pst.301

Westfall, Peter H., Randall D. Tobias, and Russell D. Wolfinger. 2011. *Multiple Comparisons and Multiple Tests Using SAS.* 2nd ed. Cary, NC: SAS Institute.

Westfall, Peter H., and S. Stanley Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for* p-*Value Adjustment*. New York: John Wiley & Sons.