Example Languages from Grants

In the following we see three snippets from three different (funded) grant proposals. The accompanying script "grant_example_code.R" has code to conduct (most of) the listed power analyses.

Disclaimer: These examples do not have multiple testing correction as they are all focused on single outcomes. I have added a fourth example, that modifies the first example, to mimic what I would hope to see for such a case.

Grant Example 1

This is a grant for what is effectively an individually randomized trial, spread across several campuses. They capture the level 2 with fixed effects, and wrap that into the R2 at level 1 rather than specifying an ICC. The results would be the same either way.

Text in the body of the grant

Minimum Detectable Effects

The power calculations below are described in terms of the minimum detectable effect (MDE), minimum detectable effect size (MDES, or MDE divided by the control group standard deviation), and the minimum detectable τ (MDτ, described later). Intuitively, the MDE is "the smallest true impact that an experiment has a good chance of detecting" (Bloom, 1995). The smaller the MDE, the more likely a study will be able to detect impacts of a small magnitude.

Power for Overall Effects (RQ 1.a): Exhibit C.9 provides a table, the assumptions, and formulas for the MDE(S) calculations. Here, we provide an example. Consider SUCCESS's effect on completing a credential within three years. With N = 4,500, the evaluation's total sample size goal, and the proportion of control group members graduating (π) equal to 0.20 **the MDE for three-year graduation rates is 2.9 percentage points**.[1] If SUCCESS's true effect on three-year graduation rates is 2.9 percentage points, then there is an 80 percent chance of finding positive and statistically significant effects at the 10 percent significance level.

Power for Effects for Subgroups (RQ 1.b): The MDE depends on group size. Based on historical data, the sample is expected to include approximately 29 percent Black students, 38 percent Hispanic students, and 21 percent White students. For the smallest subgroup ($N_{White} \approx$ 945), the MDE for three-year graduation rates is around 6 percentage points. For the other confirmatory subgroup, we do not yet have estimates of the expected sample size.

Power for Cross-College Variation in Effects (RQ 1.c): First, consider the MDE at an individual college. Most colleges are aiming to recruit 450 students, yielding an MDE of 9 percentage points on graduation rates, assuming the estimator described in equation (1). Since the preferred Empirical-Bayes (E-B) estimator partially pools data from all colleges, the actual MDE(S) will be smaller using the more precise E-B estimator.

Another consideration regarding RQ 1.c is the minimum detectable τ (MDτ), or the smallest amount of *true* cross-site variation in effects that the design will be well-powered to detect. Following (Bloom & Spybrook, 2017), we estimate the MDτ in effect size units to be 0.12σ. In percentage points, **the MDτ is around 4.7 percentage points**.

[1] The national average three-year graduation rate at community colleges is 25 percent. The average three-year graduation rate at the community colleges in this study is under 20 percent.

Text in an appendix of the grant

$$MDE(Binary) = 2.49 * \sqrt{\frac{\pi(1-\pi)(1-r^2)}{T(1-T)N}} \text{ or } MDE(Continuous) = 2.49 * \sqrt{\frac{\sigma^2(1-r^2)}{T(1-T)N}}, \qquad (1)$$

where:

2.49 = the appropriate multiplier for 80 percent power and a 10 percent significance level with a two tailed hypothesis test[2]

π = the proportion of the study population that would have a value of 1 for a binary outcome in the absence of the program (various possible values are shown in the table).

r^2 = the explanatory power of the regression, which includes the random assignment block indicators as well as other covariates (assumed = 0.10 based on Adelman, 2006; Gates & Creamer, 1984; and Weiss et al, 2017).

T = the proportion of the sample that is randomly assigned to the treatment group (assumed to be 0.58)

N = the total size of the study sample

σ = the standard deviation of a continuous outcome.

The MDES for continuous outcomes can be calculated by dividing the MDE by the control group standard deviation, and can be expressed using the equation below. Notably, the MDES for a binary outcome is the same as for a continuous outcome.

$$MDES = 2.49 * \sqrt{\frac{\sigma^2(1-r^2)}{T(1-T)n}}/\sigma = 2.49 * \sqrt{\frac{(1-r^2)}{T(1-T)n}} \qquad (2)$$

Adapting from Bloom & Spybrook (2017), the following formulas were used to calculate the minimum detectable τ (MDτ) and the minimum detectable τ in effect size units (MDτ-ES).

$$MD\tau\_ES = 2.49 * \sqrt{\left(\frac{(1-\rho_C)\left(1-r^2_{C(within)}\right)}{T(1-T)\bar{N}}\right)\left(\frac{F_{0.90}}{F_{0.20}}-1\right)} \qquad (3)$$

and

$$MD\tau = 2.49 * \sqrt{\left(\frac{(1-\rho_C)\left(1-r^2_{C(within)}\right)}{T(1-T)\bar{N}}\right)\left(\frac{F_{0.90}}{F_{0.20}}-1\right)} * \sigma, \qquad (4)$$

where:

ρ_C = the proportion of the outcome variance explained by the random assignment block indicators (assumed to be 0.09, based on Weiss et al. (2017))

$r^2_{C(within)}$ = the proportion of the within block outcome variance explained by covariates. (assumed to be 0.03, based on Weiss et al. (2017))

$F_{0.90}$ = the 90th percentile value of an *F* distribution with $J-1$ numerator degrees of freedom and $J(\bar{N}_j-2)-K$ denominator degrees of freedom.

$F_{0.20}$ = the 20th percentile value of an *F* distribution with $J-1$ numerator degrees of freedom and $J(\bar{N}_j-2)-K$ denominator degrees of freedom.

[2] This "multiplier" is a good approximation when the number of degrees of freedom is greater than 30. Use of this multiplier is similar to the use of 1.645 * standard deviation for creating a 90 percent confidence interval.

$\overline{N} =$ the harmonic mean site sample size (assumed to be 450, reflecting the target for this study).

Grant Example 2

This grant is a straightforward RCT where schools were randomized into treatment or control. They account for attrition by having smaller sample sizes estimated for the classrooms. In the original grant, they let classroom variation be represented by the ICC at the school level. In the code script, a three-level version of the power analysis is also given for comparison.

Test of the Grant

Power Considerations
We used Optimal Design for Multi-level and Longitudinal Research (Raudenbush et al., 2011) to estimate power. In these estimates, we assumed that we will recruit 48 schools from two districts, with each school having an average of 3 classrooms, and that we will test 10 children from each classroom initially and retain a minimum of 6 children per classroom (18 per school) through the course of the study. Separate estimates were made for the growth models and the model testing program effects on second grade reading comprehension.

Growth models. We propose a 3-year study (D = 3) with one assessment in the spring of each year (F = 1) plus a baseline assessment for a total of 4 occasions of measurement (M). Assuming an intraclass correlation (ρ) of .08 or .10, variability of level-1 residual ($\sigma^2 = 1$), and variability of level-1 coefficient (τ) = 1, the minimal detectable effect size for the linear growth parameter is .30-.33 (see Appendix B6). This seems easily attainable for the researcher-developed measure of vocabulary and at least plausible for the standardized measures of vocabulary and listening comprehension based on previous research (Apthorp et al, 2012; Coyne et al., 2010; Elleman et al .2009; Gonzalez et al., 2011).

Impact on reading comprehension at second grade. Reading comprehension scores as measured by standardized tests may be difficult to move by more than .25 standard deviations, even with interventions lasting two or more years (e.g., Borman et al., 2007; Connor et al., 2011; James-Burdumy et al., 2012; Reis, McCoach, Little, Muller, & Kaniskan, 2011; Savage, Abrami, Hipps, & Deault, 2009; Vaughn et al., 2008). However, we are studying the cumulative impact of the program across three years and suggest growth of .083 standard deviations per year as a “minimum relevant effect size” (Dong & Maynard, 2013) for policy or practice. In first and second grade, children’s growth on standardized tests of reading achievement is about 1.0 standard deviation (Bloom, Hill, Black, & Lipsey, 2008), so an effect size of .08 is about 1/12 of a year or one month of learning, and also equivalent to raising a child from the 50th to the 53rd percentile. Thus we wish to detect an effect size of 0.25 SD by second grade.

We determined the minimum detectable effect size (MDES) for reading comprehension under four sets of assumptions: (a) Intraclass correlation (ICC) = .08 and proportion of variance accounted for by the school-level covariates (R^2) = 0.80; (b) ICC = 0.08 and R^2 = 0.60; (c) ICC = 0.10 and R^2 = 0.80, and (d) ICC = 0.10 and R^2 = 0.60. Analyzing Reading First reading comprehension data from 14 districts and a 1 state with a two-level model omitting the classroom level, Jacob, Zhu, & Bloom (2010) found school ICCs ranging from .02 to .21 with a median value of .08. Also using a two-level model, Zhu et al. (2011) found school ICCs ranging from .06 to .11 for reading comprehension in grades 1 through 5 for four large data sets. Zhu et al. (2011) found R^2 for school-level pretest covariates that ranged from .58 to .83 for the same elementary reading data sets. We used these empirical ICC and R-squared estimates to define a range of assumptions, from .08 to .10 for school ICC and from .60 to .80 for R-squared. We also assumed that the school samples will have been reduced to 18 by attrition (the worst-case scenario). We found that the MDES for power .80 was between .22 and .25 (see Appendix B7). Thus, the hypothesized cumulative impact for reading comprehension (d = .25) is equal to the MDES even in the fourth and worst case depicted.

Grant Example 3

This grant is a multisite (school) cluster (classroom) randomized experiment. They are allowing for treatment variation across the schools.

Grant Text

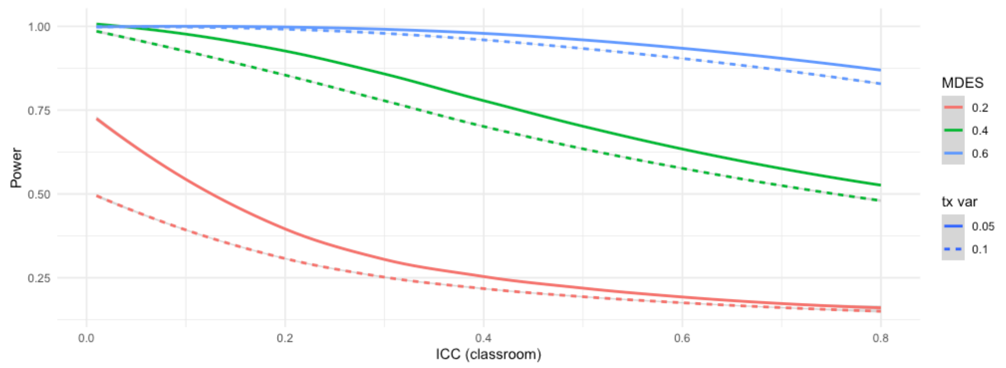
Power Analysis: In our power analysis, we fix the sample sizes: $n \cong 35$ students per classroom, $J = 8$ classrooms per school, and $K = 14$ schools. Estimates of the three parameters—the effect size (δ), the variability in effect size ($\sigma2$), and the intra-class correlation (ρ)—are based on the research literature and pilot evaluations of the Pathway program. We computed an effect size of .40 standard deviation (SD) units from pilot studies of the Pathway program. Therefore, our power analysis is based on a range of possible effect sizes, where effect sizes of .20, .40, and .60 SD units correspond to small, medium, and large estimate effects. For each of the three effect sizes (.20, .40, .60 SD), we estimate an effect size variability of .05 and .10. While Raudenbush and Liu also include .15 as a possible measure of the effect size variability, we do not think it will be this large. Over time, we expect to find less variability in the delivery of instruction among Pathway teachers. Such efforts should improve the consistency of program delivery across classrooms, especially in year 2 and 3 of the study, and reduce the variability in student achievement. In fact, the standard deviation of the gain scores on the Assessment of Literary Analysis has grown smaller over time from 2.63 (1996-97), to 2.56 (1998-1999), to 1.96 (200001).

Estimates of the intra-class correlation often vary in the research literature, depending on the sample used for analyses. For instance, Agodini et al. (2003) conducted a power analysis for experiments involving technology, using data from the Longitudinal Evaluation of School Change and Performance (LESCP), and found that approximately 10% of the variation in outcomes lies between classrooms. In our analyses, we have plausible values for all the parameters except the intra-class correlation, which varies from very small to very large estimates in our power analysis. To compute a power analysis involving classrooms in multiple school sites, we used the “Optimal Design” software developed by Raudenbush, Spybrook, Liu, and Congdon (2004). Table A.1 below shows that we have power of at least .80 under four scenarios. With a large-sized effect of .60 SD units, we have power at or above .80 even if the intra-class correlation is extremely high ($\rho = .80$). We still have power of .80 with moderate sized impacts, and the intra-class correlations of .36 and .27 are somewhat higher than what has been observed in prior research. Even a small effect size of .20 with an estimated variance of .05 has moderate power (.74) to detect significant program impacts if the intra-class correlation is very low ($\rho = .01$). In general, power increases as the intra-class correlation decreases (see Figure below). To improve the precision of our treatment effects, we will also include baseline covariates in our analyses.

They then had a table and figure that looked like this:

MDES	**ICC.2**	**Sigma2**	**D1indiv**
0.2	0.01	0.05	0.7547
0.2	0.01	0.10	0.5082
0.4	0.36	0.05	0.8114
0.4	0.36	0.10	0.7301
0.6	0.80	0.05	0.8704
0.6	0.80	0.10	0.8334

And



Note: The PUMP package models impact variation as a ratio of variation to the amount of variation at that level. In the grant, they did not mention school-level variation, but for fixed effects at the school level, this variation would be minimal. We therefore set omega to 1 and let the ICC.3 vary across 0.05 and 0.10. The numbers are quite close to the reported grant.

Revisiting Example 1 with multiple outcomes

Here is some sample grant language that accounts for multiple testing.

Minimum Detectable Effects

The power calculations below are described in terms of the minimum detectable effect (MDE), minimum detectable effect size (MDES, or MDE divided by the control group standard deviation), and the minimum detectable τ (MDτ, described later). Intuitively, the MDE is "the smallest true impact that an experiment has a good chance of detecting" (Bloom, 1995). The smaller the MDE, the more likely a study will be able to detect impacts of a small magnitude.

Power for Overall Effects: We have four outcomes of interest, and plan on using the Holm correction to account for multiple testing and control the familywise error rate. With N = 4,500, the evaluation's total sample size goal, we have 80% power to find significance on at least 2 of our 4 outcomes, after multiple testing correction, and assuming all outcomes have a true impact of at least 0.07 (this is the min-2 power MDES). If only two outcomes are impacted by treatment, the MDES for those two outcomes would rise to 0.094. If the true impacts are at least 0.064 for all four outcomes, we would have 80% power to detect at least one of them, after multiple-testing correction. These calculations show that as long as the intervention is effective with respect to at least two outcomes, we will detect some effects.

For an example of the power to detect specific impacts for specific outcomes, we give an example of the MDE for SUCCESS's effect on completing a credential within three years (with the proportion of control group members graduating (π) equal to 0.20). **The MDE for detecting three-year graduation rates is about 3 percentage points, after doing a Holm's correction for our 4 outcomes.**[1] This means that if SUCCESS's true effect on three-year graduation rates were 3 percentage points, then there is an 80 percent chance of finding positive and statistically significant effects at the 10 percent significance level, after correction.

[1] The national average three-year graduation rate at community colleges is 25 percent. The average three-year graduation rate at the community colleges in this study is under 20 percent.