Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. doi: 10.18637/jss.v000.i00

# Power Under Multiplicity Project (PUMP): Estimating Power, Minimum Detectable Effect Size, and Sample Size When Adjusting for Multiple Outcomes

**Kristen Hunter**
Harvard University
Department of Statistics

**Luke Miratrix**
Harvard Graduate
School of Education

**Kristin Porter**
MDRC

**Abstract**

For randomized controlled trials (RCTs) with a single intervention being measured on multiple outcomes, researchers often apply a multiple testing procedure (such as Bonferroni or Benjamini-Hochberg) to adjust $p$-values. Such an adjustment reduces the likelihood of spurious findings, but also changes the statistical power, sometimes substantially, which reduces the probability of detecting effects when they do exist. However, this consideration is frequently ignored in typical power analyses, as existing tools do not easily accommodate the use of multiple testing procedures. We introduce the PUMP R package as a tool for analysts to estimate statistical power, minimum detectable effect size, and sample size requirements for multi-level RCTs with multiple outcomes. Multiple outcomes are accounted for in two ways. First, power estimates from PUMP properly account for the adjustment in $p$-values from applying a multiple testing procedure. Second, as researchers change their focus from one outcome to multiple outcomes, different definitions of statistical power emerge. PUMP allows researchers to consider a variety of definitions of power, as some may be more appropriate for the goals of their study. The package estimates power for frequentist multi-level mixed effects models, and supports a variety of commonly-used RCT designs and models and multiple testing procedures. In addition to the main functionality of estimating power, minimum detectable effect size, and sample size requirements, the package allows the user to easily explore sensitivity of these quantities to changes in underlying assumptions.

*Keywords*: power, multiple testing, multi-level models, randomized controlled trials, R.

# 1. Introduction

The PUMP R package fills in an important gap in open-source software tools to design multi-level randomized controlled trials (RCTs) with adequate statistical power. With this package, researchers can estimate statistical power, minimum detectable effect size (MDES), and needed sample size for multi-level experimental designs, in which units are nested within hierarchical structures such as students nested within schools nested within school districts. The statistical power is calculated for estimating the impact of a single intervention on multiple outcomes. The package uses a frequentist framework of mixed effects regression models, which is currently the prevailing framework for estimating impacts from experiments in education and other social policy research.[1]

To our knowledge, none of the existing software tools for power calculations allow researchers to account for multiple hypothesis tests and the use of a multiple testing procedure (MTP). MTPs adjust $p$-values to reduce the likelihood of spurious findings when researchers are testing for effects on multiple outcomes. This adjustment can result in a substantial change in statistical power, greatly reducing the probability of detecting effects when they do exist. Unfortunately, when designing studies, researchers who plan to test for effects on multiple outcomes and employ MTPs frequently ignore the power implications of the MTPs.

Also, as researchers change their focus from one outcome to multiple outcomes, multiple definitions of statistical power emerge (Chen, Luo, Liu, and Mehrotra (2011); Dudoit, Shaffer, and Boldrick (2003); Senn and Bretz (2007); Westfall, Tobias, and Wolfinger (2011)). The PUMP package allows researchers to consider multiple definitions of power, selecting those most suited to the goals of their study. The definitions of power include:

- **individual power**: the probability of detecting an effect of a particular size (specified by the researcher) or larger for each hypothesis test. Individual power corresponds to how power is defined when there is focus on a single outcome.
- $1-$**minimal power**: the probability of detecting effects of at least a particular size on at least one outcome. Similarly, the researcher can consider $d-$**minimal power** for any $d$ less than the number of outcomes, or fractional powers, such as $1/2-$minimal power.
- **complete power**: the power to detect effects of at least a particular size on *all* outcomes.

As noted in Porter (2018), the prevailing default in many studies—individual power—may or may not be the most appropriate type of power. If the researcher's goal is to find statistically significant estimates of effects on most or all primary outcomes of interest, then their power may be much lower than anticipated when multiplicity adjustments are taken into account. On the other hand, if the researcher's 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. In both of these cases, by not accounting for both the challenges and opportunities arising from multiple outcomes, a researcher may find they have wasted resources, either by designing an underpowered study that cannot detect the desired effect sizes, or by designing an overpowered study that had a larger sample size than necessary. We introduce the PUMP package to allow for directly answering questions that take multiple outcomes into account, such as:

[1] Other options include nonparametric or Bayesian methods, but these are less prevalent in applied research (for example, see Gelman, Hill, and Yajima (2012), Gelman, Hill, and Yajima (2007)).

- How many schools would I need to detect a given effect on at least three of my five outcomes?
- What size effect can I reliably detect on each outcome, given a planned MTP across all my outcomes?
- How would the power to detect a given effect change if only half my outcomes truly had impact?

The methods in the PUMP package build on those introduced in Porter (2018). This earlier paper focused only on a single RCT design and model — 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 all units. This earlier paper also did not produce software to assist researchers in implementing its methods. With this current paper and with the introduction of the PUMP package, we extend the methodology to nine additional multi-level RCT designs and models. Also, while Porter (2018) focused on estimates of power, PUMP goes further to also estimate MDES and sample size requirements that take multiplicity adjustments into account.

`PUMP` extends functionality of the popular PowerUp! `R` package (and its related tools in the form of a spreadsheet and Shiny application), which compute power or MDES for multi-level RCTs with a single outcome (Dong and Maynard (2013)). For a wide variety of RCT designs with a single outcome, researchers can take advantage of closed-form solutions and numerous power estimation tools. For example, in education and social policy research, see Dong and Maynard (2013); Hedges and Rhoads (2010); Raudenbush, Spybrook, Congdon, Liu, Martinez, Bloom, and Hill (2011); Spybrook, Bloom, Congdon, Hill, Martinez, and Raudenbush (2011). However, closed-form solutions are difficult or impossible to derive when a MTP is applied to a setting with multiple outcomes. Instead, we use a simulation-based approach to achieve estimates of power.

In order to calculate power, the researcher specifies information about the sample size at each level, the minimum detectable effect size for each outcome, the level of statistical significance, and parameters of the data generating distribution. The minimum detectable effect size is the smallest true effect size the study can detect with the desired statistical significance level, in units of standard deviations. An "effect size" generally refers to the standardized mean difference effect size, which "equals the difference in mean outcomes for the treatment group and control group, divided by the standard deviation of outcomes across subjects within experimental groups" (Bloom (2006)). Researchers often use effect sizes to standardize outcomes so that outcomes with different scales can be directly compared.

The package includes three core functions:

- `pump_power()` for calculating power given a experimental design and assumed model, parameters, and minimum detectable effect size.
- `pump_mdes()` for calculating minimum detectable effect size given a target power and sample sizes.
- `pump_sample()` for calculating the required sample size for achieving a given target power for a given minimum detectable effect size.

For any of these core functions, the user begins with two main choices. First, the user chooses the assumed design and model of the RCT. The `PUMP` package covers a range of multi-level designs, up to three levels of hierarchy, that researchers typically use in practice, in which research units are nested in hierarchical groups. Our power calculations assume the user will be analyzing these RCTs using frequentist mixed-effects regression models, containing a combination of fixed or random intercepts and treatment impacts at different levels, as we explain in detail in Section~**??** and in the Technical Appendix. Second, the user chooses the MTP to be applied. `PUMP` supports five common MTPs — Bonferroni, Holm, single-step and step-down versions of Westfall-Young, and Benjamini-Hochberg. After these two main choices, the user must also make a variety of decisions about parameters of the data generating distribution.

The package also includes functions that allow users to easily explore power over a range of possible values of parameters. This exploration encourages the user to determine the sensitivity of estimates to different assumptions. `PUMP` also visually displays results. These additional functions include:

- `pump_power_grid()`, `pump_mdes_grid()`, and `pump_sample_grid()` for calculating the given output over a range of possible parameter values.
- `update()` to re-run an existing calculation with a small number of parameters updated.
- `plot()` on `PUMP`-generated objects to generate plots (including grid outputs).

The authors of the `PUMP` package have also created a web application built with R Shiny. This web application calls the `PUMP` package and allows users to conduct calculations with a user-friendly interface, but it is less flexible than the package, with a focus on simpler scenarios (e.g., 10 or fewer outcomes). The app can be found at `https://mdrcpump.shinyapps.io/pump_shiny/`.

The remainder of this paper is organized as follows. In Section~**??**, we introduce Diplomas Now, an educational experiment, to be used as a running example throughout the paper. We note, however, that the problem of power estimation for multi-level RCTs is not exclusive to the educational setting. In Section~**??**, we provide a summary of the multiple testing problem. Also in Section~**??**, we present an overview of the statistical challenges introduced by multiple hypothesis testing and how MTPs protect against spurious impact findings. In Section~**??**, we introduce our methodology for estimating power when taking the use of MTPs into account. This section also briefly discusses our validation process. Section~**??** discusses the various choices a user must make when using the package, including the designs and models, MTPs, and key design and model parameters. Section~**??** provides a detailed presentation of the `PUMP` package with multiple examples of using the packages functions to conduct calculations for our education RCT example. Section~**??** is a brief conclusion.

## 1.1. Code formatting

In general, don't use Markdown, but use the more precise LaTeX commands instead:

- Java
- **plyr**

One exception is inline code, which can be written inside a pair of backticks (i.e., using the Markdown syntax).

If you want to use LaTeX commands in headers, you need to provide a `short-title` attribute. You can also provide a custom identifier if necessary. See the header of Section 2 for example.

# 2. R code

Can be inserted in regular R markdown blocks.

```
R> x <- 1:10
R> x

 [1]  1  2  3  4  5  6  7  8  9 10
```

## 2.1. Features specific to rticles

- Adding short titles to section headers is a feature specific to **rticles** (implemented via a Pandoc Lua filter). This feature is currently not supported by Pandoc and we will update this template if it is officially supported in the future.
- Using the `\AND` syntax in the `author` field to add authors on a new line. This is a specific to the `rticles::jss_article` format.

# References

Bloom HS (2006). "The Core Analytics of Randomized Experiments for Social Research."

Chen J, Luo J, Liu K, Mehrotra D (2011). "On Power and Sample Size Computation for Multiple Testing Procedures." *Computational Statistics and Data Analysis*, **55**, 110–122.

Dong N, Maynard R (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. ISSN 1934-5747.

Dudoit S, Shaffer J, Boldrick J (2003). "Multiple Hypothesis Testing in Microarray Experiments." *Statistical Science*, **18**(1), 71–103.

Gelman A, Hill J, Yajima M (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman A, Hill J, Yajima M (2012). "Why We (Usually) Don't Have to Worry About Multiple Comparisons." *Journal of Research on Educational Effectiveness*, **5**, 189–211.

Hedges LV, Rhoads C (2010). "Statistical Power Analysis in Education Research." *Report*, National Center for Special Education Research. Http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED509387, URL `http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED509387`.

Porter KE (2018). "Statistical Power in Evaluations That Investigate Effects on Multiple Outcomes: A Guide for Researchers." *Journal of Research on Educational Effectiveness*, **11**, 267–295.

Raudenbush S, Spybrook J, Congdon R, Liu X, Martinez A, Bloom H, Hill C (2011). "Optimal Design Plus Empirical Evidence (Version 3.0)." *Report*. URL http://wtgrantfoundation.org/resource/optimal-design-with-empirical-information-od.

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, Congdon R, Hill CJ, Martinez A, Raudenbush SW (2011). "Optimal Design Plus Empirical Evidence: Documentation for the "Optimal Design" Software Version 3.0." *Report*. URL http://wtgrantfoundation.org/resource/optimal-design-with-empirical-information-od.

Westfall PH, Tobias R, Wolfinger RD (2011). *Multiple Comparisons and Multiple Tests using SAS*. The SAS Institute. ISBN 9781607648857, 9781607647836, 9781642955187.

**Affiliation:**

Kristin Porter
Universitat Autònoma de Barcelona
MDRC
475 14th Street
Suite 750
Oakland, CA 94612-1900

MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd