Running head: POWER, DOMINANCE, AND CONSTRAINT

Power, Dominance, and Constraint: A Note on the Appeal of Different Design Traditions.

Jeffrey N. Rouder

Univeristy of California Irvine

University of Missouri

Julia M. Haaf

University of Missouri

Revision, Version 3, 8/31/17

Jeff Rouder

jrouder@uci.edu

Abstract

The recent field-wide emphasis on power has brought the number of participants used in experiments into focus. Cognitive psychologists follow a design tradition where few participants perform many trials each. We ask if one wishes to increase power, is it better to add trials or to add participants. The answer is straightforward—greatest power is achieved by using more people, and the gain from adding people is greater than the gain from adding trials. In light of these results, the cognitive design tradition seems less than ideal. Yet, there are conditions where one may trade people for trials with only a minor decrement in power. Under these conditions, the limiting factor is the trial variability rather than variability across people in the population. These conditions are highly plausible, and we present a stochastic-dominance theoretical argument as to why. We think dominance holds in most cognitive effects, for example, in the Stroop effect. Dominance here is the statement that all people truly identify congruent colors faster than incongruent ones. Under this dominance assumption where everyone's true effect is in the same direction, small mean effects imply a small degree of variability across the population. It is this degree of homogeneity, the consequence of dominance, that licenses the cognitive and psychophysical design traditions.

Power, Dominance, and Constraint: A Note on the Appeal of Different Design Traditions.

The practice of psychological science is going through a period of rapid transition where methodological concerns are front and center. One longstanding, salient concern is that too many experiments are underpowered (Cohen, 1962; Maxwell, 2004; Szucs & Ioannidis, 2017). There are two consequences of underpowered designs. First, if a design is underpowered, then there is an increased chance of failing to detect an effect. When this failure happens, the interpretation is muddied as it is unclear if such failures reflect a lack of power or a truly null effect. Second, and perhaps more perniciously, the prevalence of underpowered studies as a whole is taken as a signal that the literature is not trustworthy. Indeed, it may be an indicator that researchers may be massaging variability to produce significance (Button et al., 2013; Gelman & Loken, 2014; Ioannidis, 2005)

One solution to the problem of underpowered designs is simply to add more participants. As Baumeister (2016) notes, sample sizes have risen over the decades from 10 to 20 to 50, and now to more. Indeed, if one takes power seriously, then experiments for typically small effects should have several hundred observations. Table 1 provides the minimum sample sizes per group for independent and paired *t*-tests at .80 power for a .05 level for Cohen's (1988) small, moderate and large effects. Should experimental psychologists be running hundreds of participants in every experiment?

One of the most fruitful traditions in cognitive psychology is the *psychophysical tradition*, and psychophysical experiments have provided some of the clearest and most persuasive insights. Consider, for example, Blakemore and Campbell's (1969) classic adaptation experiment that first showed orientation-and-frequency selective neural responses. There were only two participants, C. B. and F. W. C., who maybe not so surprisingly have the same initials as the authors. Logan and Cowan's (1984) classic stop process paper, the one that launched a subfield of action control, had only three participants (including G. L.). The first author of the current paper has published twice experiments with only three people (Ratcliff & Rouder, 1998; Rouder, Lu, Speckman, Sun, & Jiang, 2005).

The *psychophysical-design tradition* may be described by three properties: 1. The use of very small numbers of participants; 2. the use of within-subject manipulations; and 3. the use of very large numbers of trials per participant. This tradition may be compared with two other traditions. In the *cognitive-design tradition* there are usually a moderate number of participants, say 20, a mix of within-subject and between-subject manipulations, and moderate numbers of trials per participants, say from 10 to 100. In the *social-psychological design tradition* there are a great many participants, often between-subject groupings, and a handful of observations per participant.

In this paper, we explore the overall power to detect an effect for these traditions in within-subject designs. Suppose, for example, one wishes to partition 2000 observations across 2 conditions. The psychophysical option might be to run two participants for 1000 trials each, dividing those 1000 evenly across the two conditions. Another option, say the cognitive-option might be to run 20 participants with 50 observations in each of the two conditions. Or perhaps we would be best off running 1000 people and gathering a single observation in each condition. For the ensuing formal analysis, we seek the option that has the highest power to detect an effect across the two conditions at a fixed level. There are other criteria for assessing the usefulness of design options of course, but the highest power criterion seems like a good start.

Which Design Tradition Leads to Higher Power?

We evaluate the trade-off between the number of trials per participant and the number of participants as follows: Suppose that participants provide continuously-valued observations in two conditions, generically called *treatment* and *control.* Examples might be a priming experiment where primed and unprimed stimuli are the treatment and control condition, respectively. We assume that each of I participants provides K observations in each condition. Our goal is to provide computations of power as a function of I, the number of participants, and K, the number of observations per participant per condition.

The derivation is a bit long and provided in the section "In Detail" at the end of this manuscript. The key expression for computing power is the *noncentrality* of the test. In brief, the noncentrality describes where the t is centered. The greater the value, the larger the expected t-values and the greater the power. So, by studying the expression for noncentrality, we can gain insight whether it is better to increase the number of participants or the number of trials per participant.

The noncentrality parameter, denoted λ is:

$$\lambda = \sqrt{\mu^2 \times \frac{IK}{K\sigma_\beta^2 + 2\sigma^2}}. \tag{1}$$

The noncentrality depends on three quantities besides I and K. One is μ, which is the size of the effect. It is a population mean—the true average effect across all individuals in the population. There are two variabilities: σ^2 is the variability for replicates, that is for observations from a person in a condition; σ_β^2 is the variability of the true effect across people. These quantities are defined formally in "In Detail."

From Eq. (1), we may reach a preliminary answer to the question of whether adding more people or adding more trials better powers experiments. It may be seen that increasing the number of people, I, always results in greater power than increasing the number of replicates, K. The reason is that although I and K enter into the numerator, K also enters into the denominator. Hence, noncentrality must increase at a greater rate with I than with K. Adding more people results in a more powerful design than adding more trials. The social-psychology design parameters are best from a power perspective.

Unfortunately, the social-psychology design parameters are not as appealing as the cognitive-psychology design parameters. The reason is the cost in money and time of running the experiment. The marginal cost of adding more trials—say asking each participant to run an additional few minutes—is often far less than the marginal cost of recruiting more participants. Before recommending the social-psychology trade-off, it is wise to understand how much power is gained.

We show here that for some cases the gain in power from adding participants rather than adding trials is marginal. Our approach is to pick reasonable values for the inputs in Eq. 1 and study how power changes as the numbers of people and trials are varied. Let's take a priming effect where treatment and control conditions refer to primed and unprimed stimuli, respectively. The dependent measure is response time, and overall, responses are relatively accurate and fast (say under a second). In RT tasks like these, there is a fair amount of variation across trials, and it is not uncommon to see responses as fast as 250 ms and as slow as 1500 ms. A reasonable value of σ, the variability across repeated trials for the same participant in the same condition, is $\sigma = 300$ ms. Next, we consider the overall true effect, μ. Priming effects can be quite subtle, and means are on the order of 40 ms, which is the value we use for μ. The remaining setting is for σ_β, which denotes how variable the effect is across people. Let's consider the case when $\sigma_\beta = 28$ ms, which is reasonable for a 40 ms effect.

The trade-off for this case is shown in Figure 1A. Power to detect this 40 ms effect is plotted for a range of values for numbers of people (I) and replicates (K). The x-axis is the number of trials per condition (K); the lines are for different numbers of participants (I). The critical question is about trading I for K. The points show power for different values of I and K where $I \times K = 1000$ observations in total. These points form an iso-sample-size power curve, and as can be seen, it is better to have large numbers of people and smaller numbers of trials per participant. The size of this effect is relatively minor in Panel A; I and K trade fairly well. If recruiting additional people is resource expensive, then for these settings, it makes sense to forgo these additional people and run additional trials.

Yet, there are cases where the gain in power from the social psychological settings are dramatically better than the cognitive or psychophysical design settings. In these cases, researchers cannot overcome a small number of people with large numbers of trials. Again, we chose inputs for Eq (1) and vary I and K. Let's take a preference-ratings task to assess whether caffeine in soft drinks serve as a flavor enhancer. We take a certain decaffeinated soft drink as a standard, say Decaffeinated Coke. To make a well-controlled caffeinated version, we add in anhydrous caffeine in sufficient amounts to make the flavor noticeably more bitter. Then we ask people to judge how well they like a series of samples, some of which are the decaffeinated Coke and others are the caffeinated Coke. The dependent variable is a Likert rating, say on a 5 point scale, and we assume that repeated ratings for the same person and same beverage have a relatively small variability, say a standard deviation of 1 pt on the Likert scale ($\sigma = 1$). Let's also assume there is a net effect where people prefer the sweeter Coke, and they rate on average the decaffeinated version 1 pt higher ($\mu = 1$). Finally, and critically, lets assume that tastes vary. Some people may prefer the sweetness of the decaffeinated version, others may not have a preference, and still some others may prefer the more bitter, caffeinated version. Therefore, the range of effects of adding the anhydrous caffeine may be large, and we set $\sigma_\beta^2 = 3$ for this case. Figure 1B captures this case, and as can be seen there is no way of trading people for trials. One must use large numbers of people even with large numbers of trials to gain sufficient power.

The key quantity for determining whether people can safely be traded for trials is the variability of the effect in the population, σ^2_β, relative to the trial-by-trial variability, σ^2. When this ratio is small, as it is for the above priming example, people may be traded safely for trials. When this ratio is large, however, as it is for the beverage preference example, then lowering the number of people results in degraded power.

Dominance and Constraint in Action

The previous development shows the appeal of the social psychological design tradition where there are many participants per experiment. Adding observations by adding people results in a better accounting of variability across people more than adding trials. We suspect this dynamic may not be known to many cognitive psychologists.

Fortunately, the loss in power for cognitive-design parameters is marginal when there is little variability in the effect across people relative to the trial-by-trial variability. Given the appeal of the cognitive-design tradition in cost, it is helpful to know whether people can be safely traded for trials. In the remainder, we provide a theoretical argument why we think so in a great many cases.

Dominance Defined

We start by noting that for most effects, there is an anticipated or positive direction. For example, if we startle people, we would expect more intense startle responses for stronger startle stimuli. In fact, perhaps there are no individuals who respond with more intensity to weaker than stronger stimuli. We call this condition where all people have the same ordering *dominance*. In the startle case, for every person the distribution of startle responses for stronger stimuli dominates that for weaker stimuli.

This dominance constraint may well be a hallmark of many tasks. For example, we suspect dominance holds when stimuli differ in strength, say in perception and memory tasks. We also suspect it holds in priming and context tasks. Take, for example, the Stroop task. It is plausible that all individuals truly respond more quickly to congruent than incongruent items, and that no individual truly responds quicker to incongruent than congruent items.

In the above set up, dominance means that each individual's true effect, β_i, is positive for all people. In the Stroop task for example, if $\beta_i > 0$ we say that the ith person identifies incongruent colors truly more slowly than congruent ones. Here the constraint applies to true values (those in the limit of many trials). It does not apply to sample effects. We may observe a few negative sample differences due to noise, but if dominance holds, these sample effects tend to be positive more often than negative. Once sample noise is modeled, the resulting latent true values are positive.

Dominance and Power

Our main result is that if dominance holds, then the experimenter can trade people for items while maintaining power. Figure 2 shows how, and for concreteness, let's apply it to the Stroop case. In Figure 2A, two candidate distributions are placed on β_i, the effect across individuals. One is the normal distribution (dashed) with a mean of 40 ms and a standard deviation of 30 ms. Although normals are popular in linear models, they imply indominance. Some people have $\beta_i > 0$, indicating they obey the usual constraint, e.g., responses to incongruent stimuli are slower than those to congruent ones. But others, a minority for sure, have the opposite ordering! These people are Stroop anomalous in the sense that they truly respond quicker to incongruent than congruent items.

Figure 2A shows an alternative setup that we find far more reasonable for many tasks. We use a skewed distribution over true effects rather than the normal. Importantly, for these skewed distributions, β_i are constrained to be positive for all people. One of the key differences across the indominant and the dominant distributions in Panel A is the relationship between mean and variance. With the normal, there is no relationship. The variability and the mean are both statistically and conceptually independent. With the skewed distributions, in contrast, the mean and the variability are positively related. This fact is shown in Figure 2B. As the effect becomes smaller, not only does the mean decrease but the variability decreases as well. And this is where considerations for power come into play. Effects sizes, the ratio between mean and standard deviation, are constant! When effects are small, the population variation in effects are small. When effects are big, the population variation in effects are big as well. The dominance constraint implies that it is impossible to have small effects with big variances.

Most effects in psychological sciences are relatively small compared to the trial-by-trial variability. The dominance principle implies that the variation across people is relatively small as well. And in this case, researchers may be able to trade people for trials. Figure 2C shows the corresponding power for the skewed distribution in Figure 2A. Close formed values for power are not available; the values shown come from simulations of 10,000 iterations per combination of number of participants and number of trials per condition. The points again show the tradeoff values. As can be seen, there is just a marginal trade-off of power when trading people for trials.

Indominance

Dominance of course need not hold. There may be tasks where some people engage in a different strategy than others or have a different set of preferences. We gave an example previously of whether people prefer the flavor of caffeine in soft drinks. Indominance is plausible here as some will truly prefer the added bitterness of caffeine while others will not.

Strict dominance is not needed to safely trade people for trials. Indeed, the same basic result holds even for indominance so long as the effect is relatively homogeneous across people.

Discussion

In this paper we address whether the main feature in the cognitive and psychophysical tradition—a limited number of people that perform a great many trials—leads to well-powered designs. The answer is nuanced. Adding people *always* leads to higher power than adding trials. Yet, there are conditions under which this gain is marginal, and where researchers can safely use fewer people performing more trials. The key, not too surprisingly, is homogeneity. When people are not too different, then trading people for trials is reasonable. In this case, the limiting factor is trial variability rather than variability across participants.

We make a theoretical argument about why trial variability might be the limiting factor. We start with the concept of dominance. Dominance can be framed as a *does everyone* question. For example, does everyone truly respond more quickly to bright flashes than dim ones. We think for large classes of effects, this *everyone does* constraint holds. All people have a true positive effect, and the consequence of this restriction is that the variance across people cannot be arbitrarily large for small average effects. This limit, a lower limit on the population effect size, implies in turn that trial variability rather than population variability is the limiting factor when exploring typically small effects. And, it is precisely in these cases that researchers can safely use the cognitive and psychophysical design parameters.

We think the concept of dominance is more than just a handy assumption to support trading people for trials. It helps shape the scientific inquiry. If dominance holds, the mean effect captures roughly what all people do, and the focus can remain on this mean. If dominance does not hold, however, the mean becomes less useful. Different people are qualitatively affected differently by the manipulation, and the critical question is why.

We think psychologists have a pretty good idea about whether effects are dominant or not. Some effects, say stimulus strength or common context effects will plausibly be dominant. Other effects, say preference effects, will plausibly be indominant. These lead to a rule-of-thumb. In cases where dominance is likely, for example in priming and in Stroop, then researchers can safely trade people for trials. In cases where it is not, then many people need to be run.

Formally, assessing dominance is not so easy. It requires many trials and many people. Even with sufficient data, assessing many order constraints simultaneously is difficult in classical frameworks (Silvapulle & Sen, 2011). Fortunately, it is convenient in the Bayesian framework (Gelfand, Smith, & Lee, 1992), and Haaf & Rouder (2017) provide a Bayes factor solution for the setup we consider here. We hope a focus on dominance becomes more timely and topical in experimental psychological science.

In this paper, we focus on the frequentist notion of power. Yet, in most of our work, we recommend Bayesian inference (Rouder, Morey, & Wagenmakers, 2016). We use the frequentist power in a limited sense. Our main analysis is of the noncentrality parameter, and this parameter plays the same role in Bayesian and frequentist analysis. The larger it is when there are effects, the easier it is to state definitive evidence for null and alternative hypotheses when each holds. The results here hold regardless of whether one uses Bayesian or frequentist inference; these results are about how design parameters affect the resolution of effects.

In Detail

To derive expressions for noncentrality, we start with a statistical model on observations. Let Y_{ijk} denote the kth replicate for the ith participant in the jth condition. We model Y_{ijk} as

$$Y_{ijk} \sim \text{Normal}(\nu_{ij}, \sigma^2). \tag{2}$$

We refer to σ^2 as *trial variability*—it is the variability across replicate trials from the same condition for the same participant.

We follow an approach similar to ANOVA. The true cell means ν_{ij} can be expressed as:

$$\begin{aligned}
\nu_{i1} &= \alpha_i - \beta_i/2, \\
\nu_{i2} &= \alpha_i + \beta_i/2.
\end{aligned}$$

Here, α_i is the true participant-specific overall mean across both conditions; β_i is the true participant-specific effect, which is the main target of interest. We treat α_i and β_i as random effects that are normally distributed:

$$\begin{aligned}
\alpha_i &\sim \text{Normal}(\mu_\alpha, \sigma^2_\alpha), \\
\beta_i &\sim \text{Normal}(\mu, \sigma^2_\beta).
\end{aligned}$$

The term μ is the true population effect. It is the population mean of effects across individuals. The term σ^2_β is called the *population variability*—it is the variability of true effects across people.

With these specifications, conditional sampling distributions on individual's sample means may be derived:

$$\begin{aligned}\bar{Y}_{i1}|\alpha_i,\beta_i &\sim \text{Normal}(\alpha_i-\beta_i/2,\sigma^2/K),\\ \bar{Y}_{i2}|\alpha_i,\beta_i &\sim \text{Normal}(\alpha_i+\beta_i/2,\sigma^2/K).\end{aligned}$$

The difference, d_i may be defined as $d_i = barY_{i2} - \bar{Y}_{i1}$. Conditional on α_i and β_i, it is $d_i|\alpha_i,\beta_i \sim \text{Normal}(\beta_i, 2\sigma^2/K)$. Note that the dependence on α_i has dropped out. Marginalizing this random variable across β_i yields

$$d_i \sim \text{Normal}\left(\mu, \sigma^2_\beta + \frac{2\sigma^2}{K}\right).$$

The resulting t value of the paired t-test is distributed as

$$t \sim \text{T}\left(I-1, \frac{\sqrt{I}\mu}{\sqrt{\sigma^2_\beta + 2\sigma^2/K}}\right).$$

The critical quantity, the noncentrality parameter, is

$$\lambda = \frac{\sqrt{I}\mu}{\sqrt{\sigma^2_\beta + 2\sigma^2/K}}.$$

Readers more familiar with the regression approach to ANOVA may wonder why there is no reference to the correlation in the above derivation. The answer is that there are many ways to skin a cat, and whether correlation is treated as a parameter or derived from the primary parameters is a matter of choice. We parameterize the noncentrality parameter directly by decomposing cells $(\bar{Y}_{i1}, \bar{Y}_{i2})$ into a common component, α_i, and an effect component, β_i. Noncentrality is based on the effect, and the common component, α_i, drops out as it did above. Hence, one needs as inputs only the mean effect (μ), the variance of the effect across people (σ^2_β), and the trial noise of the sample mean (σ^2/K).

In the regression framework, however, cells are not decomposed this way. A different parameterization is chosen, and the sample means are modeled generically as:

$$\begin{pmatrix} \bar{Y}_{i1} \\ \bar{Y}_{i2} \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right],$$

where N_2 is a bivariate normal. In this formulation, parameters μ_i, μ_2, σ_1, σ_2 and ρ serve as inputs. To derive noncentrality, one must compute the mean and standard deviation of the difference $d_i = \bar{Y}_{i2} - \bar{Y}_{i1}$ from these inputs, and ρ is used in this computation as $\text{SD}(d_i) = \sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}$.

We can place our setup in this regression framework as well:

$$\begin{pmatrix} \bar{Y}_{i1} \\ \bar{Y}_{i2} \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_\alpha - \frac{\mu}{2} \\ \mu_\alpha + \frac{\mu}{2} \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{K} + \sigma_\alpha^2 + \frac{\sigma_\beta^2}{4} & \sigma_\alpha^2 - \frac{\sigma_\beta^2}{4} \\ \sigma_\alpha^2 - \frac{\sigma_\beta^2}{4} & \frac{\sigma^2}{K} + \sigma_\alpha^2 + \frac{\sigma_\beta^2}{4} \end{pmatrix} \right].$$

Here, there is correlation, but rather than being primary, it is derived as a function of the parameters:

$$\rho = \frac{\sigma_\alpha^2 - \frac{\sigma_\beta^2}{4}}{\frac{\sigma^2}{K} + \sigma_\alpha^2 + \frac{\sigma_\beta^2}{4}}$$

The choice between two different parameterizations is reflected in G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) as well. There are two different input methods for the calculating the power of a paired t-test. The first, based on differences, corresponds to our parameterization. It does not require input of the correlation. The second, based on a regression parameterization, requires correlation as input, and this correlation is used to compute the variability of the difference between sample means.

We prefer our parameterization not only because it is more direct, but because it is easier to reason with. We have strong intuitions about the input parameters, the size of trial variation (σ^2), the size of effects (μ), and the variability of these effects across people

(σ^2_β). We have almost no intuition about the inputs in the regression case, and feel that they are difficult to reason with.

References

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology*, *66*, 153 - 158. Retrieved from http://www.sciencedirect.com/science/article/pii/S002210311600007X (Rigorous and Replicable Methods in Social Psychology)

Blakemore, C. T., & Campbell, F. W. (1969). On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *The Journal of physiology*, *203*(1), 237.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013, apr). Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*, *advance online publication*. Retrieved from http://dx.doi.org/10.1038/nrn3475

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145.

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

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods*, *39*, 175-191.

Gelfand, A. E., Smith, A. F. M., & Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*(418), 523–532. Retrieved from http://www.jstor.org/stable/2290286

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460.

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*. (in press)

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 0696-0701.

Logan, G. D., & Cowan, W. B. (1984). On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review*, *91*(3), 295.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, *9*, 147-163.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for decisions between two choices. *Psychological Science*, *9*, 347-356.

Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin and Review*, *12*, 195-223.

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, *2*, 6. Retrieved from `http://doi.org/10.1525/collabra.28`

Silvapulle, M. J., & Sen, P. K. (2011). *Constrained statistical inference: Order, inequality, and shape constraints* (Vol. 912). John Wiley & Sons.

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3), e2000797.

Author Note

J.N.R. and J.M.H. conceptualized the project, derived the expressions, and wrote the manuscript. Email: jrouder@uci.edu, Web: pcl.missouri.edu; Twitter: @JeffRouder.

Table 1

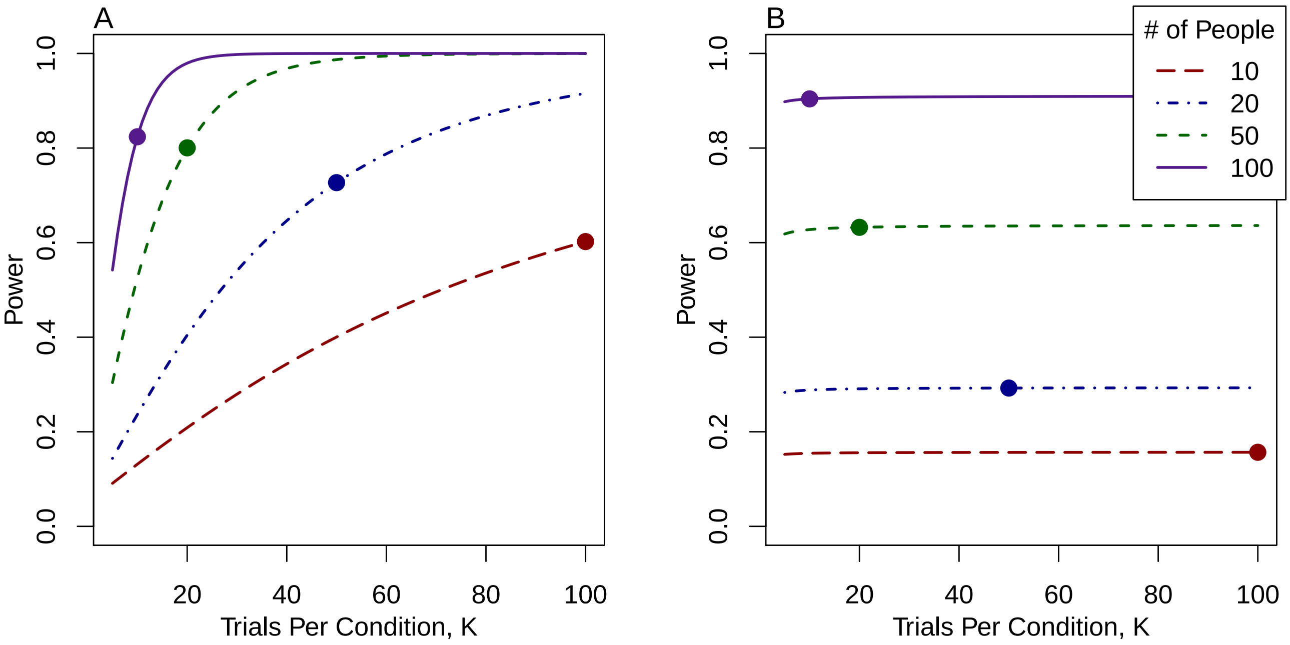
Minimum sample sizes per group for independent and paired t-tests for small, medium and large effects. Here, $\alpha = .05$ and the power is fixed at .8.

	d = 0.2	d = 0.5	d = 0.8
Two samples	394	64	26
Paired	199	34	15

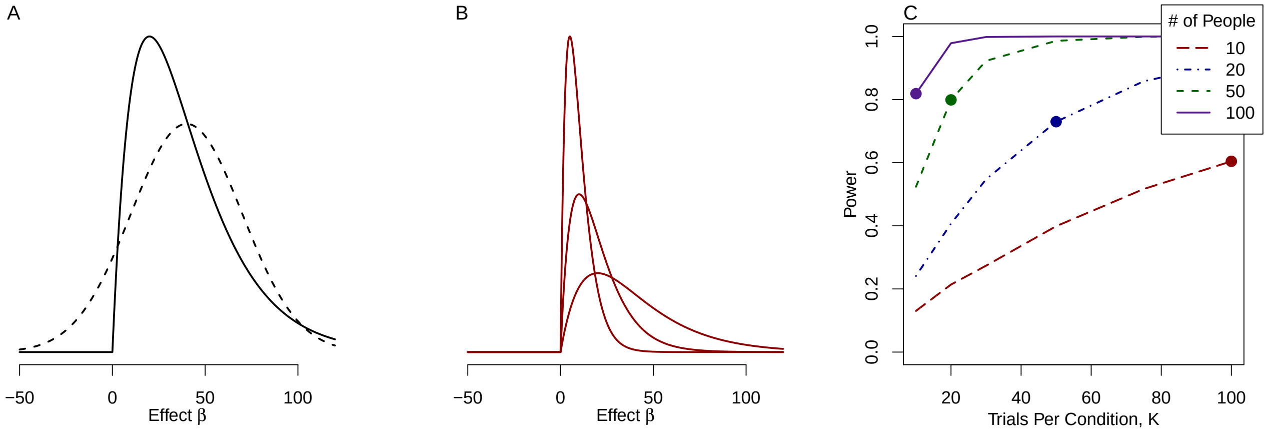
Figure Captions

Figure 1. Power as a function of I, the number of participants, and K, the number of observations per condition per person. **A**. Population variability is small relative to trial variability ($\sigma_\beta = 28$ ms, $\sigma = 300$ ms). **B**. Population variability is large relative to trial variability ($\sigma_\beta = 3000$ ms, $\sigma = 1000$ ms).

Figure 2. Stochastic dominance and power. **A**. The normal distribution over true individual effects is indominant as the distribution has mass on both positive and negative effects. The skewed distribution is dominant because there are no individuals with true negative effects. **B.** For some families of skewed distributions, in this case the gamma family, the mean and standard deviation are proportional; small means correspond to small standard deviations. **C**. Power for a 40 ms effect with skewed distribution. The consequence of this dominance setup is that researchers may trade people for trials.


A
B
Power
1.0
0.8
0.6
0.4
0.2
0.0
20
40
60
80
100
Trials Per Condition, K
# of People
10
20
50
100

Power, Dominance, and Constraint, Figure 2


A
Effect β
B
Effect β
C
Power
Trials Per Condition, K
# of People