PENNSTATE

The Methodology Center
advancing methods, improving health

FactorialPowerPlan
SAS Macro Users' Guide
Version 1.0

John J. Dziak
Linda M. Collins
Aaron T. Wagner

Please send questions and comments to *MChelpdesk@psu.edu*.

The development of the SAS %FactorialPowerPlan macro was supported by National Institute on Drug Abuse Grant P50-DA10075 to the Center for Prevention and Treatment Methodology.

The suggested citation for this users' guide is

Dziak, J. J., Collins, L. M., & Wagner, A. T. (2013). *FactorialPowerPlan SAS macro suite users' guide* (Version 1.0). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Table of Contents

1. Overview of the %FactorialPowerPlan Macro

Factorial experiments and fractional factorial experiments can have many advantages in the social and behavioral sciences (see Collins, Dziak, & Li, 2009). However, as with all experiments, they require planning, especially for choosing an adequate number of participants (also known as sample size or *N*). Researchers try to find an *N* that is large enough to offer a high probability of a useful result while not being so large as to waste money or other resources. The %FactorialPowerPlan macro is intended to help with this process.

This users' guide describes how to use the %FactorialPowerPlan macro. The macro can be used to do sample size and power calculations for planning either a factorial or fractional factorial experiment. The calculations can be done for either posttest-only or pretest-posttest designs. Participants can either be assumed to be independent, or nested within existing clusters as discussed in Dziak, Nahum-Shani, and Collins (2012). The factors (**independent variables) are assumed to be dichotomous** (each has only two levels) and the **outcome variable is assumed to be normally distributed.**

This guide assumes you have some knowledge of factorial and fractional factorial experiments. Collins et al. (2009) provides an introduction to factorial experiments for audiences in the social, behavioral, prevention, and intervention sciences. Dziak et al. (2012) provides further information on the implications of nesting within clusters for factorial experiments.

1.1 %FactorialPowerPlan Macro Features

- Facto

The macro can calculate the effect of the following special conditions on power:

- clustered samples, with treatment assignment at either the individual or the cluster level;
- an individual-level pretest, or no pretest; and
- different alpha levels from 0 to .5 (e.g., .01, .05).

The experiment may be either a complete factorial or fractional factorial design. Power for main effects or for interactions can be calculated. However, as is always the case in sample size and power planning, the **accuracy of the macro's predictions is of course dependent on a model and on the values of various user-provided parameters that represent the user's assumptions**. The parameters required are described later in this manual.

1.2 %FactorialPowerPlan Macro Requirements

The macro requires SAS version 9.1 or higher for Windows. It requires SAS/STAT and SAS/IML procedures in order to perform the calculations, but most university licenses include

these, and most users will have them automatically as part of their SAS installation. Therefore, the only thing that most users will need to do is to download the macro and then direct SAS to read it using an INCLUDE statement, as described in section 5.1 of this manual.

2. Factorial Experiments and Effect Coding

The macro assumes that there are *K* **dichotomous** factors. (The current version of the macro does not accommodate factors with more than two levels.) Thus, using effect coding, we can express the levels of any given factor as +1 and -1 (see Myers & Well, 2003). For example, +1 might indicate "on" and -1 might indicate "off," or +1 might indicate "high" and -1 might indicate "low." These numbers are just arbitrary labels to show that they are opposite sides of a contrast defined by the factor of interest. The positive and negative signs do not represent any expectation that one level is "good" or that the other is "bad."

Given this coding, we can represent any of the 2^K treatment conditions or "cells" as a list of *K* numbers. For example, if there are three factors, and subject *i* is assigned to the high level of the first and the low level of the other two, then subject *i* is in the cell (+1,-1,-1). Thus $x_{1i} = +1$, $x_{2i} = -1$, $x_{3i} = -1$. We can also treat each factor as a regression variable and represent their interactions as products. This gives us a way to represent any subject's expected value (i.e., the estimated population mean of the cell this subject is in) in terms of the factors. For example, if there are three factors then the expected value μ_i of the response y_i from respondent i is

$$\mu_i = E(y_i) = \begin{array}{c} \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \\ \beta_{1,2} x_{1i} x_{2i} + \beta_{1,3} x_{1i} x_{3i} \\ +\beta_{2,3} x_{2i} x_{3i} + \beta_{1,2,3} x_{1i} x_{2i} x_{3i} \end{array} \qquad (1)$$

It is possible to simplify the model by assuming that interactions above a certain amount of complexity (e.g., three-way interactions) are negligible.

2.1 Effect Size Measures for Main Effects

The main effect of any factor k ($1 \le k \le K$) can be defined as the difference between the average[1] responses for the two levels of the factor; that is, $ME_k = \mu_{(x_k=+1)} - \mu_{(x_k=-1)}$ (see Myers & Well, 2003). Because of effect coding, the other terms in (1) cancel out when we take the difference and we get

[1] We assume in this macro that the number of subjects in each cell are equal, so we do not have to worry about whether $\mu_{(x_k=+1)}$ is the average of all *subjects* with $x_k = +1$ or the average of all *cells* with $x_k = +1$. If cell sizes were unequal, however, it is more sensible to define the main effect as the average of all cells, i.e., the weighted average of all subjects. In that case, all of our effect size formulas still make sense (i.e., expression (1) still implies expression (2).). NOTE: power may be lower due to imbalance.

$$\begin{aligned} ME_k &= \mu_{(x_k=+1)} - \mu_{(x_k=-1)} \\ &= +1\beta_k - (-1\beta_k) \\ &= 2\beta_k \end{aligned} \qquad (2)$$

If we assume subjects each have response standard deviations σ_y, then we could also calculate the standardized effect size $ME_k^{std} = ME_k/\sigma_y = 2\beta_k/\sigma_y$, which would be similar to d as used by Cohen (1988) for t-tests. We could also calculate a standardized regression coefficient $\beta_k^{std} = \beta_k/\sigma_y = 2d$. Lastly, for a balanced design, it can be shown algebraically that Cohen's (1988) signal to noise ratio f^2 (also known as η^2), the ratio of the variance accounted for by the effect of interest to the error variance, is equal to β_k^2/σ_y^2 and to $\left(\beta_k^{std}\right)^2$. Any of these quantities could be considered a measure of the effect size of factor k. However, you need to specify only one of them.

2.2 Effect Size Measures for Two-Way Interactions

The interaction between factor a and factor b is sometimes defined as a difference between differences, as follows:

$$\begin{aligned} I \times N_{a,b} &= \left(\mu_{(x_a=+1,x_b=+1)} - \mu_{(x_a=-1,x_b=+1)}\right) - \left(\mu_{(x_a=+1,x_b=-1)} - \mu_{(x_a=-1,x_b=-1)}\right) \\ &= \left(\beta_{a,b} - \left(-\beta_{a,b}\right)\right) - \left(\left(-\beta_{a,b}\right) - \beta_{a,b}\right) = 4\beta_{a,b} \end{aligned}$$

However, other sources define it as half the difference of differences:

$$\begin{aligned} I \times N_{a,b}^* &= \frac{1}{2}\left(\left(\mu_{(x_a=+1,x_b=+1)} - \mu_{(x_a=-1,x_b=+1)}\right) - \left(\mu_{(x_a=+1,x_b=-1)} - \mu_{(x_a=-1,x_b=-1)}\right)\right) \\ &= \frac{1}{2}\left(\beta_{a,b} - \left(-\beta_{a,b}\right)\right) - \left(\left(-\beta_{a,b}\right) - \beta_{a,b}\right) = 2\beta_{a,b} \end{aligned}$$

To prevent confusion that could be caused by this inconsistency in the literature, we require that the interaction be specified as a regression coefficient. It could either be the unstandardized regression coefficient

$$\beta_{a,b} = \frac{1}{4}\left(\left(\mu_{(x_a=+1,x_b=+1)} - \mu_{(x_a=-1,x_b=+1)}\right) - \left(\mu_{(x_a=+1,x_b=-1)} - \mu_{(x_a=-1,x_b=-1)}\right)\right)$$

the standardized regression coefficient

$$\beta_{a,b}^{std} = \beta_{a,b}/\sigma_y,$$

or the signal-to-noise ratio, which is the square of the standardized regression coefficient.

You could also specify regression coefficients for higher-order interactions if desired. For example, the coefficient for the interaction $\beta_{a,b,c}$ is

$$\frac{1}{8}\left(\begin{array}{l}\left(\left(\mu_{(x_a=+1,x_b=+1,x_c=+1)}-\mu_{(x_a=-1,x_b=+1,x_c=+1)}\right)-\left(\mu_{(x_a=+1,x_b=-1,x_c=+1)}-\mu_{(x_a=-1,x_b=-1,x_c=+1)}\right)\right)\\-\left(\left(\mu_{(x_a=+1,x_b=+1,x_c=-1)}-\mu_{(x_a=-1,x_b=+1,x_c=-1)}\right)-\left(\mu_{(x_a=+1,x_b=-1,x_c=-1)}-\mu_{(x_a=-1,x_b=-1,x_c=-1)}\right)\right)\end{array}\right)$$

When providing a regression coefficient, you do not have to specify whether it represents a main effect or an interaction, nor specify the kind of interaction. That is because this does not matter for power in a balanced design being analyzed with effect coding; all that matters is the size of the regression coefficient.

3. Macro Assumptions

The field of factorial experiments is very broad. Therefore, some assumptions about the kind of experiment being considered were necessary in implementing %FactorialPowerPlan.

3.1 Randomized Factorial Experiment With Dichotomous Factors

The %FactorialPowerPlan macro assumes a **randomized factorial experiment** with K factors, $K \geq 1$ and $K \leq 99$. For most users, K will probably be between 2 and 8. Each factor is assumed to be **dichotomous** (having only two levels). Thus, the total number of possible "cells" (experimental conditions defined by the factors) is $2 \times 2 \times ... \times 2$, or 2^K.

It is assumed that the K factors are conceptually independent, so that each of these 2^K cells is possible to implement, at least in theory. For example, an experiment in which one of the "factors" is medication versus no medication, and the other "factor" is low dose versus high dose, is not allowed. This would not really be a factorial experiment, because it is impossible to interpret a difference between a low dose of nothing and a high dose of nothing. Such an experiment makes perfect sense, but rather than a factorial experiment, it would be more logically viewed as an experiment involving a single three-level factor with levels none, low, and high (see Collins et al., 2009). The %FactorialPowerPlan macro does not apply to this kind of experiment.

3.2 Balanced Assignment

Assignment of participants to experimental conditions is assumed to be **balanced**. That is, the same (or essentially the same) number of participants is assumed to be assigned to each cell. If this does not hold then the experiment may be somewhat less efficient than expected and power will be somewhat lower.

3.3 Effects Assumed to be Negligible in Fractional Factorial Experiments are Actually Negligible

The experiment may be **either complete factorial or fractional factorial**. Which of these it is, however, does not have to be specified when using the macro. This is because a fractional factorial has exactly the same theoretical power as the corresponding complete factorial for the tests of the effects it includes, as long as all of the interactions which are assumed to be negligible in order to design the fractional factorial are in fact negligible (see Collins et al., 2009). If a fractional factorial experiment is done but the assumptions behind it are not satisfied (i.e., an effect of interest is aliased with a non-negligible interaction), power may be either lower or spuriously higher, and Type I error may also be higher than expected. However, aliasing is not considered by the %FactorialPowerPlan macro, because it would be very

difficult to predict the exact effect on power that a violation of assumptions would have, since so many different effects are possible.

3.4 Response Variable is Normally Distributed

It is also assumed that the **response variable is normally distributed**, with the same error variance in all cells, and no "floor" or "ceiling" effects caused by a restricted range of possible outcomes. These technical assumptions are required for the calculations, but they are unlikely to hold exactly in practice. The extent to which they may be violated without causing important changes in power is a complicated question that deserves further research.

3.5 Power Assumptions

Last, because "power" can mean different things in different contexts, it is useful to clarify what our macro calculates.

1. The macro calculations assume that the user is not trying to reduce experimentwise Type I error rate among the main effects and interactions being tested by using a Bonferroni or other multiple comparison correction. We do not recommend that such a correction be done for planned tests of main effects or interactions in a factorial design.
2. Likewise, the macro estimates the probability of finding a given factor significant given that its true effects are nonzero, not the probability of finding each of them significant given that their true effects are all nonzero. That is, we define power as one minus Type II error rate for a single factor, not the experimentwise Type II error rate. Unlike experimentwise Type I error, experimentwise Type II error has received little attention in the literature.
3. Also, power is *not* being predicted for an omnibus or overall test (all effects zero versus at least one nonzero), but instead for each test of a single effect (this effect zero versus this effect nonzero).
4. Finally, power is not being calculated for pairwise contrasts; that is, comparisons of individual cells with other individual cells with or without a multiple comparison correction. It has been argued (e.g., McAlister, Straus, Sackett, & Altman, 2003) that inference about comparing individual cell means is important. This is often true, especially in a confirmatory experiment with only two or three factors. However, powering contrasts among all 2^k cells becomes infeasible in the context of an experiment with more than two or three factors, as in a factor-screening experiment (see Collins et al., 2009). In other words, it is not helpful to think of a 2^5 factorial as

simply a giant randomized controlled trial with 32 arms (Dziak et al., 2012); it is much more feasible in that case to focus on main effects and some of the simpler interactions.

4. Optional Design Features

The %FactorialPowerPlan macro can also take into account the effects of clustered assignment and/or the presence of a pretest. We describe these below one at a time.

4.1 Clustering

As described by Dziak et al., (2012), sometimes subjects are not independent because they are nested within natural clusters (schools, clinics, etc.). Power is different depending on whether treatment is assigned at the individual level ("within-clusters" assignment) or the cluster level ("between-clusters" assignment). In the case of within-clusters assignment, participants are individually randomly assigned, and therefore not everyone in a cluster will have the same treatment. For example, different patients at the same clinic can independently receive different pills. In the case of between-clusters assignment, the randomassignment is done for the cluster as a whole, and as a result everyone within a given cluster will receive the same treatment as everyone else in that cluster. For example, all students within a given classroom must receive the same educational curriculum. The %FactorialPowerPlan macro calculates power for either kind of clustered experiment, using the Dziak et al. approach.

In theory, other designs are possible which would combine features of between- and within-clusters assignment. For example, suppose that certain treatment conditions are unavailable in a given cluster, but that within this constraint individuals are otherwise assigned randomly. Our macro currently does not compute power for more complicated situations like this one. Instead, in the within-cluster case, we assume for simplicity that all conditions are potentially available to all participants.

%FactorialPowerPlan assumes that **there are no interactions between cluster effects and treatment**; that is, that factor effects are the same for each cluster even though there may be random cluster effects. This assumption may be too simplistic (see Raudenbush & Liu, 2000), but further research is needed about how to relax it when there are multiple factors.

4.2 Pretest

A pretest can often improve the power of an experiment. A pretest may be used in different ways, reflecting different assumptions and potentially having different effects on power. Specifically, there are different power formulas depending on whether the pretest is adjusted for as a regression covariate as in analysis of covariance (ANCOVA) or entered into a multilevel model as a repeated measurement (see Allison, 1990; Murray, 1998).

In covariate-adjusted analysis or ANCOVA the main effects and interactions of interest are no longer defined in terms of differences between cells in the mean raw responses, but in the mean adjusted responses. In the repeated measures analysis, the main effects and interactions of interest are defined in terms of differences between conditions in the mean change scores (posttest minus pretest; see Murray 1998, p. 181, who called this approach the "time×condition" analysis). In many situations with large samples and random assignment, the two methods can give almost identical performance (as in the simulations in Dziak et al., 2012). However, they reflect subtly different assumptions (see Allison, 1990) and can have very different performance in some situations (see Janega et al., 2004; Murray, 1998; Murray & Blitstein, 2003). When assignment is not random, ANCOVA and repeated measures approaches can give substantively very different answers, and they have different potential biases. However, **for this macro we assume random assignment,** and so the main concern is not bias but power. According to existing power formulas, the repeated measures approach is better than having no pretest at all only when the pretest-posttest correlation is greater than 0.5, and is worse than no pretest if the pretest-posttest correlation is less than 0.5. Also, according to power formulas, ANCOVA is at best much more powerful and at worst not much less powerful, than having no pretest or using repeated measures. Thus, one could argue that ANCOVA is the best choice. However, an opposing argument is that this power advantage may sometimes be an artifact of the ANCOVA power formula not accounting for measurement error in the pretest. More information is given in Frison and Pocock (1992), Vickers (2001), Oakes and Feldman (2001), and Smolkowski (2010). Because the situation is so complicated, %FactorialPowerPlan offers either kind of model for the pretest, so it is the user's choice. The exception is that for between-clusters assignment, the pretest-as-covariate option is not allowed because power is difficult to predict in this setting (Murray, 1998; Dziak et al., 2012). Thus, the following combinations of options are available:

	No pretest	**Pretest as covariate**	**Pretest as repeated measure**
No clustering	✓	✓	✓
Within-clusters assignment	✓	✓	✓
Between-clusters assignment	✓	*Not provided*	✓

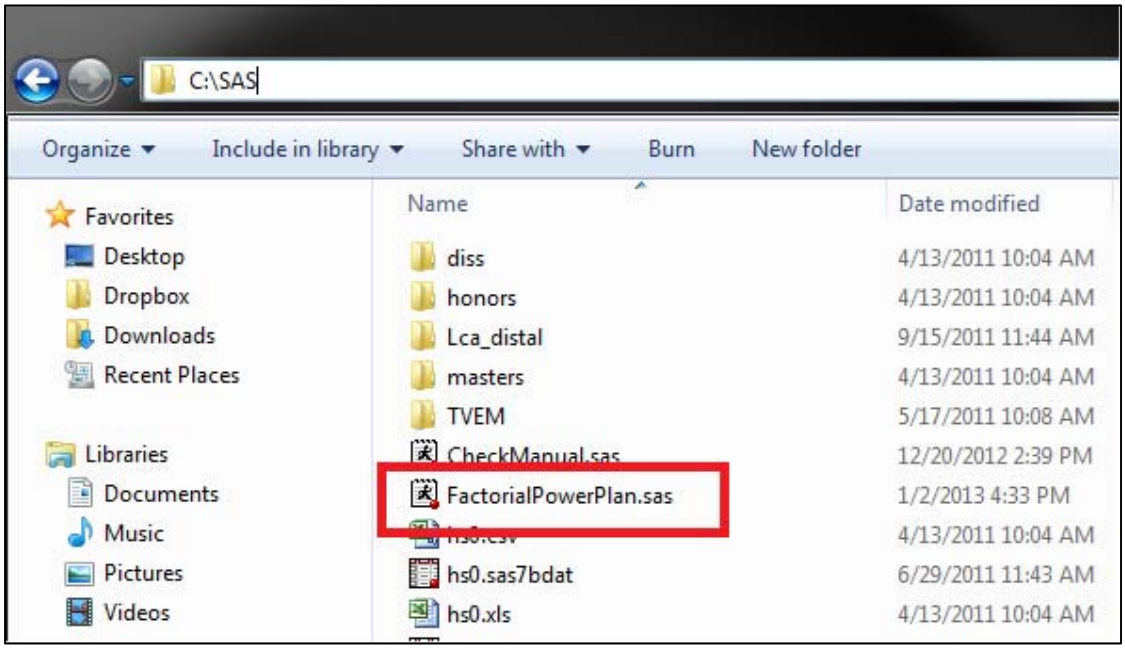
5. Preparing to Use the %FactorialPowerPlan Macro

5.1 Loading and Running the Macro

A SAS macro is a special block of SAS commands. First the block is defined, and then it is called when needed. Two steps need to be completed before running %FactorialPowerPlan. First, if you haven't already done so, download and save the macro to the designated path (e.g., *S:\myfolder*). Second, direct SAS to read the macro code from the path, using a SAS %INCLUDE statement. For example, suppose that the SAS macro file FactorialPowerPlan.sas is saved in the folder "`S:\myfolder\` ". Then the include statement would be

```
%INCLUDE "S:\myfolder\FactorialPowerPlan.sas";
```

You must replace "`S:\myfolder\`" with the location where you downloaded and saved the macro.



If this is the location of the macro, the include statement would be
%INCLUDE "C:\SAS\FactorialPowerPlan.sas";

5.2 Specifying Design Information

There are three ways to use %FactorialPowerPlan.

1. *Estimate power* available from a given sample size and a given effect size.
2. *Estimate sample size* needed for a given power and a given effect size.
3. *Estimate effect size* detectable from a given power at a given sample size.

There are three main pieces of information: power, sample size, and effect size. The user provides two of them, and the macro calculates the third.

If the participants are independent, then "sample size" is the number of participants. If the participants are clustered, "Sample size" is measured differently. In this case, we assume that the average number of participants per cluster is given, and then we calculate the needed number of clusters. For example, if an investigator is randomly assigning classrooms and believes that classrooms have about 20 students on average, he/she may need to know how many classrooms are needed. Of course, the total number of participants could then be estimated as the product of the number of clusters and the mean number of members per cluster (e.g., 30 classrooms x 20 students/ classroom = 600 students).

The exact usage of the macro is slightly different depending on which of the three pieces of information listed above is being calculated, and on whether or not the participants are assumed to be nested within clusters. This defines **3** (power, *N*, effect size) **× 2** (clustered, unclustered) **= 6** main ways to use the macro. After providing a brief overview of the macro syntax, we then provide six sections considering each of these possible use scenarios in turn.

6. Syntax

To run the macro, you should specify information about your model assumptions as "arguments" or inputs in a list of key words. Upper or lower case is not important. The order of the arguments is also not important, because the arguments are distinguished by their names. The possible arguments are listed in the following table. You do not have to use all of the arguments. No single argument is always required, although at least some information must be provided. Examples for using the code are given in the following section.

Argument		Value
alpha		Two-sided Type I error level for the test to be performed (default=0.05)
nfactors		The number of factors (independent variables) in the planned experiment (default=1)
sigma_y		The assumed standard deviation σ_y of the response variable after treatment, within each treatment condition (i.e., adjusting for treatment but not adjusting for post-test). This statement must be used if the effect size argument used is either "raw_main" or "raw_coef"
model_order		The highest order term to be included in the regression model in the planned analysis (1=main effects, 2=two-way interactions, 3=three-way interactions, etc.); must be ≥ 1 and ≤ nfactors. (default=1)
Pretest items	pretest	One of three options: • "no" or "none" for no pretest • "covariate" for pretest to be entered as a covariate in the model • "repeated" for pretest to be considered as a repeated measure The option "yes" is also allowed and is interpreted as "repeated." The option "covariate" is not allowed if assignment is between clusters.
	pre_post_corr	Relevant only if there is a pretest. The correlation between the pretest and the posttest.
Clustering items	assignment	One of three options: (default=unclustered) • "independent" or equivalently "unclustered" • "within" or equivalently "within_clusters" • "between" or equivalently "between_clusters"
	change_score_icc	Relevant only if assignment is between clusters and there is a pretest. The intraclass correlation of the change scores (posttest minus pretest).
	cluster_size	Relevant only if assignment is between clusters or within clusters. The mean number of members in each cluster.
	cluster_size_sd	Relevant only if assignment is between clusters. The standard deviation of the number of members in each cluster. (default=0)
	icc	Relevant only if assignment is between clusters or within clusters. The intraclass correlation of the variable of interest in the absence of treatment.
	nclusters	The total number of clusters available (for between clusters or within clusters assignment).
Two of the following three items must be specified: sample size, power, and effect size.		
	ntotal	The total sample size available (for unclustered assignment. For clustered assignment, use "cluster_size" and "nclusters.")
	power	If specified: The desired power of the test. If returned: The expected power of the test.
	Any one of the following may be used to specify the effect size:	
	d_main	Effect size measure: standardized mean difference ME/σ_y.
	effect_size_ratio	Effect size measure: signal to noise ratio β^2/σ_y^2.
	std_coef	Effect size measure: standardized effect-coded regression coefficient β/σ_y.
	raw_coef	Effect size measure: unstandardized effect-coded regression coefficient β.
	raw_main	Effect size measure: unstandardized mean difference ME.

7. Calculating Power Without Clustering

7.1 Example

Suppose you want to calculate power for a 5-factor experiment, analyzed using a 2nd order model (i.e., main effects and 2-way interactions). Suppose that the smallest main effect expected to be of interest for any active factor is 3 units, on a response variable whose error standard deviation is expected to be 10. Then use the code

```
%INCLUDE "S:\MyFolder\FactorialPowerPlan.sas";
%FactorialPowerPlan(assignment=independent,
          model_order=2,
          nfactors=5,
          ntotal=300,
          raw_main=3,
          sigma_y=10);
```

- "assignment=independent" indicates that participants are not in clusters,(default)
- "model order" denotes the order of the highest order term to be included in the regression
- "nfactors" denotes the number of factors in the experiment
- "ntotal" denotes the total sample size (not the sample size per condition)
- "sigma_y" denotes the standard deviation of y within each treatment condition
- "raw_main" denotes the unstandardized main effect

The output is

```
------------------------------------------------------------
FactorialPowerPlan Macro
The Methodology Center
(c) 2012 Pennsylvania State University
------------------------------------------------------------
Assumptions:
There are  5 dichotomous factors.
There is independent random assignment.
Analysis will be based on main effects and 2-way interactions.
Two-sided alpha:   0.05
Total number of participants:   300
Effect size as unstandardized difference in means:   3.00
Assumed standard deviation for the response variable is   10.00
Attempting to calculate the estimated power.
------------------------------------------------------------
Results:
The calculated power is:  0.7354
------------------------------------------------------------
```

Experimenting with different values of "`nfactors`" and "`model_order`" will show that they have almost no effect on power in this situation, and this is to be expected, because

for any reasonable model size there are many fewer than 300 parameters and so the loss of degrees of freedom due to extra model parameters is of little or no importance (see Collins, Dziak, & Li, 2009). The values of "`nfactors`" and "`model_order`" might have been somewhat more important if the data were clustered, as in some of the later examples. However, "`raw_main`" and "`sigma_y`" have important effects on power in any case.

7.2 Specifying the Effect Size

There are several ways to express effect size for a main effect. All of the following four calculations give the same answer, 0.7354. "raw_main" represents ME_k, "raw_coef" represents β_k, "d_main" represents ME_k/σ and "std_coef" represents β_k/σ.

```
%FactorialPowerPlan(raw_main=3,
          sigma_y=10,
          nfactors=5,
          model_order=2,
          ntotal=300);

%FactorialPowerPlan(d_main=.3,
          nfactors=5,
          model_order=2,
          ntotal=300);

%FactorialPowerPlan(raw_coef=1.5,
          sigma_y=10,
          nfactors=5,
          model_order=2,
          ntotal=300);

%FactorialPowerPlan(std_coef=.15,
          nfactors=5,
          model_order=2,
          ntotal=300);
```

7.3 Including a Pretest

It is easy to change the power calculation to assume a pretest is present. Suppose that the pretest-posttest correlation is expected to be .6. Then we could say

```
%FactorialPowerPlan(assignment=independent,
          model_order=2,
          nfactors=5,
          ntotal=300,
          pre_post_corr=.6,
          pretest=covariate,
          raw_main=3,
          sigma_y=10);
```

or alternatively we could say

```
%FactorialPowerPlan(assignment=independent,

          model_order=2,
          nfactors=5,
          ntotal=300,
          pre_post_corr=.6,
          pretest=repeated,
          raw_main=3,
          sigma_y=10);
```

The calculated power for the `covariate` approach is 0.8991 and the calculated power for the "repeated" approach is 0.8251 (output omitted here to save space). As mentioned above, it is controversial whether to conclude from this that (1) the "covariate" approach is better, (2) that both methods are the same but the power formula for the "covariate" approach is too optimistic, or (3) somewhere in between. However, it is clear that either way, some power is gained relative to the no-pretest model, which offered a power of 0.7354.

8. Calculating Sample Size Without Clustering

Instead of calculating power for a given sample size, one might be interested in sample size for a given power. For example, running the code

```
%FactorialPowerPlan(std_coef=.15,
          nfactors=5,
          model_order=2,
           power=.8);
```

- "std_coef" is the standardized effect-coded regression coefficient measure of effect size
- "model order" denotes the order of the highest order term to be included in the regression
- "nfactors" denotes the number of factors in the experiment
- "power" denotes the expected power of the experiment

provides the output

```
------------------------------------------------------------
FactorialPowerPlan Macro
The Methodology Center
(c) 2012 Pennsylvania State University
------------------------------------------------------------
Assumptions:
There are  5 dichotomous factors.
There is independent random assignment.
Analysis will be based on main effects and 2-way interactions.
Desired power:   0.80
Two-sided alpha:   0.05
Effect size as standardized regression coefficient:   0.15
Attempting to calculate the estimated required sample size.
------------------------------------------------------------
Results:
The calculated sample size is   351
------------------------------------------------------------
```

Internally, %FactorialPowerPlan implements this calculation by using the power formula iteratively. It successively tries a number of sample sizes, calculates power from each, and compares the expected and target power for each sample size. The proposed sample size is increased or decreased based on whether the expected power is below or above target. This is a binary search and allows the macro to quickly find the exact sample size that corresponds most closely to the target power.

8.1 Specifying Effect Size

As before, effect size can be specified in multiple ways. Any of the following will give the same answer of 351 subjects.

```
%FactorialPowerPlan(raw_main=3,
          sigma_y=10,
          nfactors=5,
          model_order=2,
          power=.8);

%FactorialPowerPlan(raw_coef=1.5,
          sigma_y=10,
          nfactors=5,
          model_order=2,
          power=.8);

%FactorialPowerPlan(d_main=.3,
          nfactors=5,
          model_order=2,
          power=.8);

%FactorialPowerPlan(std_coef=.15,
          nfactors=5,
          model_order=2,
          power=.8);

%FactorialPowerPlan(effect_size_ratio=.0225,
          nfactors=5,
          model_order=2,
          power=.8);
```

8.2 Including a Pretest

It is reasonable to suppose that the needed sample size could be reduced by adding a pretest. To see how much, we could run the code

```
%FactorialPowerPlan(std_coef=.15,
          nfactors=5,
          model_order=2,
          power=.8,
          pre_post_corr=.6,
          pretest=covariate);
```

or the code

```
%FactorialPowerPlan(std_coef=.15,
          model_order=2,
          nfactors=5,
          power=.8,
          pre_post_corr=.6,
          pretest=repeated);
```

The calculated sample size is 226 in the first case or 282 in the second. This is less than the 351 which was required without a pretest.

As previously mentioned, it is controversial whether to conclude from this that (1) the "covariate" approach is better, (2) that both methods are the same but the power formula for the "covariate" approach is too optimistic, or (3) somewhere in between. However, it is clear that either way, some power is gained relative to the no-pretest model.

8.3 A Possible Complication: Avoiding Non-Integer Cell Sizes

A slight complication can sometimes occur in calculating power for factorial experiments. Specifically, when the hypothesized effect size is very large, the sample size corresponding to the desired per-factor power is less than the number of cells in the complete factorial. For example, consider an 8-factor study with a hypothesized standardized effect size of 1. Running the code

```
%FactorialPowerPlan(d_main=1,
          nfactors=8,
          model_order=3,
          power=.8);
```

- "d_main" is the standardized mean difference measure of effect size
- "model order" denotes the order of the highest order term to be included in the regression
- "nfactors" denotes the number of factors in the experiment
- "power" denotes the expected power of the experiment

produces the following output

```
Attempting to calculate the estimated required sample size.
The calculated sample size is    96
However, a complete factorial requires    256 subjects.
```

This is because a *t*-test comparing two conditions of size 48 each would have enough power. However, a complete factorial with 8 factors cannot have less than 256 subjects, because it has 256 conditions (cells) that must be non-empty. Therefore, in this situation the investigator has a choice of either (1) using the larger sample size and having more power than was considered necessary, or (2) using the smaller sample size, which requires using a fractional factorial instead of a complete factorial. Because 96/256 is less than ½ but greater than ¼, a 2^{8-2} quarter-factorial would be needed if only 96 subjects were to be used (see Collins, Dziak and Li, 2009). In that case, the 96-64=32 "extra" subjects should be distributed as evenly as possible among

the 64 cells, so that each cell gets either 1 or 2 subjects. However, it would be more desirable to have multiple participants per cell in case of dropout. **In any case, with only one or a few participants per cell, direct pairwise comparisons between cells are not feasible, and inferences would have to focus on main effects and two- or three- way interactions.**

9. Calculating Detectable Effect Size Without Clustering

Sometimes the number of subjects one can afford is fixed, and the desired target power is known, and all that remains is to calculate the smallest detectable effect size in order to decide whether the experiment is worth doing. %FactorialPowerPlan can handle this scenario. For example, the code

```
%FactorialPowerPlan(nfactors = 5,
          model_order = 2,
          power = .8,
          ntotal = 300,
          sigma_y = 10);
```

- "nfactors" denotes the number of factors in the experiment
- "model order" denotes the order of the highest order term to be included in the regression
- "power" denotes the expected power of the experiment
- "ntotal" denotes the total sample size available (unclustered)
- "sigma_y" denotes the standard deviation of y within each treatment condition)

produces the following output.

```
------------------------------------------------------------
FactorialPowerPlan Macro
The Methodology Center
(c) 2012 Pennsylvania State University
------------------------------------------------------------
Assumptions:
There are  5 dichotomous factors.
There is independent random assignment.
Analysis will be based on main effects and 2-way interactions.
Desired power:   0.80
Two-sided alpha:   0.05
Total number of participants:   300
Assumed standard deviation for the response variable is   10.00
Attempting to calculate the estimated detectable effect size.
------------------------------------------------------------
Results:
The detectable effect size is estimated as follows:
 As an unstandardized regression coefficient for either
 a main effect or an interaction:              1.6230
 As an unstandardized mean difference for a main effect:  3.2459
 As an unstandardized difference in differences for
 a 2-way interaction:                    6.4919
 As a standardized regression coefficient for either
 a main effect or an interaction:              0.1623
 As a standardized mean difference (Cohen d) for a
 main effect:                          0.3246
 As a standardized difference in differences for
 a 2-way interaction:                    0.6492
 As a standardized effect size ratio (Cohen f squared)
 for a main effect or interaction:             0.0263
------------------------------------------------------------
```

The seven answers given are not actually different, but are the same effect size expressed in different ways. They represent β, $ME = 2\beta$, $4\,\beta$, β/σ_y, $2\beta/\sigma_y$, $4\beta/\sigma_y$, and β^2/σ_y^2.

9.1 Including a Pretest

As before, a pretest makes the situation somewhat better. One could use either the code

```
%FactorialPowerPlan(nfactors = 5,
          model_order = 2,
          power = .8,
          ntotal = 300,
          sigma_y = 10,
          pre_post_corr=.6,
          pretest=covariate);
```

or

```
%FactorialPowerPlan(nfactors = 5,
       model_order = 2,
       power = .8,
       ntotal = 300,
       sigma_y = 10,
       pre_post_corr=.6,
       pretest=repeated);
```

The detectable effect size (expressed as standardized difference relative to σ_y) goes down from 0.32 to 0.26 or 0.29. In general, a lower detectable effect size is desirable because it means that treatment effects do not have to be as large in order to get a statistically significant result.

As previously mentioned, it is controversial whether to conclude from this that (1) the "covariate" approach is better, (2) that both methods are the same but the power formula for the "covariate" approach is too optimistic, or (3) somewhere in between. However, it is clear that either way, some power is gained relative to the no-pretest model.

10. Calculating Power With Clustering

When subjects are nested into clusters, power calculations are different depending on whether assignment is at the individual subject level (within clusters) or the cluster level (between clusters).

10.1 Example: Within Clusters

Within-clusters assignment is available by specifying "assignment=within." The following code:

```
%FactorialPowerPlan(raw_main=3,
          sigma_y=10,
          nfactors=5,
          model_order=2,
          assignment=within,
          cluster_size=10,
          icc=.1,
          nclusters=30);
```

- "raw_main" denotes the unstandardized mean difference (effect-size measure)
- "sigma_y" denotes the standard deviation of y within each treatment condition
- "nfactors" denotes the number of factors in the experiment
- "model order" denotes the order of the highest order term to be included in the regression
- "assignment" indicates the within-cluster assignment of treatment
- "cluster_size"denotes mean number of members in each cluster
- "icc" denotes intraclass correlation of variable of interest when not treated
- "nclusters" denotes the total number of clusters available

will produce the output

```
------------------------------------------------------------
FactorialPowerPlan Macro
The Methodology Center
(c) 2012 Pennsylvania State University
------------------------------------------------------------
Assumptions:
There are  5 dichotomous factors.
There is random assignment of individuals for each cluster (within-clusters
effects).
Analysis will be based on main effects and 2-way interactions.
Two-sided alpha:   0.05
Cluster size:  10.00
Number of clusters:   30
Effect size as unstandardized difference in means:   3.00
```

```
Intraclass correlation of response variable:   0.10
Assumed standard deviation for the response variable is   10.00
Attempting to calculate the estimated power.
------------------------------------------------------------
Results:
The calculated power is  0.7354
------------------------------------------------------------
```

Notice that because of the *within-clusters* assignment, and our assumption of no treatment-by-cluster interactions, the calculated power for 30 clusters of 10 members is the same as the calculated power for 300 individuals found earlier. The assumed additive random effects for cluster simply cancel out when comparing factor means.

10.2 Including a Pretest

Just as before, you can also include a pretest.

```
%FactorialPowerPlan(raw_main=3,
          sigma_y=10,
          nfactors=5,
          model_order=2,
          assignment=within,
          cluster_size=10,
          icc=.1,
          pre_post_corr=.6,
          pretest=covariate,
          nclusters=30);

%FactorialPowerPlan(raw_main=3,
          sigma_y=10,
          nfactors=5,
          model_order=2,
          assignment=within,
          cluster_size=10,
          icc=.1,
          pre_post_corr=.6,
          pretest=repeated,
          nclusters=30);
```

Adding a pretest with `pre_post_corr=.6` gives a calculated power of 0.8991 or 0.8625 for pretest as covariate or repeated measure, respectively. The former is the same as the equivalent model in the unclustered case, and the latter is actually a little higher. The unintuive, slightly higher power when clustering is technically accurate but conceptually misleading, because in general, independent random assignment provides the most statistical information. The reason for this seeming paradox is that we assumed that there were the same total variance σ_y^2, but now there is a cluster-level variance, so that the assumed individual-level variance becomes

lower in order to add up to the same number (Dziak et al., 2012). This shows that it can be important to think carefully about assumptions.

10.3 Example: Between Clusters

When assignment is between-clusters, then power calculations must consider not only the response variable intraclass correlation (ICC) and the mean cluster size, but also the change-score ICC (if there is a pretest) and the standard deviation of the cluster sizes. The code

```
%FactorialPowerPlan(raw_main=3,
          sigma_y=10,
          nfactors=5,
          model_order=2,
          assignment=between,
          cluster_size=10,
          cluster_size_sd=2,
          icc=.1,
          change_score_icc=.05,
          nclusters=30);
```

- "raw_main" denotes the unstandardized mean difference (effect-size measure)
- "sigma_y" denotes the standard deviation of y within each treatment condition
- "nfactors" denotes the number of factors in the experiment
- "model order" denotes the order of the highest order term to be included in the regression
- "assignment" indicates between-clusters assignment of treatment
- "cluster_size"denotes mean number of members in each cluster
- "cluster_size_sd" denotes standard deviation of the number of members in each cluster
- "icc" denotes intraclass correlation of variable of interest when not treated
- "changescore" denotes the posttest minus the pretest
- "nclusters" denotes the total number of clusters available

will produce the output

```
FactorialPowerPlan Macro

The Methodology Center
(c) 2012 Pennsylvania State University
-------------------------------------------------------------
Assumptions:
There are  5 dichotomous factors.
There is random assignment of clusters (between-clusters effects).
Analysis will be based on main effects and 2-way interactions.
Two-sided alpha:   0.05
```

```
Cluster size:  10.00
Cluster size standard deviation:   2.00
Number of clusters:   30
Effect size as unstandardized difference in means:   3.00
Intraclass correlation of response variable:   0.10
Intraclass correlation of change scores:   0.05
Assumed standard deviation for the response variable is   10.00
Attempting to calculate the estimated power.
------------------------------------------------------------
Results:
The calculated power is:  0.4121
However, a complete factorial requires          32 clusters.
------------------------------------------------------------
```

Notice that the same number of clusters now leads to a power of .4121 for the between-clusters experiment instead of .7354 for a within-clusters experiment. However, sometimes between-clusters assignment is necessary and so a sufficiently large number of clusters must somehow be obtained.

Adding a pretest to the between-clusters scenario as follows:

```
%FactorialPowerPlan(raw_main=3,
          sigma_y=10,
          nfactors=5,
          model_order=2,
          assignment=between,
          cluster_size=10,
          cluster_size_sd=2,
          icc=.1,
          change_score_icc=.05,
          pre_post_corr=.6,
          pretest=repeated,
          nclusters=30);
```

produces the slightly better (but still too low) estimated power of 0.6295. To incorporate the pretest as a repeated measure, we would use the following code. For between-clusters assignment, the pretest-as-covariate option is not allowed because power is difficult to predict in this setting (Murray, 1998; Dziak et al., 2012).

11. Calculating Sample Size With Clustering

When discussing sample size in the context of clustered data, we must consider both the size and the number of the clusters. For example, if there are 20 clusters each of size 10, then there are 20×10=200 total participants. In the kinds of experiments that we consider, the size of the clusters is not entirely under the investigator's control. For instance, the clusters may represent classrooms, schools, clinics, hospitals, or corporations, and their characteristic size is known before the study begins. The investigator then has to decide how many of these clusters need to be included. Therefore, the %FactorialPowerPlan macro requires that *cluster size* (average number of members per cluster) be input by the user and will then calculate the *number of clusters* required. The total number of participants, of course, can then easily be obtained as the product of these numbers. As mentioned earlier, all else being equal, between-clusters experiments require more clusters than within-clusters experiments. The macro also requires slightly more information for between-clusters experiments than for within-clusters experiments: the change score ICC (if there is a pretest) and the standard deviation in cluster sizes are required.

11.1 Example: Within Clusters

The following code:

```
%FactorialPowerPlan(raw_main=3,
          sigma_y=10,
          nfactors=5,
          model_order=2,
          assignment=within,
          cluster_size=10,
          icc=.1,
          power=.80);
```

- "raw_main" denotes the unstandardized mean difference (effect-size measure)
- "sigma_y" denotes the standard deviation of y within each treatment condition
- "nfactors" denotes the number of factors in the experiment
- "model order" denotes the order of the highest order term to be included in the regression
- "assignment" indicates within-cluster assignment of treatment
- "cluster_size"denotes mean number of members in each cluster
- "icc" denotes intraclass correlation of variable of interest when not treated
- "power" denotes the expected power of the experiment

will produce the output

```
------------------------------------------------------------
FactorialPowerPlan Macro
The Methodology Center
(c) 2012 Pennsylvania State University
------------------------------------------------------------
Assumptions:
There are  5 dichotomous factors.
There is random assignment of individuals for each cluster (within-clusters
effects).
Analysis will be based on main effects and 2-way interactions.
Desired power:   0.80
Two-sided alpha:   0.05
Cluster size:  10.00
Effect size as unstandardized difference in means:   3.00
Intraclass correlation of response variable:   0.10
Assumed standard deviation for the response variable is   10.00
Attempting to calculate the estimated required sample size.
------------------------------------------------------------
Results:
The calculated sample size is    36 clusters.
------------------------------------------------------------
```

Thus, 36 clusters (or 360 total participants) are recommended.

Including a Pretest

Somewhat fewer clusters are needed if there is a pretest. The code

```
%FactorialPowerPlan(raw_main=3,
      sigma_y=10,
      nfactors=5,
      model_order=2,
      assignment=within,
      cluster_size=10,
      icc=.1,
      pre_post_corr=.6,
      pretest=repeated,
      power=.80);
```

now produces a recommendation for only 26 clusters, not 36 clusters. If we include the pretest as a covariate

```
%FactorialPowerPlan(raw_main=3,
          sigma_y=10,
          nfactors=5,
          model_order=2,
          assignment=within,
          cluster_size=10,
          icc=.1,
          pre_post_corr=.6,
          pretest=covariate,
          power=.80);
```

then 23 clusters are recommended.

As previously mentioned, it is controversial whether to conclude from this that (1) the "covariate" approach is better, (2) that both methods are the same but the power formula for the "covariate" approach is too optimistic, or (3) somewhere in between. However, it is clear that either way, some power is gained relative to the no-pretest model.

11.3 Example: Between Clusters

All else being equal, between-clusters experiments require more clusters than within-clusters experiments. The macro also requires slightly more information for between-clusters experiments than for within-clusters experiments: the change score ICC (if there is a pretest) and the standard deviation in cluster sizes are required.

The code

```
%FactorialPowerPlan(raw_main=3,
          sigma_y=10,
          nfactors=5,
          model_order=2,
          assignment=between,
          cluster_size=10,
          cluster_size_sd=2,
          icc=.1,
          power=.80 );
```

produces the output

```
------------------------------------------------------------
FactorialPowerPlan Macro
The Methodology Center
(c) 2012 Pennsylvania State University
------------------------------------------------------------
Assumptions:
There are  5 dichotomous factors.
There is random assignment of clusters (between-clusters effects).
Analysis will be based on main effects and 2-way interactions.
Desired power:   0.80
Two-sided alpha:   0.05
Cluster size:  10.00
Cluster size standard deviation:   2.00
Effect size as unstandardized difference in means:   3.00
Intraclass correlation of response variable:   0.10
Intraclass correlation of change scores:   0.05
Assumed standard deviation for the response variable is   10.00
Attempting to calculate the estimated required sample size.
------------------------------------------------------------
Results:
The calculated sample size is    71 clusters.
------------------------------------------------------------
```

Including a Pretest

Fully 71 clusters of average size 10 are required. Fortunately, a pretest reduces this somewhat. The code

```
%FactorialPowerPlan(raw_main=3,
          sigma_y=10,
          nfactors=5,
          model_order=2,
          assignment=between,
          cluster_size=10,
          cluster_size_sd=2,
          icc=.1,
          change_score_icc=.05,
          pre_post_corr=.6,
          pretest=repeated,
          power=.80 );
```

produces a recommendation of only 42 clusters of average size 10. For between-clusters assignment, the pretest-as-covariate option is not allowed because power is difficult to predict in this setting (Murray, 1998; Dziak et al., 2012).

12. Calculating Detectable Effect Size with Clustering

Finally, one can also calculate detectable effect size under clustering.

12.1 Example: Within Clusters

The following code

```
%FactorialPowerPlan(sigma_y=10,
          nfactors=5,
          model_order=2,
          assignment=within,
          cluster_size=10,
          icc=.1,
          power=.80,
          nclusters=50);
```

- "sigma_y" denotes the standard deviation of y within each treatment condition
- "nfactors" denotes the number of factors in the experiment
- "model order" denotes the order of the highest order term to be included in the regression
- "assignment" indicates within-cluster assignment of treatment
- "cluster_size"denotes mean number of members in each cluster
- "icc" denotes intraclass correlation of variable of interest when not treated
- "power" denotes the expected power of the experiment
- "nclusters" denotes the total number of clusters available

will produce the output

```
------------------------------------------------------------
FactorialPowerPlan Macro
The Methodology Center
(c) 2012 Pennsylvania State University
------------------------------------------------------------
Assumptions:
There are  5 dichotomous factors.
There is random assignment of individuals for each cluster (within-clusters
effects).
Analysis will be based on main effects and 2-way interactions.
Desired power:   0.80
Two-sided alpha:   0.05
Cluster size:  10.00
Number of clusters:   50
Intraclass correlation of response variable:   0.10
Assumed standard deviation for the response variable is   10.00
Attempting to calculate the estimated detectable effect size.
------------------------------------------------------------
Results:
The detectable effect size is estimated as follows:
 As an unstandardized regression coefficient for either
 a main effect or an interaction:              1.2554
 As an unstandardized mean difference for a main effect:  2.5108
```

```
 As an unstandardized difference in differences for
 a 2-way interaction:                    5.0217
 As a standardized regression coefficient for either
 a main effect or an interaction:              0.1255
 As a standardized mean difference (Cohen d) for a
 main effect:                          0.2511
 As a standardized difference in differences for
 a 2-way interaction:                    0.5022
 As a standardized effect size ratio (Cohen f squared)
 for a main effect or interaction:             0.0158
------------------------------------------------------------
```

If a pretest is used, it may become feasible to detect smaller effects. Here, the pretest is included as a repeated measure.

```
%FactorialPowerPlan(sigma_y=10,
          nfactors=5,
          model_order=2,
          assignment=within,
          cluster_size=10,
          pre_post_corr=.6,
          pretest=repeated,
          icc=.1,
          power=.80,
          nclusters=50);
```

The output is

```
------------------------------------------------------------
Results:
The detectable effect size is estimated as follows:
 As an unstandardized regression coefficient for either
 a main effect or an interaction:              1.0653
 As an unstandardized mean difference for a main effect:  2.1305
 As an unstandardized difference in differences for
 a 2-way interaction:                    4.2610
 As a standardized regression coefficient for either
 a main effect or an interaction:              0.1065
 As a standardized mean difference (Cohen d) for a
 main effect:                          0.2131
 As a standardized difference in differences for
 a 2-way interaction:                    0.4261
 As a standardized effect size ratio (Cohen f squared)
 for a main effect or interaction:             0.0113
------------------------------------------------------------
```

Including a Pretest

The detectable effect size, as a standardized difference d relative to σ_y, goes from .25 down to .21. If we include the pretest as a covariate,

```
%FactorialPowerPlan(sigma_y=10,
          nfactors=5,
          model_order=2,
```

```
          assignment=within,
          cluster_size=10,
          pre_post_corr=.6,
          pretest=covariate,
          icc=.1,
          power=.80,
          nclusters=50);
```

the output is

```
------------------------------------------------------------
Results:
The detectable effect size is estimated as follows:
 As an unstandardized regression coefficient for either
 a main effect or an interaction:              1.0043
 As an unstandardized mean difference for a main effect:  2.0086
 As an unstandardized difference in differences for
 a 2-way interaction:                     4.0173
 As a standardized regression coefficient for either
 a main effect or an interaction:              0.1004
 As a standardized mean difference (Cohen d) for a
 main effect:                           0.2009
 As a standardized difference in differences for
 a 2-way interaction:                     0.4017
 As a standardized effect size ratio (Cohen f squared)
 for a main effect or interaction:             0.0101
------------------------------------------------------------
```

The detectable effect size, as a standardized difference d relative to σ_y, goes from .25 down to .20 (compared to .21 when the test is included as a repeated measure).

As previously mentioned, it is controversial whether to conclude from this that (1) the covariate approach is better, (2) that both methods are the same but the power formula for the covariate approach is too optimistic, or (3) somewhere in between. However, it is clear that either way, some power is gained relative to the no-pretest model.

12.2 Example: Between Clusters

The code

```
%FactorialPowerPlan(sigma_y=10,
          nfactors=5,
          model_order=2,
          assignment=between,
          cluster_size=10,
          cluster_size_sd=2,
          icc=.1,
          change_score_icc=.05,
          power=.80,
          nclusters=50);
```

- "sigma_y" denotes the standard deviation of y within each treatment condition
- "nfactors" denotes the number of factors in the experiment
- "model order" denotes the order of the highest order term to be included in the regression
- "assignment" indicates within-cluster assignment of treatment
- "cluster_size"denotes mean number of members in each cluster
- "cluster_size_sd" denotes standard deviation of the number of members in each cluster
- "icc" denotes intraclass correlation of variable of interest when not treated
- "changescore" denotes the posttest minus the pretest
- "power" denotes the expected power of the experiment
- "nclusters" denotes the total number of clusters available

produces the output

```
------------------------------------------------------------
Results:
The detectable effect size is estimated as follows:
 As an unstandardized regression coefficient for either
 a main effect or an interaction:             1.7963
 As an unstandardized mean difference for a main effect:  3.5927
 As an unstandardized difference in differences for
 a 2-way interaction:                    7.1854
 As a standardized regression coefficient for either
 a main effect or an interaction:             0.1796
 As a standardized mean difference (Cohen d) for a
 main effect:                          0.3593
 As a standardized difference in differences for
 a 2-way interaction:                    0.7185
 As a standardized effect size ratio (Cohen f squared)
 for a main effect or interaction:            0.0323
```

As expected, the detectable effect size is larger now than it was in the within-clusters case; that is, the experiment is less sensitive to small but possibly meaningful effects. However, a pretest may help somewhat. For between-clusters assignment, the pretest-as-covariate option is not allowed because power is difficult to predict in this setting (Murray, 1998; Dziak et al., 2012). To incorporate the pretest as a repeated measure, the following code

```
%FactorialPowerPlan(sigma_y=10,
          nfactors=5,
          model_order=2,
          assignment=between,
          cluster_size=10,
          cluster_size_sd=2,
          pre_post_corr=.6,
          pretest=repeated,
          icc=.1,
          change_score_icc=.05,
```

```
          power=.80,
          nclusters=50);
```

produces the output

```
------------------------------------------------------------
Results:
The detectable effect size is estimated as follows:
 As an unstandardized regression coefficient for either
 a main effect or an interaction:              1.3613
 As an unstandardized mean difference for a main effect:  2.7225
 As an unstandardized difference in differences for
 a 2-way interaction:                    5.4451
 As a standardized regression coefficient for either
 a main effect or an interaction:              0.1361
 As a standardized mean difference (Cohen d) for a
 main effect:                           0.2723
 As a standardized difference in differences for
 a 2-way interaction:                    0.5445
 As a standardized effect size ratio (Cohen f squared)
 for a main effect or interaction:             0.0185
------------------------------------------------------------
```

The detectable effect size, as a standardized difference relative to σ_y, goes from .35 down to .27.

13. References

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20*, 93–114. doi:10.2307/271083

Anderson, S., Auquier, A., Hauck, W., Oakes, D., Vandaele, W., & Weisberg, H., (1980). *Statistical methods for comparative studies.* New York: John Wiley & Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collins, L. M., Dziak, J. J., & Li, R. (2009). Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. *Psychological Methods, 14*(3), 202-224.

Dziak, J., Nahum-Shani, I. R., & Collins, L. M. (2012). Multilevel factorial experiments for developing behavioral interventions: Power, sample size, and resource considerations. *Psychological Methods, 17,* 153-175.

Eldridge, S. M., Ashby, D., & Kerry, S. (2006). Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology, 35*, 1292-1300.

Frison, L., & Pocock, S. J. (1992). Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Statistics in Medicine, 11*, 1685-1704.

Janega, J. B., Murray, D. M., Varnell, S. P., Blitstein, J. L., Birnbaum, A. S., & Lytle, L. A. (2004). Assessing intervention effects in a school-based nutrition intervention trial: Which analytic model is most powerful? *Health Education & Behavior, 31,* 756 –774.

McAlister, F. A., Straus, S. E., Sackett, D. L., & Altman, D. G. (2003). Analysis and reporting of factorial trials: A systematic review. *Journal of the American Medical Association, 289,* 2545–2553.

Murray, D. M., & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in cluster-randomized trials. *Evaluation Review, 27,* 79–103.

Murray, D. M. (1998). *Design and analysis of cluster-randomized trials*. New York, NY: Oxford University Press.

Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis* (2nd ed.). Mahwah, NJ: Erlbaum.

Oakes, J. M., & Feldman, H. A. (2001). Statistical power for nonequivalent pretest-posttest designs : The impact of change-score versus ANCOVA models. *Evaluation Review, 25*, 3-28.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*, 199–213.

Smolkowski, K. (2010, April 30). *Gain score analysis*. Accessed at http://homes.ori.org/~keiths/Files/Tips/Stats_GainScores.html

Spybrook, J., Raudenbush, S. W., Congdon, R., & Martinez, A. (2009). *Optimal design for longitudinal and multilevel research: Documentation for the optimal design software* (Version 2.0). Retrieved from *www.wtgrantfoundation.org*

Vickers, A. J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: A simulation study. *BMC Medical Research Methodology*, *1*, 6. doi:10.1186/1471-2288-1-6. Retrieved from http://www.biomedcentral.com/1471-2288/1/6

14. Appendix: Models and Formulas

In this section we describe the formulas used by the %FactorialPowerPlan macro in the different scenarios described in the main body of the users' guide.

No Clustering and No Pretest

We assume the regression model

$$y_{ij} = \beta_0 + \sum \beta_k x_{ik} + e_i,$$

where e_i are random normal errors with variance σ_e^2. (Interactions between the factors may be in the model also, but for simplicity they are not shown in the expression above, as we are mainly interested here in the sampling distribution of any given regression coefficient estimate.) The variance σ_y^2 of y_i after adjusting for the treatment is equivalent to σ_e^2. The noncentrality parameter for calculating power for the effect represented by a given coefficient β (specifically, the hypothesis that this $\beta = 0$ for this effect) is

$$\lambda = \frac{N\beta^2}{\sigma_e^2} = \frac{N\beta^2}{\sigma_y^2},$$

where N is the total sample size. This is expression (4.3) in Spybrook et al. (2011), rescaled by 4 because they are using $+\frac{1}{2}, -\frac{1}{2}$ coding and we are using $+1, -1$ coding. Arguably, following Cohen (1988), it may be more conservative to use $\lambda = (N - d)\beta^2/\sigma_{\text{error}}^2$ where d is the number of regression coefficients in the model. This is based on the distinction between the denominators in the maximum likelihood (n) and unbiased ($n - d$) formulas. However, there is not a clear right answer; it should not matter much for practical sample sizes.

No Clustering, Pretest as Covariate

If there is a pretest entered as a covariate, then we assume

$$p_i = \mu_p + p_i^*$$
$$y_i - \beta_p p_i = \beta_0 + \sum \beta_k x_{ki} + e_{ij} \quad ,$$

where p_i is the pretest, y_i is the posttest, and $p_i^* \sim N(0, \sigma_p^2)$ and $e_i \sim N(0, \sigma_e^2)$ are independent random error. The overall (marginal) variance of y_i adjusting for the treatment but not adjusting for the pretest is denoted $\sigma_y^2 = \beta_p^2\sigma_p^2 + \sigma_e^2$. It also turns out that $\sigma_e^2 = \left(1 - \rho_{\text{pre,post}}^2\right)\sigma_y^2$ where $\rho_{\text{pre,post}}$ is the correlation between the pretest and posttest.

The noncentrality parameter for testing whether a particular regression coefficient β equals zero is

$$\lambda = \frac{N\beta^2}{\sigma_e^2} = \frac{N\beta^2}{\left(1-\rho_{\text{pre,post}}^2\right)\sigma_y^2}.$$

This is (4.11) in Spybrook et al. (2011), rescaled as discussed above. A very technical point is that, if the pretest is a latent variable measured with error, and $\rho_{\text{pre,post}}$ is misinterpreted as the correlation between the latent pretest and the posttest, then the power estimate based on the second formula may be overly optimistic (see Oakes & Feldman, 2001). Therefore, it is better to instead use the correlation of the observed pretest and the observed posttest; this will be the correlation of the latent and observed pretests times the correlation of the latent pretest and posttest.

No Clustering, Pretest as Repeated Measure

If there is a pretest entered as a repeated measure, then we assume

$$p_i = \mu_p + r_i + e_{0i}$$
$$y_i - p_i = \beta_0 + \sum \beta_k x_{ki} + (e_{1i} - e_{0i}) \ ,$$

where $r_{ij} \sim N(0, \sigma_r^2)$, $e_{0ij} \sim N(0, \sigma_e^2)$ and $e_{1ij} \sim N(0, \sigma_e^2)$. That is, the repeated measures model works like a regression model on gain scores (Anderson et al., 1980). The overall pretest variance is denoted $\sigma_y^2 = \sigma_r^2 + \sigma_e^2$, and the pretest-posttest correlation is $\rho_{\text{pre,post}} = \sigma_r^2/\sigma_y^2$. The noncentrality parameter is

$$\lambda = \frac{N\beta^2}{2\sigma_e^2} = \frac{N\beta^2}{2(1-\rho_{\text{pre,post}})\sigma_y^2} \ .$$

Note that in this expression $\rho_{\text{pre,post}}$ is not squared here.

Clustering and No Pretest

For individual i in cluster j, we assume the regression model

$$y_{ij} = \beta_0 + \sum \beta_k x_{kij} + u_j + e_{ij} \ ,$$

where $u_j \sim N(0, \tau_u^2)$ independently for each cluster and $e_{ij} \sim N(0, \sigma_e^2)$ independently for each individual.

Within-clusters assignment. If assignment is within-clusters then the noncentrality parameter is

$$\lambda = \frac{N\beta^2}{2\sigma_y^2} \ ,$$

where $\sigma_y^2 = \tau_u^2 + \sigma_e^2$ is the overall variance. This is (5.4) in Spybrook et al. (2011), rescaled, with treatment-by-cluster variance assumed to be negligible. This assumption may not be appropriate, but more research is needed regarding how to deal with this issue in a multiple-factor situation. Spybrook et al. (2011) were following Raudenbush and Liu (2000), who were considering the single-factor case.

Between-clusters assignment. If assignment is between-clusters then the noncentrality parameter is

$$\lambda = \frac{N\beta^2}{\sigma_y^2\left(1 + (\tilde{n} - 1)\rho_y\right)}$$

(Murray 1998, Dziak et al., 2012). Here N is the total number of participants and $\tilde{n}$ is an adjusted measure of the size of each cluster. If every cluster has the same number n of members, then $\tilde{n} = n$. Otherwise, Dziak et al. (2012) suggested $\tilde{n} = ((CV_n)^2 + 1)n$ based on Eldridge, Ashby, and Kerry (2006), where CV_n is the expected standard deviation of cluster size divided by the expected mean of cluster size. Last, $\sigma_y^2 = \tau_u^2 + \sigma_e^2$ is the overall variance, and ρ_y is the intraclass correlation of the clusters (at pretest, or at posttest after adjusting for treatment). This assumes treatment-by-cluster variance is negligible.

Clustering and a Pretest as Covariate

Within-clusters assignment. If assignment is within-clusters and the pretest is a covariate, then Dziak et al. (2012,) recommend estimating the noncentrality parameter as

$$\lambda = \frac{N\beta^2}{\left(1-\rho_{\text{pre,post}}^2\right)\sigma_y^2},$$

where $\rho_{\text{pre,post}}^2$ is calculated ignoring clusters rather than within clusters. As before, this formula ignores the possibility of random cluster-by-treatment interaction, so the formulas of Raudenbush and Liu (2000), although more complicated, are likely more realistic.

Clustering and a Pretest as Repeated Measure

For a repeated measures clustering design, we assume model (8) or (9) in Dziak et al., for within-clusters and between-clusters factors respectively. Either model implies that the posttest minus pretest difference or "change score" is

$$Y_{1ij} - Y_{0ij} = \gamma_{100} + u_j + \sum \beta_k x_{kij} + u_j + e_{ij}$$

for the *i*th individual in the *j*th cluster. (The above expression is a somewhat simpler way to write the model than the multilevel notation in Dziak et al.; here we write u_j in place of u_{10j}, and β_k in place of γ_{1k0} or γ_{10k}.) We also assume that in the absence of treatment effects, the overall variance σ_y^2 and intraclass correlation ρ_y would be the same for Y_{0ij} as for Y_{1ij}. Finally, we denote the intraclass correlation of the change scores as ρ_{change}.

Within-clusters assignment. If assignment is within-clusters then the noncentrality parameter expression from Dziak et al. (2012) is equivalent to

$$\lambda = \frac{N\beta^2}{2\sigma_y^2(1-\rho_{\text{pre,post}})(1-\rho_y)} \ .$$

This assumes the treatment-by-cluster variance is negligible.

Between-clusters assignment. If assignment is within-clusters then the noncentrality parameter expression from Dziak et al. (2012) is equivalent to

$$\lambda = \frac{N\beta^2(1-\rho_{\text{change}})}{2\sigma_y^2(1-\rho_{\text{pre,post}})(1-\rho_y)(1+(\tilde{n}-1)\rho_{\text{change}})} \ .$$

This assumes the treatment-by-cluster variance is negligible.