# PUMP manuscript draft

Power Under Multiplicity Project (PUMP): An R package for estimating statistical power, minimum detectable effect sizes (MDES's) and sample sizes when adjusting for multiple hypothesis tests Draft outline

## TODO list

Topics to possibly discuss:

- effect size: what it is, how to decide what it is
- correlation between test statistics, relationship to correlation between outcomes

## Introduction

### Overview

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.

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 (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 can vary by outcome) on at least one outcome. Similarly, one might consider 1/2-minimal power: the probability of detecting effects of at least a particular size on at least 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. The choice of definition also affects the overall extent of power.

The methodological developments implemented in our R package, PUMP, and described in this paper are focused on the multiplicity problem that arises in randomized control trials (RCT's) that test an intervention's impacts on a modest number of outcomes. For example, in education a researcher might design a trial to

investigate the effects of a mentoring program on three outcomes related to social and emotional development — measures of social competence, emotional competence and self-regulation. For this type of research, the literature and tools for estimating statistical power, or for estimating sample size requirements of minimum detectable effect sizes (MDES's) that achieve a desired level of statistical power, are extensive.[1] However, to our knowledge, no tools exist that take multiplicity into account.

The PUMP package fills this gap. It allows users to estimate statistical power, sample size requirements or MDES's 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 and when using any of the following RCT designs and estimation models:

- Individual random assignment
- Blocked individual random assignment
  - Constant effects model (2 levels - e.g., students blocked within teachers)
  - Fixed effects model (2 levels and 3 levels - e.g., students blocked within teachers within schools)
  - Random effects model (2 levels, 3 levels)
- Cluster random assignment
  - 2-level model (treatment at Level 2 - e.g., at teacher level)
  - 3-level model (treatment at Level 3 - e.g., at school level)
- Blocked cluster random assignment
  - 3-level fixed-effects model (treatment at Level 2)
  - 3-level random-effects model (treatment at Level 2)

For details about the assumptions for each of the estimation models, see X.

## Review of the multiple testing problem in a frequentist framework

This paper focuses on the frequentist framework of hypothesis testing, as it is currently the prevailing framework in education and other social policy research. In the frequentist framework, when framing impacts in terms of effect sizes, one typically tests a null hypothesis of no effect, $H0_m : ES_m = 0$, against an alternative hypothesis of $H1_m : ES_m \neq 0$ for a two-sided tests or $H1_m : ES_m > 0$ or $H1_m : ES_m < 0$ for a one-sided test. For the purposes of computing power researchers specify an alternative hypothesis of at least a particular effect size. For example, researchers may specify 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

$$t_m = \frac{\hat{ES}_m}{SE(\hat{ES}_m)}, \tag{1}$$

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 the discussion going forward (although the PUMP package also allows for one-sided tests), the raw p-value for the $m^{th}$ test is $p_m = 2 * PrT_m \geq |t_m|$.[2] 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.

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, $\alpha$. A Type I error is the probability of erroneously rejecting the null hypothesis when it is true. The quantity $\alpha$ is also referred to as the significance level. If $\alpha = 0.05$, then the null hypothesis is rejected if the p-value is

[1] For example, power estimating tools frequently used in social science research include Dong and Maynard (2013) Dong & Maynard, 2013; Hedges & Rhoads, 2010; Raudenbush et al., 2011; Spybrook et al., 2011).

[2] For a one-sided test, depending on the direction of our alternative hypothesis, the raw p-value for the $m^{th}$ test is computed as $p_m = PrT_m \leq t_m$ or $p_m = PrT_m \geq t_m$.

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 null hypothesis tests will be erroneously rejected is 1-Pr(none of the null hypotheses will be erroneously rejected) $= 1 - (1 - \alpha)^m$. Therefore, if researchers are estimating effects on three outcomes, if these outcomes are assumed independent, the probability of at least one false positive finding is 0.14. If the researchers were instead estimating effects on five independent outcomes, the probability of at least one false positive finding is 0.23. 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.[3] 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). There-fore, 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).

## 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 $\alpha$ . 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 $\alpha$. In Table 1, A, B, C and D are four possible scenarios: the numbers of true or false hypotheses not rejected or rejected. $M0$ and $M1$ 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.

[3] Alternatively, 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.

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 $\alpha$ 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 $E(\frac{B}{R})$ 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 $FWER = P(R > 0) = E(\frac{B}{R}) = FDR$. 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 $\alpha$ 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 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.

# Estimating Power, MDES's and Sample Sizes in Studies of Impacts on Multiple Outcomes (Kristen - first draft October 1)

- Approach for estimating statistical power
  - General strategy
  - Table of all designs (showing naming convention)
- Approach for estimating MDES's and sample size
  - fun with optimization
- Validation (how much/what should we present?)
  - paragraph on what we did
  - make validation results available on Github
    - include validation template
    - include our simulation code
    - discrepancy between our results and power-up

## Power estimation strategy

In order to estimate power for a single outcome, or even a small set of outcomes, we can often use closed-form algebraic expressions derived from our assumed model. However, with multiple outcomes, finding such expressions can be quite difficult, or even impossible depending on the multiple testing procedure. Instead, we rely on simulation to calculate estimated power.

If we were to rely on a full simulation approach, we could use the following method to estimate power:

1. Simulate data according to the alternative hypotheses (assuming true nonzero effects).
2. Calculate test statistics $t_1$ under the alternative hypothesis.
3. Use these test statistics to calculate $p$-values.

4. Calculate power using the distribution of $p$-values.

However, we can simplify this process by skipping the first step. Given an assumed model and correlation structure for the test statistics, we can directly sample from $f(t_1)$, the joint alternative distribution of the test statistics. This shortcut improves both the simplicity and the speed of computation.

We now describe how to sample from $f(t_1)$ directly. First, we assume a particular research design and model. Define $\psi_m$ as the treatment effect for outcome $m$. Then, we can also express the treatment effect in terms of effect size:

$$ES_m = \frac{\psi_m}{\sigma_{Y,m}}$$

where $\sigma_{Y,m}$ is the standard deviation of outcome $Y_m$. In order to calculate power, we are interested in the standard error of the estimated effect size, which we denote as

$$Q_m = SE(\hat{ES}_m).$$

The quantity $Q_m$ is defined by the assumed model, and can be a function of the number of units at different levels, the percent of units treated, the assumed $R^2$, and other parameters. Finally, when testing the hypothesis for outcome $m$, the test statistics for a $t$ test is:

$$t_m = \frac{\hat{ES}_m}{\hat{Q}_m}$$

with degrees of freedom $df$, also defined by the assumed model. Under the alternative hypothesis, $t_m$ has a $t$ distribution with degrees of freedom $df$ and mean $\hat{ES}_m/\hat{Q}_m$. We choose the correlation between test statistics $\rho$ to sample jointly from $t_m, m = 1, \ldots M$.

For Westfall-Young procedures, this method is augmented by also sampling test statistics under the joint null distribution. Under the null hypothesis, $t_m$ has a $t$ distribution with degrees of freedom $df$ and mean 0.

## RCT Designs

When designing a study, the researcher has two main choices. First, the researcher chooses the design of the experiment, including the number of levels, and at which level randomization occurs. Second, and separately, the researchers chooses an assumed model, including whether intercepts and treatment effects should be treated as constant, fixed, or random. The same experimental design can have different modeling choices, and these two decisions should be conceptually separated from each other. The PUMP package supports models with 1, 2, or 3 levels, with randomization occurring at any level. For example, a design with 2 levels and randomization at level 1 is a blocked design. A design with 3 levels and randomization at level 3 is a cluster design.

For modeling, we have the following choices.

- Whether level 2 and level 3 intercepts are:
  - fixed: intercepts are fixed effects constrained to have mean 0.
  - random: intercepts are considered to be Normally distributed, allowing for partial pooling.
- Whether level 2 and level 3 treatment effects are:
  - constant: all units in a level have the same single average impact.
  - fixed: each unit within a level has an individual estimated impact, with an additional mean impact.
  - random: treatment impacts are Normally distributed around a mean impact.

The research design is denoted by $d$, followed by the number of levels and randomization level, so `d3.1` is a 3-level design with randomization at level 1. The model is denoted by $m$, followed by the level and the assumption for the intercept, either $f$ or $r$ and then the assumption for the treatment impacts, $c$, $f$, or $r$. For example, `m3ff2rc` means at level 3, we assume fixed intercepts and treatment impacts, and at level 2

we assume random intercepts and constant treatment impacts. The full design and model are specified by concatenating these together, e.g. `d3.2_m3ff2rc`.

The full list of supported models is below. We also including the corresponding names from the PowerUP! package where appropriate. For more details about each model, see the appendix.

| Code | Design | Model | PowerUp |
|---|---|---|---|
| d1.1_m2cc | d1.1 | m2cc | n/a |
| d2.1_m2fc | d2.1 | m2fc | blocked_i1_2c |
| d2.1_m2ff | d2.1 | m2ff | blocked_i1_2f |
| d2.1_m2fr | d2.1 | m2fr | blocked_i1_2r |
| d3.1_m3rr2rr | d3.1 | m3rr2rr | blocked_i1_3r |
| d2.2_m2rc | d2.2 | m2rc | simple_2c_2r |
| d3.3_m3rc2rc | d3.3 | m3rc2rc | simple_c3_3r |
| d3.2_m3ff2rc | d3.2 | m3ff2rc | blocked_c2_3f |
| d3.2_m3rr2rc | d3.2 | m3rr2rc | blocked_c2_3r |

## Estimating MDES and sample size

Frequently, a researcher's main concern with power is prospectively calculating either the minimum detectable effect size (MDES) from a possible study, or determining the necessary sample size. With closed-form power expressions, it is easy to invert the formula to instead calculate these quantities. However, given our simulation approach, we instead perform a search algorithm to calculate MDES and sample size.

The user provides a particular target power, say 80%. To perform a search, we calculate power over a range of different MDES or sample size values, and then find the value that is within a specified tolerance of the target power. We discuss the algorithm for sample size, although the approach for MDES is the same.

First, we begin by bounding the possible range of sample size values. Bonferroni is the most conservative correction, so it would result in the largest possible sample size, and thus a Bonferroni correction provides our upper bound. If we are interested in complete power, we have a larger upper bound; in order to have complete power of 0.8, we would need each outcome to have an individual power of $0.8^{(1/M)}$. On the other hand, our least conservative correction is doing no correction at all, and would result in the smallest possible sample size, so no adjustment provides our lower bound. If we are interested in minimal power, we must have a smaller lower bound; in order to have 1-minimal power of 0.8, each outcome needs to have individual power of $1 - (1 - 0.8)^{(1/M)}$.

Second, given our lower and upper bounds, we now use optimization to find the sample size. **TODO** this will be updated once we have settled on an optmization procedure!

## Validation

We completed extensive validation checks to ensure our power calculation procedure was correct. First, we compared our power estimates in scenarios with only one outcome, $M = 1$, to PowerUpR!. Without a multiple testing procedure adjustment, our estimates match. Second, in order to validate our estimates under multiplicity, we used a simulation approach. We generated many iterations of data according to the assumed design and model, calculated p-values, and calculated an empirical estimate of power and error. Then, we compared the PUMP results to validate that the PUMP estimates fall in the confidence interval for the simulations.

A more detailed explanation of the validation procedure can be found in the Appendix, and full validation code and results are in our github repository **TODO**. For some scenarios, we have discrepancies from PowerUp resulting from different modeling choices. For example, for certain models PowerUp assumes the intraclass correlation is zero, while we allow for nonzero values. When there are discrepancies, these are noted in the appendix.

**TODO** update the appendix.

## PUM-P Package (Luke - first October 1)

- Overview of power, mdes, sample size and grid functions + plotting function
- Example(s) - LUKE'S VIGNETTE

## Guidance for practice

- Reflections on issues that come up in vignette
  - e.g., how to come up with correlation assumption - what's needed by future researchers so can come to our tool with good assumptions
  - e.g., reinformce importance of looking at ranges when it matters a lot

## Appendices

- specifications/derivations for all designs
- how data were generated for simulations
- something about validation maybe? (one example or template?)

## References

Dong, Nianbo, and Rebecca Maynard. 2013. "PowerUP!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies." Journal Article. *Journal of Research on Educational Effectiveness* 6 (1): 24–67.