Running head: PANGEA

PANGEA: Power ANalysis for GEneral Anova designs

Jacob Westfall

University of Texas at Austin

Working paper: Manuscript last updated October 11, 2016

Table of Contents

Abstract (155 words)

In this paper I present PANGEA (Power ANalysis for GEneral Anova designs; http://jakewestfall.org/pangea/), a user-friendly, open source, web-based power application that can be used for conducting power analyses in general ANOVA designs. A general ANOVA design is any experimental design that can be described by some variety of ANOVA model. Surprisingly, a power analysis program for general ANOVA designs did not exist until now. PANGEA can estimate power for designs that consist of any number of factors, each with any number of levels; any factor can be considered fixed or random; and any possible pattern of nesting or crossing of the factors is allowed. I demonstrate how PANGEA can be used to estimate power for anything from simple between- and within-subjects designs, to more complicated designs with multiple random factors (e.g., multilevel designs and crossed-random-effects designs). I document the statistical theory underlying PANGEA and describe some experimental features to be added in the near future.

Keywords: statistical power, experimental design, mixed models.

For decades, methodologists have warned about the low statistical power of the typical psychology study. Cohen (1962) originally estimated that, in the year 1960, the average statistical power of studies in social and abnormal psychology to detect a typical or "medium" effect size (fatefully defined by Cohen as a standardized mean difference of 0.5) was about 46%. There is scant evidence that the situation has improved since then (Marszalek, Barber, Kohlhart, & Holmes, 2011; Maxwell, 2004). Sedlmeier & Gigerenzer (1989) estimated the average statistical power in the year 1984, for the same research literature and effect size investigated by Cohen, to be about 37%. More recent analyses of the average statistical power in social psychology (Fraley & Vazire, 2014) and neuroscience (Button et al., 2013) find estimates of about 50% and 21%, respectively, for detecting typical effect sizes in those fields. Thus, despite persistent warning, the concept of statistical power has remained largely neglected in practice by scientific psychologists.

In the last few years, however, there has been renewed interest in statistical power and its implications for study design, fueled in large part by a "replication crisis" or "reproducibility crisis" gripping much of science, but psychology in particular (e.g., Pashler & Wagenmakers, 2012). It may not seem immediately obvious why such a crisis should lead to increased concern about statistical power. Indeed, when considered in the isolated context of a single study, the problems of low statistical power seem rather unimpressive; while it would clearly seem to be in the experimenter's own best interest that a study have a reasonably high chance of detecting some predicted effect (assuming the prediction is correct), it is not obvious whether it is ultimately anyone else's concern if the experimenter chooses, for whatever reasons, to run a statistically inefficient study. However, when considered in the broader context of entire programs of research built on many, many low-powered studies, the problems accruing from a

policy of running low-powered studies suddenly loom much larger, and it is now widely agreed that this has been a major factor precipitating the current crisis (Bakker, Dijk, & Wicherts, 2012; Ioannidis, 2005, 2008; Schimmack, 2012).

If researchers are to begin taking statistical power seriously, then minimally they need to understand and be able to compute statistical power for the kinds of experiments they are actually running. However, while complicated designs are entirely commonplace in psychology and neuroscience—for example, mixed (split plot) designs with predictors varying both within- and between-subjects (Huck & McLean, 1975), multilevel designs with hierarchically nested units (Raudenbush & Bryk, 2001), and designs employing random stimulus samples (Wells & Windschitl, 1999)—issues of statistical power tend only to be widely understood for relatively simple designs. For example, both the most popular textbook on power analysis (Cohen, 1988) and the most popular software for power analysis (Faul, Erdfelder, Lang, & Buchner, 2007) cover statistical power up to fixed-effects ANOVA, multiple regression models, and tests of the difference between two dependent means (i.e., matched pairs), but neither handle any of the three classes of more complicated designs just mentioned[1]. Some literature on statistical power does exist for certain special cases of these designs (Raudenbush, 1997; Raudenbush & Liu, 2000; Westfall, Kenny, & Judd, 2014), but more general treatments have remained inaccessible to psychologists, and there is often no accompanying software for researchers to use.

The purpose of this paper is to fix this situation. I present PANGEA (Power ANalysis for GEneral Anova designs), a user-friendly, open source, web-based power application that can be used for conducting power analyses in general ANOVA designs. A general ANOVA design is any experimental design that can be described by some variety of ANOVA model. Surprisingly,

[1] As of this writing, the software cited in the text, G*Power version 3.1.9.2, supports power analysis for omnibus tests in mixed designs, but does not support tests of general within-subject contrasts.

a power analysis program for general ANOVA designs has not existed until now. PANGEA can estimate power for designs that consist of any number of factors, each with any number of levels; any factor can be considered fixed or random; and any possible pattern of nesting or crossing of the factors is allowed. PANGEA can be used to estimate power for anything from simple between- and within-subjects designs, to more complicated designs with multiple random factors (e.g., multilevel designs and crossed-random-effects designs), and even certain dyadic designs (e.g., social relations model; Kenny, 1994), all in a single unified framework.

The rest of this paper is structured as follows. First I walk through demonstrations of how to specify several, progressively more complex designs in PANGEA. Next I describe the statistical theory and procedures underlying PANGEA. Finally, I give some of the technical details of PANGEA's software implementation and briefly describe some features that I plan to add in future versions. PANGEA can be accessed at http://jakewestfall.org/pangea/, where users also can download the source code to run PANGEA locally if they wish.

Specifying General ANOVA Designs

The general ANOVA model encompasses the models classically referred to as fixed-effects, random-effects, and mixed-model ANOVA (Winer, Brown, & Michels, 1991). I refer to any experimental design that can be described by a general ANOVA model as a general ANOVA design. This includes experiments involving any number of factors, each with any number of levels; any factor in the experiment can be considered fixed or random; and any possible pattern of nesting or crossing of the factors is allowed. Not included in the class of general ANOVA designs are designs involving continuous predictors or unequal sample sizes. Despite these limitations, this is clearly a very broad class of experimental designs, and PANGEA can be used to compute statistical power for any design within this class. While this makes PANGEA quite a

general power analysis tool, this very generality can also make PANGEA difficult to use at first, since it requires the user to be able to exactly specify the design of the study. In this section I give a brief tutorial on specifying general ANOVA designs in terms of what is nested or crossed, fixed or random, and what the replicates in the study are. I first give abstract definitions of these terms, and then I illustrate these concepts concretely by describing the specification of a series of progressively more complex study designs.

Terminology

Factors. The first and most fundamental term to understand is the concept of a factor. A factor is any categorical variable—measured by the experimenter—that can potentially explain variation in the response variable. The individual categories comprising the factor are called the *levels* of that factor. Importantly, factors refer not only to the treatment factors or predictors that are of primary interest (e.g., experimental group; participant gender), but may also refer to classification or grouping factors that are presumably not of primary interest, but which nevertheless may explain variation in the outcome (e.g., participants that are repeatedly measured; blocks or lists of stimulus materials; laboratories in a multi-site experiment).

Crossed vs. nested. The crossed vs. nested distinction is similar to the within-subject vs. between-subject distinction that is more familiar to most psychologists and neuroscientists. Factor A is said to be nested in factor B if each level of A is observed with one and only one level of B. For example, if each participant in a study is randomly assigned to a single group, then we can say that the Participant factor is nested in the Group factor. Or if we are studying students who attend one and only one school, we can say that the Student factor is nested in the School factor. In both of these examples, the levels of the containing factor (Group in the first case, School in the second case) vary “between” the levels of the nested factor.

Factors A and B are said to be crossed if every level of A is observed with every level of B and vice versa. For example, if we administer *both* an active treatment drug and an inert placebo drug to each participant in a study, then we can say that the Participant and Drug factors are crossed. If we further suppose in this experiment that we measured each participant twice per drug—once before administering the drug, and again after administering the drug—then we can say that the Drug and Time-point factors are crossed. In this example, the levels of each factor vary "within" the levels of the other factors.

Fixed vs. random. The distinction between fixed factors and random factors is probably the most conceptually subtle of the terms presented here. This situation is not helped by the fact that the distinction is not always defined equivalently by different authors (Gelman & Hill, 2006, p. 245). The definition given here is the one that is standard in the literature on analysis of variance (Cornfield & Tukey, 1956; Winer et al., 1991). For this definition, we start by imagining that, for each factor in the experiment, there is a theoretical population of potential levels that we might have used, the number of which, N, could be very large (e.g., approaching infinity). Say that our actual experiment involved n of these potential levels. If $n/N = 1$, so that the factor levels in our experiment fully exhaust the theoretical population of levels we might have used, then the factor is said to be fixed. If $n/N \approx 0$, so that the factor levels in our experiment are a sample of relatively negligible size from the theoretical population of levels we might have used, then the factor is said to be random[2].

One conceptual ambiguity with this definition is what exactly is meant by "a theoretical population of potential levels that we might have used." In what sense might we have used these unobserved factor levels? To answer this, it is useful to consider which factors it would *in*

[2] The in-between case, where the observed factor levels are an incomplete but non-negligible proportion of the population of potential levels (e.g., $n/N = 0.5$), has been studied in the ANOVA literature (e.g., Cornfield & Tukey, 1956), but is rarely discussed in practice.

principle be acceptable to vary or exchange in future replications of the study in question (Westfall, Judd, & Kenny, 2015). Random factors are ones for which it would be acceptable to exchange the levels of the factor for new, different levels in future replications of the experiment; that is, there are, in principle, other possible levels of the factor that could have served the experimenter's purposes just as well as those that were actually used. Examples of random factors include the participants in a study; the students and schools in an educational study; or the list of words in a linguistics study. Fixed factors are ones that we would necessarily require to remain the same in each replication of the study; if the levels were to be exchanged in future replications of the study, then the new studies would more properly be considered entirely different studies altogether. A factor is also fixed if no other potential levels of that factor are possible other than those actually observed. Examples of fixed factors include the experimental groups that participants are randomly assigned to; the socioeconomic status of participants on a dichotomous low vs. high scale; and participant gender.

Replicates. In traditional analysis of variance terminology, the number of replicates in a study refers to the number of observations in each of the lowest-level cells of the design; lowest-level in that it refers to the crossing of all fixed *and* random factors in the design, including e.g. participants. For example, in a simple pre-test/post-test style design where we measure each participant twice before a treatment and twice after the treatment, the number of replicates would be two, since there are two observations in each Participant-by-Treatment cell.

Example Designs

Two independent groups. We begin with the simplest possible design handled by PANGEA: an experiment where the units are randomly assigned to one of two independent groups. Interestingly, there are two equivalent ways to specify this design, depending on the unit

of analysis that we consider to be the replicates in the design. These two perspectives on the design are illustrated in Table 1.

Table 1
Two equivalent specifications of a two-group between-subjects design. The numbers in each cell indicate the number of observations in that cell of the design. Blank cells contain no observations.

Two-group between-subject design: Participants as replicates
Factors: **G**roup (fixed; 2 levels).
Design:
Replicates: 5

g_1	5
g_2	5

Two-group between-subject design: Participants as explicit factor
Factors: **G**roup (fixed; 2 levels), **P**articipant (random; 5 levels *per* ***G***).
Design: **P** nested in **G**.
Replicates: 1

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
g_1	1	1	1	1	1					
g_2						1	1	1	1	1

In the first specification, we simply have a single fixed factor with two levels, and the observations in each cell (i.e., the replicates) are the experimental participants. Thus, the participants are only an implicit part of the design. In the second specification, the participants are explicitly given as a random factor in the design, one that is nested in the fixed Group factor, and the replicates refer to the number of times we observe each subject. Thus, this latter specification is more general than the first specification in that it allows for the possibility of repeated measurements of each subject; when the number of replicates is one, as it is in Table 1, then it equivalent to the first specification.

Designs with a single random factor. Designs in which the participants (or, more generally, the experimental units) are observed multiple times require that the participants be given as an explicit, random factor in the design. Table 2 gives two examples of such designs.

Table 2
Examples of designs with a single random factor. The numbers in each cell indicate the number of observations in that cell of the design. Blank cells contain no observations.

2×2 within-subjects design: Two-color Stroop task
Factors: **P**articipant (random; 10 levels), **I**nk Color (fixed; 2 levels), **W**ord Color (fixed; 2 levels)
Design: **P** crossed with **I**, **P** crossed with **W**, **I** crossed with **W**.
Replicates: 10

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
i_1w_1	10	10	10	10	10	10	10	10	10	10
i_1w_2	10	10	10	10	10	10	10	10	10	10
i_2w_1	10	10	10	10	10	10	10	10	10	10
i_2w_2	10	10	10	10	10	10	10	10	10	10

2×3 mixed (split plot) design: Pre-test/Post-test assessment of three drugs
Factors: **T**ime (fixed; 2 levels), **D**rug (fixed; 3 levels), **P**articipant (random; 3 levels *per* ***D***)
Design: **T** crossed with **D**, **T** crossed with **P**, **P** nested in **D**.
Replicates: 1

	d_1			d_2			d_3		
	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
t_1	1	1	1	1	1	1	1	1	1
t_2	1	1	1	1	1	1	1	1	1

The first example is a 2×2 within-subjects design based on a simplified Stroop task involving only two colors (MacLeod, 1991). In this experiment, participants make speeded responses to color words presented on a computer screen, and their task is to indicate the *font color* that the color word is printed in. The word printed on the screen in each trial is either "red" or "blue," and the word is printed in either a red font or a blue font. Participants make 10 responses toward each of the four stimulus types. The Stroop effect refers to the observation that

response times tend to be slower when the font color and color word are inconsistent than when they are consistent. This experiment involves one random Participant and two fixed factors, Ink Color and Word Color. All three factors are crossed, and the number of replicates is 10, because there are 10 response times in each Participant × Ink Color × Word Color cell. The test of the Stroop effect corresponds to the test of the fixed Ink Color × Word Color interaction.

The second example, illustrated in the bottom part of Table 2, is a 2 (Time: pre-test vs. post-test) × 3 (Drug: d_1, d_2, or d_3) mixed design where the Time factor varies within-subjects and the Drug factor varies between-subjects. The Time and Drug factors are both fixed, and the random Participant factor is crossed with Time and nested in Drug. Because we measure each subject only once at each Time point, the number of replicates in this design is one.

Designs with multiple random factors. In PANGEA it is simple to specify designs that involve multiple random factors, and this is the most appropriate way to think about many common designs. Three examples of such designs are illustrated in Table 3.

The first example is a three-level hierarchical design commonly encountered in education research. In this example we have some intervention that we are assessing in a large study involving a number of elementary schools. The students (henceforth *pupils*) attending each elementary school belong to one and only one classroom. For each school, we randomly assign half of the classrooms to receive the intervention, and the other half of the classrooms to undergo some placebo procedure. Thus we have a fixed, two-level Intervention factor that is crossed with a random School factor, and which has a random Classroom factor nested in it. The Classroom factor is nested in the School factor, and we may view the pupils as the replicates (i.e., the observations in each Classroom). As in the two-independent-groups example discussed earlier, it is also possible to view this design as having an explicit, random Pupil factor that is nested in all

the other factors, and this would be appropriate if we measured each Pupil multiple times.

PANGEA could easily handle such a four-level design.

Table 3
Examples of designs with multiple random factors. The numbers in each cell indicate the number of observations in that cell of the design. Blank cells contain no observations.

Three-level hierarchical design: Pupils (replicates)-in-Classrooms-in-Schools
Factors: **S**chool (random; 3 levels), **I**ntervention (fixed; 2 levels), **C**lassroom (random; 2 levels *per* ***S*×*I***).
Design: **S** crossed with **I**, **C** nested in **S**, **C** nested in **I**.
Replicates: 20

	s_1				s_2				s_3			
	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}
i_1	20	20			20	20			20	20		
i_2			20	20			20	20			20	20

Crossed random factors: Stimuli-within-Condition design
Factors: **P**articipant (random; 6 levels), **T**ype (fixed; 2 levels), **W**ord (random; 3 levels *per* ***T***).
Design: **P** crossed with **T**, **P** crossed with **W**, **W** nested in **T**.
Replicates: 1

	t_1			t_2		
	w_1	w_2	w_3	w_4	w_5	w_6
p_1	1	1	1	1	1	1
p_2	1	1	1	1	1	1
p_3	1	1	1	1	1	1
p_4	1	1	1	1	1	1
p_5	1	1	1	1	1	1
p_6	1	1	1	1	1	1

Crossed random factors: Counterbalanced design
Factors: **G**roup (fixed; 2 levels), **P**articipant (random; 3 levels *per* ***G***), **B**lock (fixed; 2 levels), **S**timulus (random; 3 levels *per* ***B***).
Design: **P** nested in **G**, **S** nested in **B**, **G** crossed with **B**, **P** crossed with **S**.
Replicates: 1

		b_1			b_2		
		s_1	s_2	s_3	s_4	s_5	s_6
	p_1	1 (t_1)	1 (t_1)	1 (t_1)	1 (t_2)	1 (t_2)	1 (t_2)
g_1	p_2	1 (t_1)	1 (t_1)	1 (t_1)	1 (t_2)	1 (t_2)	1 (t_2)
	p_3	1 (t_1)	1 (t_1)	1 (t_1)	1 (t_2)	1 (t_2)	1 (t_2)
	p_4	1 (t_2)	1 (t_2)	1 (t_2)	1 (t_1)	1 (t_1)	1 (t_1)
g_2	p_5	1 (t_2)	1 (t_2)	1 (t_2)	1 (t_1)	1 (t_1)	1 (t_1)
	p_6	1 (t_2)	1 (t_2)	1 (t_2)	1 (t_1)	1 (t_1)	1 (t_1)

The second example involves a design that has been frequently discussed in the psycholinguistics literature (Clark, 1973; Raaijmakers, Schrijnemakers, & Gremmen, 1999). In this example, a sample of participants study a set of noun words that are either abstract (e.g., "truth") or concrete (e.g., "word") and then undergo a recognition test in which they indicate their degree of recognition for each word (Gorman, 1961). In this design, every participant responds to every word. In the past we have referred to this type of design as a Stimuli-within-Condition design (Westfall et al., 2014). As has been pointed out by many authors over many years (e.g., Coleman, 1964; Judd, Westfall, & Kenny, 2012), it is appropriate to view the sample of stimulus words as a random factor in the design, and failure to do so in the analysis can, in many cases, lead to a severely inflated type 1 error rate. Thus, this design consists of a random Word factor nested in a fixed Type factor, as well as a random Subject factor that is crossed with both Word and Type.

The final example of a design involving multiple random factors is similar to the Stimuli-within-Condition design just discussed, but in this design the fixed treatment factor is counterbalanced across the stimuli, so that each stimulus is sometimes observed in one level of the treatment factor and sometimes observed in the other level. For example, we give each participant two lists of words to study; for one of the lists, they are to give a definition of each word ("deep processing"), and for the other list, they are to indicate how many letters are in the word ("shallow processing"; Craik & Lockhart, 1972). After this task they undergo a recognition memory test in which they rate their degree of recognition toward every word. The counterbalancing takes place as follows. The full set of words is divided into two blocks, b_1 and b_2. Likewise, the participants are randomly assigned to one of two groups, g_1 or g_2. Thus, there is

a random Participant factor nested in a fixed Group factor, and a random Word factor nested in a fixed Block factor. The g_1 participants receive the b_1 words with the deep processing instructions (denoted t_1—the first level of an *implicit* Treatment factor) and the b_2 words with the shallow processing instructions (denoted t_2). The g_2 participants receive the b_1 words with the shallow processing instructions and the b_2 words with the deep processing instructions. As discussed by Kenny and Smith (1980), and as illustrated at the bottom of Table 3, the test of the Group × Block interaction is equivalent to the test of t_1 vs. t_2, the levels of the implicit Treatment factor representing deep vs. shallow processing of the words.

Statistical Details of Power Computations

In this section I describe how PANGEA performs the actual power analysis once the user has specified the design. To obtain statistical power estimates, there are ultimately three pieces of information needed: (1) the noncentrality parameter for a noncentral *t* or *F* distribution; (2) the associated degrees of freedom; and (3) the alpha level of the test. Here I show how the noncentrality parameter (henceforth denoted δ) and degrees of freedom (henceforth denoted ν) are obtained from the information that PANGEA solicits from the user. I first describe the unstandardized solution, in which δ and ν are written in terms of means and variances (i.e., on the scale of the dependent variable), and then describe a standardized solution, in which δ and ν are written in terms of standardized mean differences and proportions of variance (i.e., in a dimensionless metric). In a final subsection I give some theoretical and empirical justifications for some of the default values of the input parameters used by PANGEA.

Unstandardized Solution

Noncentrality parameter. The noncentrality parameter for a noncentral *t* distribution can be written in the same form as the sample *t*-statistic, but it is based on population values. Thus, if there are f fixed cells in total, then the noncentrality parameter is equal to

$$\delta = \frac{\mathrm{E}[\hat{\beta}]}{\sqrt{\mathrm{var}(\hat{\beta})}} = \frac{\frac{\sum_i^f c_i \mu_i}{\sum_i^f c_i^2}}{\sqrt{\frac{\sigma_{diff}^2 f}{N \sum_i^f c_i^2}}} = \frac{\sum_i^f c_i \mu_i \sqrt{N}}{\sqrt{\sigma_{diff}^2 f \sum_i^f c_i^2}},$$

where $\hat{\beta}$ is the estimate of the effect, the c_i are the contrast code values that multiply the cell means (the μ_i), N is the total number of observations in the experiment, and σ_{diff}^2 is the appropriate error mean square, i.e., the variance of the mean difference implied by the contrast (Winer et al., 1991, p. 147).

Most of the terms comprising the noncentrality parameter—the whole numerator, as well as the contrast codes and sample sizes—are obtained simply by direct input from the user. Finding σ_{diff}^2 requires a little more work. To do so, PANGEA first uses the Cornfield-Tukey algorithm (Cornfield & Tukey, 1956; Winer et al., 1991, pp. 369–374) to find the expected mean square equations for the design specified by the user. Then σ_{diff}^2 can be obtained by taking the expected mean square for the effect to be tested and subtracting the term that involves the corresponding variance component, so that what remains is all the sources of variation that lead to variation in the effect *other than* true variation in the effect. This is the same logic used to select the appropriate denominator of an *F*-ratio for testing effects in an ANOVA.

There is one minor, necessary modification to the procedure described above based on the Cornfield-Tukey algorithm, which is due to the two slightly different ways that a variance component is defined in classical ANOVA (on which Cornfield-Tukey is based) compared to in the modern literature on linear mixed models (which is the notation used by PANGEA). In the

classical ANOVA literature, the variance component associated with a factor is defined as the variance in the *effects* of that factor's levels; let the variance component defined in this way be denoted as σ^2_{EMS}. In the mixed model literature, a variance component is defined as the variance in the *coefficients* associated with that factor's levels; let the variance component defined in this way be denoted as σ^2_{LMM}. As a consequence, the two definitions of the variance component for a given factor have the relationship

$$\sigma^2_{EMS} = \sum_{i}^{f} c_i^2 \, \sigma^2_{LMM}.$$

Practically, this simply means that after applying the Cornfield-Tukey algorithm in the manner described above, one must also multiply the variance components by the sum of squared contrast codes associated with that factor, where applicable.

As an example, consider the Stimuli-within-Condition design illustrated in the middle part of Table 3. The expected mean squares for this design, and the associated degrees of freedom, are given in Table 4. So when computing power for a test of the Treatment effect in this design, we find σ^2_{diff} by taking the expected mean square for Treatment, subtracting the term involving the Treatment variance component (σ^2_T), and multiplying the $\sigma^2_{T\times S}$ term by the sum of squared contrasts for the Treatment factor, leaving us with

$$\sigma^2_{diff} = \sigma^2_E + r\sigma^2_{W\times S} + rw \sum_{i}^{f} c_i^2 \, \sigma^2_{T\times S} + rs\sigma^2_W.$$

PANGEA would require the user to enter the values of the variance components found in the right-hand side of this equation, namely, the error variance (σ^2_E), the Word × Subject interaction variance ($\sigma^2_{W\times S}$, a.k.a. the variance of the random Word × Subject intercepts), the Treatment × Subject interaction variance ($\sigma^2_{T\times S}$, a.k.a. the variance of the random Subject slopes), and the

Word variance (σ_W^2, a.k.a. the variance of the random Word intercepts). Once these variances have been given, they can easily be combined with the contrast codes, sample sizes, and expected regression coefficient supplied by the user to form the noncentrality parameter.

Table 4
Expected mean squares for the Stimuli-within-Condition design. The lower-case labels denote the sample sizes of the corresponding factor, so that t is the number of Treatments, w is the number of Words per Treatment, and s is the number of subjects. The number of replicates is denoted by r.

Label	Source of variation	Degrees of freedom	Expected value of mean square
T	Treatment	$t-1$	$\sigma_E^2 + r\sigma_{W\times S}^2 + rw\sigma_{T\times S}^2 + rs\sigma_W^2 + rws\sigma_T^2$
W	Word	$t(w-1)$	$\sigma_E^2 + r\sigma_{W\times S}^2 + rs\sigma_W^2$
S	Subject	$s-1$	$\sigma_E^2 + r\sigma_{W\times S}^2 + rtw\sigma_S^2$
T×S	Treatment×Subject	$(t-1)(s-1)$	$\sigma_E^2 + r\sigma_{W\times S}^2 + rw\sigma_{T\times S}^2$
W×S	Word×Subject	$t(w-1)(s-1)$	$\sigma_E^2 + r\sigma_{W\times S}^2$
E	Error	$tws(r-1)$	σ_E^2

Degrees of freedom. The degrees of freedom used by PANGEA are based on the Welch-Satterthwaite approximation (Satterthwaite, 1946; Welch, 1947). The first step is to find the linear combination of mean squares whose expectation will result in the correct expression for σ_{diff}^2. Then the Welch-Satterthwaite approximation states that the degrees of freedom ν for this linear combination of mean squares is approximately equal to

$$\nu \approx \frac{(\sum k_i M_i)^2}{\sum \frac{(k_i M_i)^2}{\nu_i}},$$

where the M_i are the mean squares, the k_i are the weights for each mean square in the linear combination, and the ν_i are the degrees of freedom associated with each mean square.

The appropriate linear combination of mean squares is found by solving the system of expected mean square equations for the k_i. To do this, we first collect the expected mean square equations into a matrix **X** where the rows represent the mean squares, the columns represent the variance components, and the entries in each cell are the corresponding terms from the table of expected mean square equations. We then set

$$\mathbf{X^T k = s},$$

where **k** is the vector of weights k_i and **s** is a vector containing the terms that comprise σ^2_{diff}. Finally we solve this equation for **k**, yielding

$$\mathbf{k} = (\mathbf{X^T})^{-1}\mathbf{s},$$

To illustrate this process, consider again the Stimuli-within-Condition design. In this case for **X** and **s** we have

$$\mathbf{X} = \begin{bmatrix} \sigma_E^2 & r\sigma_{W\times S}^2 & rw\sigma_{T\times S}^2 & 0 & rs\sigma_W^2 & rws\sigma_T^2 \\ \sigma_E^2 & r\sigma_{W\times S}^2 & 0 & 0 & rs\sigma_W^2 & 0 \\ \sigma_E^2 & r\sigma_{W\times S}^2 & 0 & rtw\sigma_S^2 & 0 & 0 \\ \sigma_E^2 & r\sigma_{W\times S}^2 & rw\sigma_{T\times S}^2 & 0 & 0 & 0 \\ \sigma_E^2 & r\sigma_{W\times S}^2 & 0 & 0 & 0 & 0 \\ \sigma_E^2 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \mathbf{s} = \begin{bmatrix} \sigma_E^2 \\ r\sigma_{W\times S}^2 \\ rw\sigma_{T\times S}^2 \\ 0 \\ rs\sigma_W^2 \\ 0 \end{bmatrix},$$

so that when we solve for **k** we obtain $\mathbf{k^T} = [0 \quad 1 \quad 0 \quad 1 \quad -1 \quad 0]$, indicating that the appropriate linear combination of mean squares is $M_W + M_{T\times S} - M_{W\times S}$. And indeed we can verify that

$$\mathrm{E}[M_W + M_{T\times S} - M_{W\times S}] = \sigma_E^2 + r\sigma_{W\times S}^2 + rw\sum_i^f c_i^2\,\sigma_{T\times S}^2 + rs\sigma_W^2 = \sigma_{diff}^2.$$

With the weights k_i, the degrees of freedom ν_i, and the variance components and sample sizes input by the user, we can now simply plug values into the Welch-Satterthwaite equation to obtain the approximate degrees of freedom ν for the noncentral *t* distribution.

Standardized Solution

Standardized mean difference. The standardized effect size used by PANGEA is a generalized version of Cohen's *d*, or the standardized mean difference between conditions. Cohen's *d* is classically defined for two independent groups as

$$d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}},$$

where μ_1 and μ_2 are the means of the two groups and σ_{pooled} is the pooled standard deviation, i.e., the square root of the average of the two variances, assuming the two groups are of equal size. Our generalized *d* extends this in two ways: The numerator allows for arbitrary contrasts among the group means rather than simply a difference between two groups, and the denominator is given a corresponding definition based on the standard deviation of an observation within each group, pooled across all groups.

First we consider the numerator. One way to view the numerator of *d* is as the regression coefficient from a simple linear regression with a categorical predictor c_i, with values c_1 and c_2 such that $|c_1 - c_2| = 1$. For example, values c_1 and c_2 might be $\{0, 1\}$ or $\{-\frac{1}{2}, \frac{1}{2}\}$. So one obvious way to generalize the numerator is as $\left(\sum_i^f c_i \mu_i\right)/\left(\sum_i^f c_i^2\right)$, which is the population value of the regression coefficient for a contrast-coded predictor, where f is the total number of fixed cells and the c_i can be any set of weights that sum to 0. Generalizing the numerator in this way would create the complication that the overall value of *d* is sensitive to the choice of contrast code values even when the means and pooled standard deviation remain constant. For example, choosing contrast codes of $\{-1,1\}$ would result in a smaller effect size than choosing $\{-\frac{1}{2}, \frac{1}{2}\}$. Clearly this is an undesirable property of a standardized effect size. To correct this, we will insert a term that rescales the contrast codes so that the range of the codes is always equal to 1 as it is in

the classical case. Let a and b be the minimum and maximum values, respectively, of the c_i. Then the numerator of our generalized effect size d will be

$$\frac{\sum_i^f c_i \mu_i (b-a)}{\sum_i^f c_i^2},$$

which is invariant to the scale of the contrast codes and allows for a natural generalization to multiple groups.

Next we define the σ_{pooled} term comprising the denominator of *d*, representing the pooled standard deviation, that is, the square root of the variance of an observation in each fixed cell, averaged across all the fixed cells. To find this in the general ANOVA case, first we consider the variance of an observation in any single condition. Let $\mathrm{var}(y_i)$ be the variance of an observation in the *i*th fixed cell, from a total of *f* fixed cells. For example, in an experiment with two fixed factors, each with two levels, we have $f = 4$. This variance can be written in a general way as

$$\mathrm{var}(y_i) = \underbrace{\sum_j \sigma_j^2}_{\substack{\text{random} \\ \text{intercepts}}} + \underbrace{\sum_k c_{ik}^2 \sigma_k^2}_{\substack{\text{random} \\ \text{slopes}}} + \underbrace{2c_{ik}\sigma_k}_{\substack{\text{intercept-} \\ \text{slope} \\ \text{covariances}}} + \underbrace{\sum_{p<q} 2c_{ip}c_{iq}\sigma_{pq}}_{\substack{\text{slope-slope} \\ \text{covariances}}} + \underbrace{\sigma_E^2}_{\substack{\text{random} \\ \text{error} \\ \text{term}}}.$$

Because the variance within each cell is a function of the contrast code values in that cell, there is generally unequal variance across the cells, a fact pointed out by Goldstein, Browne, and Rasbash (2002). The pooled variance across all of the fixed cells, σ_{pooled}^2, is then equal to

$$\begin{aligned}
\frac{1}{f}\sum_i^f \mathrm{var}(y_i) &= \frac{1}{f}\sum_i^f \left(\sum_j \sigma_j^2 + \sum_k c_{ik}^2 \sigma_k^2 + 2c_{ik}\sigma_k + \sum_{p<q} 2c_{ip}c_{iq}\sigma_{pq} + \sigma_E^2 \right) \\
&= \frac{1}{f}\left(\sum_i^f \sum_j \sigma_j^2 + \sum_i^f \sum_k c_{ik}^2 \sigma_k^2 + \sum_i^f \sum_k 2c_{ik}\sigma_k + \sum_i^f \sum_{p<q} 2c_{ip}c_{iq}\sigma_{pq} + \sum_i^f \sigma_E^2 \right) \\
&= \frac{1}{f}\left(f\sum_j \sigma_j^2 + \sum_k \sigma_k^2 \sum_i^f c_{ik}^2 + \sum_k \sigma_k \sum_i^f 2c_{ik} + \sum_{p<q} \sigma_{pq} \sum_i^f 2c_{ip}c_{iq} + f\sigma_E^2 \right)
\end{aligned}$$

$$= \sum_j \sigma_j^2 + \sum_k \sigma_k^2 \left(\frac{\sum_i c_{ik}^2}{f} \right) + \sigma_E^2.$$

The last step above depends on the assumption that the predictors comprise a complete set of orthogonal contrast codes; it is an important step because it means that, under the contrast coding assumption, the pooled variance does not depend on any of the random covariances in the model.

Putting all this together, we define our generalized *d* as

$$d = \frac{\sum_i^f c_i \, \mu_i (b - a)}{\sum_i^f c_i^2 \sqrt{\sigma_{pooled}^2}},$$

which reduces to the classical definition in the case of two independent groups, but can be extended to an arbitrary number of fixed cells and allows for the inclusion of random effects.

Variance partitioning coefficients. The concept of variance partitioning coefficients (VPCs) was discussed by Goldstein et al. (2002), who define them in the context of multilevel models (i.e., mixed models with hierarchically nested random factors; Raudenbush & Bryk, 2001; Snijders & Bosker, 2011) as the proportion of random variance in the outcome that is accounted for by the different "levels" of the model. General ANOVA models do not generally involve a notion of multiple "levels" of the model, but we will make use of VPCs to partition the random variance in the outcome into the proportion due to each individual variance component.

The definition of the VPCs is simple. We saw earlier that the pooled variance can be written as a linear combination of variance components:

$$\sigma_{pooled}^2 = \sum_j \sigma_j^2 + \sum_k \sigma_k^2 \left(\frac{\sum_i c_{ik}^2}{f} \right) + \sigma_E^2.$$

The VPC for each variance component is formed by taking the ratio of the corresponding term (i.e., the variance component as well as any coefficients multiplying it) over the pooled variance. For example, the VPC for the error variance component, σ_E^2, would be

$$V_E = \frac{\sigma_E^2}{\sum_j \sigma_j^2 + \sum_k \sigma_k^2 \left(\frac{\sum_i c_{ik}^2}{f}\right) + \sigma_E^2}.$$

The sum of all the VPCs is 1, and each VPC can be interpreted simply as the proportion of variance due to that variance component.

Noncentrality parameter and degrees of freedom. It is easy to write the noncentrality parameter and degrees of freedom in terms of the standardized effect size and VPCs just defined. The noncentrality parameter is

$$\delta = \frac{\sum_i^f c_i \mu_i \sqrt{N}}{\sqrt{\sigma_{diff}^2 f \sum_i^f c_i^2}} = \frac{\sum_i^f c_i \mu_i \sqrt{N}}{\sqrt{\sigma_{diff}^2 f \sum_i^f c_i^2}} \left(\frac{(b-a)\sqrt{\sigma_{pooled}^2}}{(b-a)\sqrt{\sigma_{pooled}^2}} \right) = \frac{d\sqrt{N}\sqrt{\sum_i^f c_i^2 / f}}{(b-a)\sqrt{\frac{\sigma_{diff}^2}{\sigma_{pooled}^2}}}$$

$$= \frac{d\sqrt{N}\sigma_c}{(b-a)\sqrt{\sigma_{VPC}^2}},$$

where σ_c is the standard deviation of the contrast codes c_i, and σ_{VPC}^2 is identical to σ_{diff}^2 except the variance components in that expression are replaced by their VPCs. The degrees of freedom are approximately equal to

$$\nu \approx \frac{(\sum k_i M_i)^2}{\sum \frac{(k_i M_i)^2}{\nu_i}} = \frac{(\sum k_i M_i)^2}{\sum \frac{(k_i M_i)^2}{\nu_i}} \left(\frac{\left(\sigma_{pooled}^2\right)^2}{\left(\sigma_{pooled}^2\right)^2} \right) = \frac{\left(\sum k_i M_i^{VPC}\right)^2}{\sum \frac{\left(k_i M_i^{VPC}\right)^2}{\nu_i}},$$

where the M_i^{VPC} are identical to the M_i except the variance components in their expectations are replaced by the corresponding VPCs.

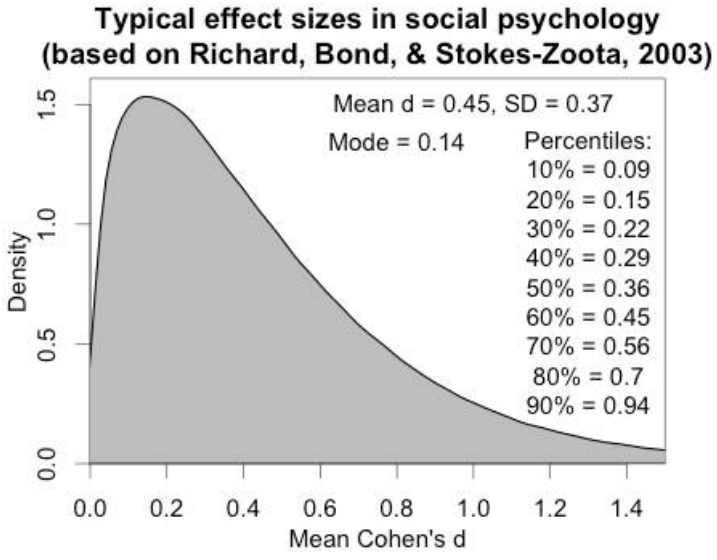
Default Inputs

When one finishes specifying the experimental design in PANGEA and begins considering the experimental parameters for the power analysis (effect size, sample sizes, etc.), one finds

some default values suggested for the standardized effect size and VPCs. In this section I give the rationale behind these default values.

Typical effect sizes. The default effect size suggested by PANGEA is $d = 0.45$, which is based on a distribution of values of Cohen's *d* derived from a meta-analysis by Richard, Bond Jr., & Stokes-Zoota (2003) and illustrated in Figure 1. Richard et al. (2003) conducted a meta-analysis of meta-analyses in the field of social psychology to determine the range of typical effect sizes across the field, involving some 25,000 individual studies published over 100 years in diverse research areas. While the focus of this meta-meta-analysis was the field of social psychology, I believe there is little reason to expect the distribution of typical effect sizes to be appreciably different in other areas of psychology (e.g., cognitive psychology), and in the absence of meta-analytic evidence of such a difference, I submit that a default of $d = 0.45$ represents a reasonable suggestion for most psychological studies if one has *no other information* about the specific effect to be studied.

Figure 1
Distribution of typical values of Cohen' *d* in social psychology as shown on the PANGEA page.



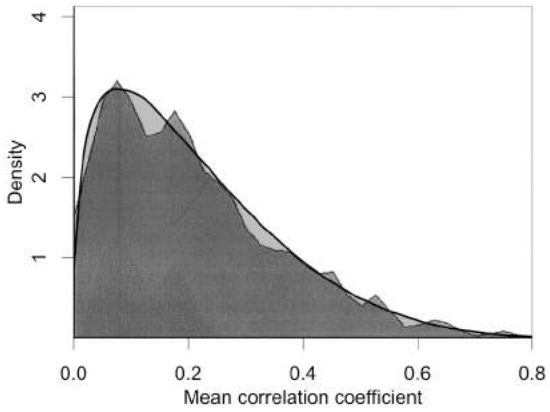
The meta-analysis by Richard et al. (2003) was actually based on average values of the correlation coefficient, rather than Cohen's *d*; some assumptions were required in order to construct the *d* distribution shown in Figure 1, which I describe here. First I sought to characterize the distribution of correlation coefficients reported by Richard et al. (2003), which is shown in Figure 2 as the bumpy density curve. Based on the shape and range of this distribution, I considered characterizing it as a beta distribution. The mean and standard deviation of the empirical distribution were reported by Richard et al. (2003) to be $\bar{x} = .21$ and $\sigma_x = .15$, respectively. The beta distribution has two parameters α and β, and I estimated these parameters by finding the values that would produce the observed mean and standard deviation, using the estimates

$$\alpha = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{\sigma_x^2} - 1\right)$$

$$\beta = (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{\sigma_x^2} - 1\right).$$

This produced the beta distribution illustrated as the smooth density in Figure 2, which appears to provide a good characterization of the empirical distribution.

Figure 2
Empirical distribution of correlation coefficients from Richard et al. (2003) along with best-fitting beta distribution.



The next step is to convert this distribution of correlation coefficients to a distribution of values of Cohen's *d*. To do this, I simulated many, many values from the best-fitting beta distribution, converted each of these values to the *d* metric using

$$d = \frac{2r}{\sqrt{1 - r^2}},$$

and computed the mean, median, and percentiles of this distribution, which is what is reported in Figure 1. This conversion from *r* to *d* is based on an assumption of two equally sized groups, and to the extent that this is not true in actual psychological studies, the *d* values produced by this conversion will be somewhat too small (McGrath & Meyer, 2006). To investigate the extent of underestimation of the *d* values, I repeated the process above using the more general formula

$$d = \frac{r}{\sqrt{(1 - r^2)p_1 p_2}},$$

where p_1 and p_2 are the proportions of participants in the two groups of the study. The values of p_1 and p_2 for each simulated study were based on assuming that participants were randomly assigned to conditions by a binomial process with probability 0.5, and number of trials equal to a typical sample size in experimental psychology (e.g., 30 to 150). The results of this simulation suggested that the degree of underestimation, at least under this assumption of binomial assignment to conditions, is negligible; the average *d* value in this new distribution was 0.46 rather than 0.45.

Hierarchical ordering. The default values of the VPCs suggested by PANGEA are based on the *hierarchical ordering principle*, a concept often invoked in discussions of fractional factorial designs in the literature on design of experiments (Montgomery, 2013). Wu and Hamada (2000) summarize this principle as "(i) lower order effects are more likely to be important than higher order effects, (ii) effects of the same order are likely to be equally

important" (p. 143). For example, consider the counterbalanced design illustrated at the bottom of Table 3, in which we have a fixed (implicit) Treatment factor, a random Participant factor, a random Stimulus factor, and all interactions thereof, all the way up to a three-way Participant × Stimulus × Treatment interaction. The idea of hierarchical ordering is that, on average, we should expect the main effects of Participant, Stimulus, and Treatment to explain more variance in the outcome than the two-way interactions, and we should expect the two-way interactions to explain more variance than the three-way interaction. Anecdotally, this does seem to concord with my own personal experience fitting mixed models to many different datasets in psychology.

As for why hierarchical ordering should tend to occur, one possible explanation is given by Li, Sudarsanam, and Frey (2006), who suggest that this phenomenon is

> "partly determined by the ability of experimenters to transform the inputs and outputs of the system to obtain a parsimonious description of system behavior […] For example, it is well known to aeronautical engineers that the lift and drag of wings is more simply described as a function of wing area and aspect ratio than by wing span and chord. Therefore, when conducting experiments to guide wing design, engineers are likely to use the product of span and chord (wing area) and the ratio of span and chord (the aspect ratio) as the independent variables" (p. 34).

This process described by Li et al. (2006) certainly happens in psychology as well. For example, in priming studies in which participants respond to prime-target stimulus pairs, it is common for researchers to code the "prime type" and "target type" factors in such an experiment so that the classic priming effect is represented as a main effect of prime-target congruency vs. incongruency, rather than as a prime type × target type interaction. And in social psychology, many studies involve a my-group-membership × your-group-membership interaction effect,

which is often better characterized and coded as a main effect of ingroup (group congruency) vs. outgroup (group incongruency). It seems natural to expect random slopes associated with these robust effects to have greater variance than the random slopes of the incidental effects, which are now coded as interactions, and this would give rise to hierarchical ordering.

The way the hierarchical ordering assumption is implemented in PANGEA is as follows. For every estimable source of random variation in the design (i.e., every random variance component) except for the random error term, I count the number of variables that comprise that source. For example, random three-way interactions are comprised of three variables, random two-interactions are comprised of two variables, and random main effects are comprised of one variable. Let this number be n_i for the *i*th variance component. I then reverse these numbers using $n_i' = \max + \min - n_i$, where max and min are the maximum and minimum n_i, respectively, and assign a value of max+1 to the random error variance. Finally I divide all these values by the sum of the values, making them proportions or VPCs. As an example, the counterbalanced design discussed above has the following default VPC values:

$$V_E = 30\%, \quad V_P = 20\%, \quad V_S = 20\%, \quad V_{P\times S} = 10\%, \quad V_{G\times S} = 10\%, \quad V_{P\times B} = 10\%.$$

References

Bakker, M., Dijk, A. van, & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554. http://doi.org/10.1177/1745691612459060

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. http://doi.org/10.1038/nrn3475

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2015). shiny: Web application framework for R (Version 0.11.1). Retrieved from http://CRAN.R-project.org/package=shiny

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. http://doi.org/10.1016/S0022-5371(73)80014-3

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. http://doi.org/10.1037/h0045186

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2 edition). Hillsdale, N.J: Routledge.

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports*, *14*(1), 219–226. http://doi.org/10.2466/pr0.1964.14.1.219

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics*, *27*, 907–949. http://doi.org/10.1214/aoms/1177728067

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. http://doi.org/10.1016/S0022-5371(72)80001-X

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. http://doi.org/10.3758/BF03193146

Fraley, R. C., & Vazire, S. (2014). The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power. *PLoS ONE*, *9*(10), e109019. http://doi.org/10.1371/journal.pone.0109019

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models* (1 edition). Cambridge ; New York: Cambridge University Press.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning Variation in Multilevel Models. *Understanding Statistics*, *1*(4), 223–231. http://doi.org/10.1207/S15328031US0104_02

Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology*, *61*(1), 23–29. http://doi.org/10.1037/h0040561

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin*, *82*(4), 511–518. http://doi.org/10.1037/h0076767

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Med*, *2*(8), e124. http://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated: *Epidemiology*, *19*(5), 640–648. http://doi.org/10.1097/EDE.0b013e31818131e7

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. http://doi.org/10.1037/a0028347

Kenny, D. A. (1994). *Interpersonal Perception: A Social Relations Analysis* (1 edition). New York: The Guilford Press.

Kenny, D. A., & Smith, E. R. (1980). A note on the analysis of designs in which subjects receive each stimulus only once. *Journal of Experimental Social Psychology*, *16*(5), 497–507. http://doi.org/10.1016/0022-1031(80)90054-2

Li, X., Sudarsanam, N., & Frey, D. D. (2006). Regularities in data from factorial experiments. *Complexity*, *11*(5), 32–45. http://doi.org/10.1002/cplx.20123

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163–203. http://doi.org/10.1037/0033-2909.109.2.163

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). SAMPLE SIZE IN PSYCHOLOGICAL RESEARCH OVER THE PAST 30 YEARS [1,2]. *Perceptual and Motor Skills*, *112*(2), 331–348. http://doi.org/10.2466/03.11.PMS.112.2.331-348

Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods*, *9*(2), 147–163. http://doi.org/10.1037/1082-989X.9.2.147

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods*, *11*(4), 386–401. http://doi.org/10.1037/1082-989X.11.4.386

Montgomery, D. C. (2013). *Design and analysis of experiments*. New York: Wiley.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science A Crisis of Confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. http://doi.org/10.1177/1745691612465253

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to Deal with "The Language-as-Fixed-Effect Fallacy": Common Misconceptions and Alternative Solutions. *Journal of Memory and Language*, *41*(3), 416–426. http://doi.org/10.1006/jmla.1999.2650

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173–185. http://doi.org/10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd edition). Thousand Oaks: SAGE Publications, Inc.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213. http://doi.org/10.1037/1082-989X.5.2.199

Richard, F. D., Bond Jr., C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. http://doi.org/10.1037/1089-2680.7.4.331

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, *2*(6), 110–114. http://doi.org/10.2307/3002019

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 551–566. http://doi.org/10.1037/a0029487

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309–316. http://doi.org/10.1037/0033-2909.105.2.309

Snijders, T. A. B., & Bosker, R. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (Second Edition edition). Los Angeles: SAGE Publications Ltd.

Welch, B. L. (1947). The Generalization of `Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, *34*(1/2), 28–35. http://doi.org/10.2307/2332510

Wells, G. L., & Windschitl, P. D. (1999). Stimulus Sampling and Social Psychological Experimentation. *Personality and Social Psychology Bulletin*, *25*(9), 1115–1125. http://doi.org/10.1177/01461672992512005

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science*.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. http://doi.org/10.1037/xge0000014

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical Principles In Experimental Design*. New York: McGraw-Hill.

Wu, C. J., & Hamada, M. S. (2000). *Experiments: Planning, analysis, and optimization*. John Wiley & Sons.