Wharton Behavioral Lab
Guidelines for Increasing Statistical Power and Choosing Sample Sizes[1]

1. **Null hypothesis statistical tests (NHSTs) are still the norm for most academic research**, even though most statisticians, research methodologists, and academic journals encourage researcher to focus more on effect sizes (e.g., the size of specific differences between means and the confidence intervals around those means; see Appelbaum, et al. 2018; Cohen 1994; Fritz, Morris, & Richler 2012; Wilkinson 1999)[2]. Thus, it is important to understand the relationships among effect sizes, NHSTs, sample size, and statistical power. A NHST assumes that the null hypothesis is true and estimates the probability that the observed effect size will occur when the true effect size is exactly zero; this probability is called the p-value. The purpose of a NHST is to evaluate whether there is evidence for the alternative hypothesis; by choosing a threshold such that we consider there to be evidence if the p-value is less than .05, we decrease the risk of making a Type I error (i.e., concluding that an effect exists when it does not) to .05. The purpose of computing statistical power is to assess the risk of making a Type II error (i.e., concluding that an effect does not exist when it does). Given decisions about expected effect sizes and acceptable levels of risk for making Type I and Type II error, it is possible to estimate the sample size necessary to meet those criteria.

   Statistical power, ϕ, is usually defined as the probability that a "true" effect will be detected by the "rejection" of a NHST, and 80% is a frequent benchmark for sufficient power. Statistical power is a function of

   (1) the **statistical model** used to generate the NHST (usually Y is a linear function of X, and either or both may be vectors for each observation),

   (2) the **estimation method** used to fit that model to observed data (usually least-squares or maximum likelihood),

   (3) the ***p*-value**, or α, that is used as the criterion for the NHST (usually $\alpha = .05$; thus, Type I Error is usually required to be 4 times smaller than Type II Error, which is 20% for $\phi = 80\%$),

   (4) the **size of the effect**, *ES*, which might be expressed in units of the dependent variable per unit of independent variable (e.g., β for standardized coefficients and B for unstandardized coefficients, respectively) or more abstractly in standard deviations of the error term (e.g., Cohen's d), and

---

[1] This research note was prepared by Wes Hutchinson, Wharton School, University of Pennsylvania, November 12, 2018. Much thanks to Professors Paul Rosenbaum, Dylan Small, and Joe Simmons for comments on an earlier version of these guidelines. Useful books on this topic included Bailar and Hoaglin (2009) and Murphy, Myors, and Wolach (2014).

[2] There are many measures of effects size. The most commonly reported ES measures (especially in meta-analyses) are Cohen'd d, the Pearson correlation coefficient (r) and, for ANOVAs, the partial eta squared (η_p^2; see Fritz, Morris, & Richler 2012). Confidence intervals (CIs) are computed for a specific confidence level, say 95%. The interpretation of a 95% CI is "Were this procedure to be repeated on numerous samples, the fraction of calculated confidence intervals (which would differ for each sample) that encompass the true population parameter would tend toward 95%,"and NOT "There is a 95% probability that the population parameter lies within the CI."

(5) the **size of the sample** (*N* for the total sample, or n_i, for each group, *i*).

Decisions about all five of these factors result from social conventions in the academic community; thus, researchers are ultimately responsible for making and defending their decisions about these factors given the specific goals of their research.

2. **Before collecting data, researchers should determine the minimum effect size that is to be detected and what null hypothesis statistical test (NHST) will be used to detect the effect**. Although a heuristic set of effect sizes (small, medium, and large) is commonly used, expected effects sizes derive from the theories and empirical results of the research area addressed by an experiment. Thus, there are no universal guidelines, although there are some general benchmarks (e.g., Simmons 2014), and insights about common research strategies (e.g., attenuating a known simple effect, Simonsohn 2014, and mediation analysis, Fritz, Kenney, & MacKinnon 2016, Weingarten & Hutchinson 2018). Most statistical packages have procedures that will compute the power (ϕ) of a given NHST for specified total sample sizes and effect sizes (e.g., the GLMPOWER and POWER procedures in SAS). These same packages will also compute the total sample size required to provide (on average) a given level of statistical power. Below are some examples the sample sizes needed to achieve ϕ = .80 and ϕ = .90 for different type of dependent measures, NHSTs, and effect sizes.

TABLE

Examples of Sample Sizes Needed for Desired Power and Effect Sizes

F-ratio test for a main effect or interaction or linear contrast for a fixed effect, between subjects ANOVA model of a DV from a balanced factorial design that has normally distributed error.

Effect Size	***r***	**Cohen's *d***	**partial η^2**	**Total Sample Size, ϕ = .80**	**Total Sample Size, ϕ = .90**
	.049	.100	.002	3,142	4,206
Small	.098	.200	.010	786	1,052
	.145	.300	.022	352	470
	.191	.400	.038	200	266
Medium	.236	.500	.059	128	210
Large	.365	.800	.138	52	68

Pearson chi-square test of the difference between proportions for two groups of equal sample size for a binary dependent variable.

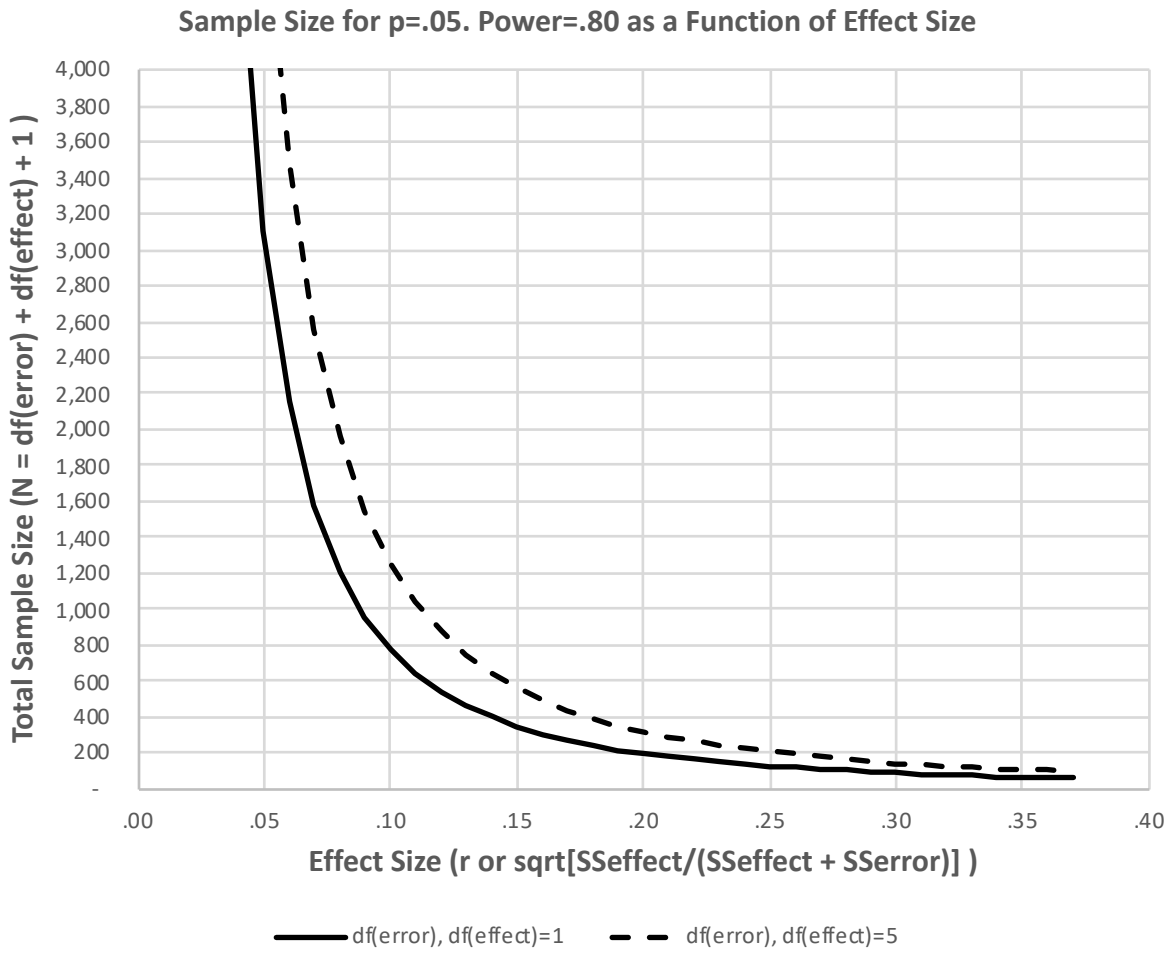
Effect Size	**r**	**Group A**	**Group B**	**Total Sample Size, ϕ = .80**	**Total Sample Size, ϕ = .90**
Medium	.218	.20	.40	128	178
Medium	.204	.30	.50	148	202
Medium	.200	.40	.60	154	212
Small	.109	.25	.35	518	716
Small	.102	.35	.45	592	820
Small	.100	.45	.55	618	854

Likelihood ratio chi-square test of a single binary predictor (Prob(A) = Prob(B) = .5) in a binary logistic regression, possibly in the presence of one or more covariates that might be correlated with the tested predictor. Note that .45 vs. .55 is an odds ratio of 1.48 and .20 vs. .40 is an odds ratio of 2.67.

Effect Size	**r**	**Group A**	**Group B**	**Total Sample Size, ϕ = .80**	**Total Sample Size, ϕ = .90**
Medium	.218	.20	.40	158	211
Medium	.204	.30	.50	187	250
Medium	.200	.40	.60	195	261
Small	.109	.25	.35	645	864
Small	.102	.35	.45	750	1,004
Small	.100	.45	.55	794	1,063

A robust approximation for the sample size (N) required for p = .05 and ϕ = .90 is the simple formula $df_{error} = 7.75/r^2$, where r is the effect size expressed as a correlation coefficient and $N = df_{error} + 2$ (Murphy, Myors, & Wolach 2014). A more robust approximation for effects with $df_{effect} = k > 1$ (e.g., an ANOVA factor with k+1 conditions or a multiple regression model with k predictors is the simple formula $df_{error} = (5.26 + 3.24 \text{ x } \text{sqrt}[df_{effect}) / \text{partial } \eta^2$, where partial $\eta^2 = SS_{effect} / (SS_{effect} + SS_{error})$ and $N = df_{error} + k + 1$ (Murphy, Myors, & Wolach 2014). The Figure below illustrates how effect size affects sample size using these approximations.

FIGURE



Based on the results shown in the Table and the Figure, **a general rule of thumb is that a sample size of *N* = 200 should be "sufficient" for NHSTs with *df* = 1 to detect medium effect sizes** (e.g., $r = .2$, $d = .4$, or partial $\eta^2 = .04$). Samples sizes in the range of 400 to 800 are generally needed for small effect sizes (e.g., $r = .1$, $d = .2$, or partial $\eta^2 = .01$), and "really" small effects require sample sizes between 1,000 and 10,000. **The reader is reminded that even at these sample sizes there is still a 10% to 20% chance that a valid effect will not be detected by a NHST.**

3. **For all data, identify outliers and other anomalous observations** (based on accepted, criteria that are unrelated to research hypotheses). Virtually all methods of data analysis assume data are consistent with some "well-behaved" model of error; thus, **anomalous observations decrease statistical power**. Anomalous observations should be omitted from statistical analyses, but the criteria used and the number of omitted observations should be fully reported.

4. **For ordinary least squares regression (OLS), when IVs are correlated, statistical power is reduced**. To see how statistical power can be increased for OLS regression analyses, consider the following.

observed $Y = B_{Y0.123...k} + B_{Y1.23...k} X_1 + B_{Y2.13...k} X_2 + B_{Y3.12...k} X_3 + ... + B_{Yk.123...k} X_k + \varepsilon$,
where $\varepsilon \sim N(0, \sigma^2)$.

predicted $Y = B_{Y0.123...k} + B_{Y1.23...k} X_1 + B_{Y2.13...k} X_2 + B_{Y3.12...k} X_3 + ... + B_{Yk.123...k} X_k$.

predicted $z(Y) = \beta_{Y1.23...k} z(X_1) + \beta_{Y2.13...k} z(X_2) + \beta_{Y3.12...k} z(X_3) + ... + \beta_{Yk.123...k} z(X_k)$.

Note that unstandardized regression coefficients, B_{Yi}, are related to standardized coefficients, β_{Yi}, as follows: $B_{Yi} = \beta_{Yi}$ SD(Y) / SD(X_i) .

$\boldsymbol{t = \beta / SE(\beta_{Yi.12...(i)...k})}$, where

$$\mathbf{SE}_{\beta_{Yi.12...(i)...k}} = \sqrt{\frac{1-R_Y^2}{N-k-1}} \sqrt{\frac{1}{1-R_i^2}},$$

$\boldsymbol{N}$ = total number of observations,
$\boldsymbol{k}$ = number of independent variables
$\boldsymbol{R_Y^2}$ = the variance in Y accounted for by the independent variables,
$\boldsymbol{R_i^2}$ = the variance in X_i accounted for by the other independent variables.

It is easy to see from these equations that, all else equal, **statistical power is increased** (i.e., t is increased when the effect is "true") when

(1) β is increased or B is increased,
(2) residual error, $(1 - R_Y^2)$, is decreased,
(3) N is increased,
(4) k is decreased, and
(5) R_i^2 is decreased.

Multicollinearity problems result when R_i^2 becomes large (see Cohen, Cohen, West and Aiken 2003).

5. **When some IVs are arithmetic functions of "simple effect" variables** that are also included in the regression (e.g., polynomials and interactions), the coding of the simple variables is very important. **Mean-centered, effects-coding of the simple variables increase statistical power for regression coefficients** because this reduces the correlations between IVs (i.e., R_i^2; see Irwin & McClelland 2001), even though R_Y^2 is not affected.

6. For between-subjects data (i.e., the error structure for each observed DV is assumed to be i.i.d. normal, and all IVs are assumed to be fixed effects), **ANOVA is a special case of regression** in which main effects and interactions are modeled as "bundles" of effects-coded IVs in such a way that (for balanced designs) the correlations between variables in different effect bundles are exactly zero. Thus, all else equal, statistical power is maximized in balanced ANOVA experimental designs compared to unbalanced designs and designs in which independent variables are measured rather than manipulated (with random assignment to conditions) because R_i^2 is structurally equal to zero. Also, when measured covariates are

included, random assignment implies that the expected value of R_i^2 is zero for the manipulated factors.

7. To better understand statistical power for between-subjects ANOVA, note that

   $$\boldsymbol{F}[\boldsymbol{df}\textbf{(effect)}, \boldsymbol{df}\textbf{(error)}] = MS(\text{effect})/MS(\text{error})$$

   $$= [SS(\text{effect})/\, df(\text{effect})]/[\, SS(\text{error})/\, df(\text{error})].$$

   This formula is central to understanding ANOVA & ANCOVA, especially how modeling decisions such as using contrasts and covariates can potentially add statistical power. **Two important properties of the F test for a balanced between-subjects design analyzed by a fixed effect ANOVA** are that (1) *MS*(error) is the same for all effects (i.e., main effects, interactions, and linear contrasts) and (2) the "complexity" if the factorial design usually does not have a large effect on needed total sample size (*N*) because *SS*(error) should not be affected by the design and *df*(error) = N - k - 1, and k + 1 is the degrees of freedom for the full ANOVA model (e.g., *df*(2x3) = *df*(A) + *df*(B) + *df*(AxB) + *df*(INT) = 1 + 2 + 2 + 1 =6). Thus, if a 2x2 design requires N=200, then a 2x3 design for the same level error, effect sizes, and *df*(effect) will require N=202 because *k* has increased by 2.

8. Most approaches to analyzing repeated measures data involves choosing alternative estimators for *MS*(error). It is generally the case that **repeated measures (i.e., within-subjects) experimental designs provide greater statistical power than between-subject designs** because the main effects of subjects do not contribute to *MS*(error), as they do in a between subject design. Of course, many experimental manipulations are pragmatically difficult to implement within-subjects (e.g., due to carry-over or demand effects).

9. **For unbalanced designs or when covariates are used**, Type I SS and Type III SS are not equivalent, and Type III SS (which removes all shared variance from the analyzed effects) is generally preferred, unless there is a theory-driven rationale for sequential analysis (i.e., Type I SS). Type III SS is also used in OLS regression models. **Type III SS tests have less statistical power than Type I SS tests**, but it is usually hard to defend the use of Type I SS.

10. **Omnibus F-tests, by themselves, are almost always conceptually too forgiving** because they test the null hypothesis that the observed pattern of means for a main effect or interaction contains differences that are greater than what would be expected if all conditions had the same true means. Almost all theoretically interesting hypotheses are much more specific about the predicted pattern of means and may or may not be tested by a single main effect or interaction. Always check the observed pattern of means to be sure it is consistent with theory. Whenever possible, test a specific linear contrast that represents the predicted pattern. **Linear contrasts have greater statistical power than omnibus tests** because *df*(effect) = 1 and, if valid, the predicted pattern of means will account for most of the systematic variance in the omnibus test (i.e., *MS*(contrast)/1 > *MS*(omnibus effect) / *df*(omnibus effect)). For within-subjects effects, a simple and robust test is to (a) compute a single number for each subject that represents the predicted pattern (however complicated) and then (b) test the mean across subjects against the expected value under the null hypothesis (usually zero).

11. **Traditional ANOVA models are saturated** in the sense that the within-cell sample means are perfectly predicted. When there are unequal numbers of observations in each cell (i.e., an unbalanced design), least-squares marginal means are computed using these predicted cell-mean values. Least-squares means usually better represent the population means than do the raw sample means.

12. **For experimental designs that have pretest and posttest observations** as repeated measures, a simple between-subjects ANOVA of difference scores (sometimes called gain scores) is identical to repeated measures ANOVA. However, a more general and statistically powerful model is to use the pretest measure to predict the posttest measure in the absence of a treatment effect. This is usually done by using an ANCOVA model with the pretest measure as the covariate. **The covariate method is more statistically powerful than the difference-score method** because it includes the difference score method as a special case (i.e., the coefficient of the pretest measure is 1). In some cases, the conceptual interpretability of the differences score may outweigh the statistical advantage of the ANCOVA approach, however.

13. **Almost all within subjects experimental designs, must "control" for possible order effects using a Latin Square or some other fractional factorial design.** Such designs typically make the main effect of order uncorrelated with all other effects (and interactions) of interest. However, some higher-order effects are necessarily confounded with order and/or each other.

14. **The traditional multivariate approach to repeated measures ANOVA (e.g., PROC GLM in SAS) is recommended for balanced experimental designs** (especially when sample sizes are small) because the resulting F-tests are "exact." In such models, each within-subjects effect (and its interactions with all between subjects effects) is tested using the interaction of the "residual" within-subjects effect with the between-subjects effects as the error term in an F-test. For unbalanced designs, when there is missing data for some but not all conditions for some subjects, or when within-subjects covariates are to be used, a more general approach that incorporates a specific model of the within-subject variances and covariances (e.g., PROC MIXED in SAS; see Wolfinger and Chang 1998) is recommended (as long as sample sizes are reasonably large). The F-tests are usually estimated by (restricted) maximum likelihood, and interpreting them requires some care. This approach accommodates the unbalanced design and missing data much better than the traditional multivariate approach (and yields identical results for balanced designs), and it provides greater statistical power because is uses more data and explicitly models the error structure. However, the variance/covariance model must be taken seriously, and several conceptually plausible such models should be estimated before a final model is selected.

15. When there are **multiple dependent variables** and it is desirable to correct for family-wise error (i.e., getting at least one measure significant by testing many measures), then **sample sizes can be computed using a Bonferroni-corrected value for α** (e.g., for $\alpha = .05$ with 5 measures, an $\alpha = .01$ should be used). However, it should be noted that the Bonferroni correction is very conservative, emphasizing Type I over Type II errors, and in that sense, reduces statistical power somewhat. More statistically powerful methods include the

Bonferroni/Holm sequential test procedure (Holm 1979; also see Rosebaum 2008, Small, Volpp, and Rosenbaum 2011), and the MANOVA-based method developed by Lehmacher, Wassmer, and Reitmeir (1991). An alternative that imposes a somewhat weaker constraint on Type I error is the widely used false discovery rate (FDR) test developed by Benjamini and Hochberg (1995).

REFERENCES

Appelbaum, M., H. Cooper, R. B. Kline, E. Mayo-Wilson, A. M. Nezu, and S. M. Rao (2018), "Journal Article Reporting Standards for Quantitative Research in Psychology: The APA Publications and Communications Board Task Force Report," *American Psychologist*, 73 (1), 3-25.

Bailar J. C. and D. C. Hoaglin, *Medical Uses of Statistics*, 3rd. ed., 2009, Wiley.

Benjamini, Yoav & Yosef Hochberg (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57 (1), 289-300.

Cohen, Jacob (1994), "The Earth Is Round *(p* < .05)," American Psychologist, 49(12),997-1003.

Cohen, Jacob, Patricia Cohen, Stephen G. West and Leona S. Aiken, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed., 2003, Lawrence Erlbaum Associates.

Fritz, M.S., Kenny, D.A., & MacKinnon, D.P. (2016). The opposing effects of simultaneously ignoring measurement error and omitting confounders in a single-mediator model. *Multivariate Behavioral Research*, 51 (5), 681-97.

Fritz, C. O., P. E. Morris, and J. J. Richler (2012), "Effect Size Estimates: Current Use, Calculations, and Interpretation," *Journal of Experimental Psychology: General*, 141 (1), 2-18.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* 6, 65–70.

Irwin, Julie R. and Gary H. McClelland (2001) "Misleading heuristics and moderated multiple regression models," *Journal of Marketing Research,* 38, 100-110.

Lehmacher, Walter , Gernot, Wassmer, & Peter Reitmeir, (1991), "Procedures for Two-Sample Comparisons with Multiple Endpoints Controlling the Experimentwise Error Rate," *Biometrics*, 47, 511-521.

Murphy, Kevin R., Brett Myors and Allen Wolach, *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*, 4th ed., 2014 Routledge.

Rosenbaum, P. R. (2008). Testing hypotheses in order. *Biometrika*, 95(1), 248-252.

Simmons, Joseph (2014), "MTurk vs. The Lab: Either Way We Need Big Samples." *DataColada*, [18].

Simonsohn, Uri (2014), "No-way Interactions." *DataColada*, [17].

Small, D. S., K. Volpp, and P. R. Rosenbaum (2011), "Structured Testing of 2 × 2 Factorial Effects: An Analytic Plan Requiring Fewer Observations," *The American Statistician*, 65 (1), 11-15.

Weingarten, Evan and J. Wesley Hutchinson (2018), "Does Ease Mediate the Ease-of-Retrieval Effect? A Meta-Analysis," *Psychological Bulletin*, February online.

Wilkinson, Leland and the Task Force on Statistical Inference (1999), "Statistical Methods in Psychology Journals: Guidelines and Explanations," *American Psychologist*, Vol. 54, No. 8, 594-604.

Wolfinger, R. D., and M. Chang, (1998). Comparing the SAS GLM and MIXED procedures for repeated measures. Cary, NC: SAS Institute Inc.