Running head: POWER, DOMINANCE, AND CONSTRAINT

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

Jeffrey N. Rouder & Julia M. Haaf

University of Missouri

Original Submission, Version 2, 5/24/17

Jeff Rouder

rouderj@missouri.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 noise 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 noise 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. Let Y_{ijk} denote the kth replicate, $k = 1, \ldots, K$, for the ith participant, $i = 1, \ldots, I$, in the jth condition, $j = 1, 2$, for control and treatment, respectively.

The usual course of analysis is to aggregate across replicates to produce $\bar{Y}_{ij}$, a participant-by-condition sample mean. Then, these sample means are submitted to a paired t-test. In the paired t-test, the difference, d_i, the participant's observed effect given by $d_i = \bar{Y}_{i2} - \bar{Y}_{i1}$, is modeled as a normal:

$$d_i \sim \text{Normal}(\mu_d, \sigma_d^2). \tag{1}$$

The question answered is whether the data are discordant with the supposition that $\mu_d = 0$. The t-statistic in this case is

$$t = \frac{\sqrt{I} \times \bar{d}}{s_d},$$

where $\bar{d}$ and s_d are the sample mean and sample standard deviation, respectively, of the observed participant effects, $d_1, \ldots, d_I$.

The key quantity for power is not the t statistic in any given experiment, but the distribution of t values across repeated experiments with the same design parameters. The distribution of the t statistic follows the noncentral T distribution:

$$t \sim \text{T}\left(I - 1, \frac{\sqrt{I}\mu_d}{\sigma_d}\right),$$

where the first argument, $I - 1$, is the familiar degrees of freedom. The second argument is known as the noncentrality parameter and it is critical for understanding power. We denote it as λ, where for this case,

$$\lambda = \frac{\sqrt{I}\mu_d}{\sigma_d} = \sqrt{I}\delta_d,$$

where $\delta_d = \mu_d/\sigma_d$ is the standardized true effect size. The noncentrality parameter provides insight into the power of the test. The larger the noncentrality, the more powerful the test.[1] Figure 1 shows the relationship between power and I, the number of participants for a few values of the true effect size, δ_d.

Unfortunately, from this development it is not obvious how power depends on the number of replicates K. To understand the role of replicates, it is helpful to consider a model on the observations themselves rather than the observed difference: We model Y_{ijk} as

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

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

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 latent data that are normally distributed:

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

The term σ^2_β is called the *population noise*—it is the variablility of true effects across the population.

With these specifications, conditional sampling distributions on 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, $d_i = \bar{Y}_{i2} - \bar{Y}_{i1}$ is distributed as $d_i|\beta_i \sim \text{Normal}(\beta_i, 2\sigma^2/K)$. Marginalizing this random variable across β_i yields

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

The resulting t value is distributed as

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

The critical quantity, the noncentrality parameter, is

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

After algebraic rearrangement, the noncentrality can be expressed as

$$\lambda = \sqrt{\mu_\beta^2 \times \frac{IK}{K\sigma_\beta^2 + 2\sigma^2}}. \tag{3}$$

From Eq. (3), it is clear 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 K.

The effectiveness of increasing K depends on the amount of trial noise (σ) and population noise (σ_β). When population noise is relatively small, that is all people have a similarly sized effect, then increasing K is about as effective as increasing I. However, when the population noise is relatively large, increasing K may have a limited effect.

This differential behavior is shown in Figure 2. We take a case in response time in a simple task where there is a true 40 ms effect. Power to detect this effect is plotted for a range of values for I and K. In Panel A, there is relatively small population noise ($\sigma_\beta = 28$ ms, $\sigma = 300$ ms); in Panel B, there is relatively large population noise ($\sigma_\beta = 120$ ms, $\sigma = 100$ ms). Th 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. 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 where I and K trade fairly well but is dramatic in Panel B where increases in K cannot make up for decreases in I.

Dominance and Constraint in Action

The previous development showed the appeal of the social psychological design tradition where there are many participants per experiment. In retrospect, the reason is obvious. Adding observations by adding people results in a better accounting of variability across people. Adding observations by adding trials does not. We suspect this dynamic may not be known to most cognitive psychologists.

Unfortunately, the social-psychology design parameters are not as appealing as the cognitive-psychology design parameters. The reason is the cost of running the experiments in money and time. 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. Fortunately, there are conditions under which the cognitive and psychophysical design traditions remain powerful. This happens when the population noise is small relative to the trial noise. In this case, the loss of power by trading number of trials for number of participants may be marginal. Given the appeal of these design traditions in terms of cost and ability to address higher-order properties, it is helpful to know whether these conditions hold. In the remainder, we provide a theoretical argument why we think so; that is, why trial noise is the limiting factor, and consequently, why researchers can often increase their power without hassle by increasing the numbers of replicate trials.

We start with the concept of *stochastic dominance*, or dominance for short. Dominance implies that distributions of observations have a set order. Take, for example, the responses to bright and dim flashes of light. Responses to bright flashes are faster than those to dim ones. The appropriate constraint here is dominance. Not every response to the bright flash is necessarily faster than every response to the dim one. Instead, we say the distribution dominates, that is, at every probability index, the corresponding quantile for the bright flash is smaller in value than that for the dim flash.

We think it is plausible that this dominance holds for all people. For example, for any individual, the pth percentile of response time to the bright flash is less than the pth percentile to the dim flash.

Dominance may well be a hallmark of many tasks in cognition. For example, we suspect dominance holds when stimuli differ in strength. We also suspect it hold in priming and context tasks. Take, for example, the Stroop task. It is plausible that for all individuals, responses to incongruent stimuli dominate (are greater in response time) than responses to congruent ones.

In the above set up, dominance implies $\beta_i \geq 0$, 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 term "truly" or "true" refers to the limit of infinite trials, and in any finite sample, the sample difference d_i may be positive or negative though in dominant cases they tend to be positive more often than negative. If one makes the normal assumptions above, it is possible to construct principled tests of whether all people show the dominances constraint, and Haaf & Rouder (2017) develop the Bayes factor approach.

Dominance has implications for power, and in particular, if dominance holds, then power may be relatively unaffected by increasing trials rather than increasing participants. Figure 3A shows how. In the figure, several candidate distributions are placed on β_i. One is the normal distribution, and it is indominant. 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 pathological in the sense that in the limit of trials, they average quicker responses in the incongruent than congruent conditions.

We are unaware that anyone has documented such pathological individuals in common cognitive tasks. To our knowledge there is no strange lesion or unappealing genetic variation that renders people quicker to identify incongruent Stroop stimuli than congruent ones or quicker to detect dim flashes than bright ones. So it is with some surprise that the usual normal model, the one shown in Panel A, undergirds hierarchical linear models and structural equation models. Models that place a normal on latent effects imply that some people have true pathologies of this type.

Figure 3A also shows an alternative setup that we find far more reasonable for cognitive tasks. We use skewed distributions over true effects rather than the normal. Importantly, for these skewed distributions, β_i are constrained to be positive for all people. The different distributions show the case for differently sized effects. For example, the distribution over the smaller values, dashed, might correspond to a small difference in luminance between a bright and dim flash; the distribution over the larger values, solid, might correspond to a large difference in luminance.

One of the key differences across the indominant and the dominant distributions in Panel A is the relationship between the overall effect and the variability of the effect in the population. With the normal, there is no relationship. The variability and the overall mean are both statistically and conceptually independent. With the skewed distributions, in contrast, the overall mean and the variability are positively related. And this is where considerations for power come into play. For these distributions, the mean and standard deviation are proportional, that is $\text{E}(\beta) = \tau \times \text{SD}(\beta)$, where τ is a constant. The remarkable implication of this setup is that the true effect size, δ_β, is τ, and it does not change with task variables! 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! One of the implications of this dominance principle is that population effect sizes, δ_β cannot be too small.

Figure 3B shows corresponding power for the solid-line skewed distribution in

Figure 3A. Close formed values for power are not easily attainable; the values shown come about 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, and as can be seen, power is relatively stable when one trades participants for trials. The figure provides the demonstration that cognitive and psychophysical designs may be well powered without hundreds of participants.

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 trials for people is reasonable. In this case, the limiting factor is trial noise rather than noise across participants.

We make a theoretical argument about why trial noise might be the limiting factor. We start with a concept of stochastic dominance. In distribution, responses in one condition might dominate that from another. 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 cognitive 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 limit on the population effect size, implies in turn that trial noise rather than population noise 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.

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.

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 aptitudes. Perhaps the clearest example is handedness. Take, for example, the distance a ball may be thrown with the left and right hand. Most people have a right hand advantage, that is, the distances thrown with their right hand dominate those thrown wit the left hand. But this dominance does not hold for all people as there are left-handed individuals in the population. The question then is how to tell if dominance holds. Haaf & Rouder (2017) describe a Bayes factor approach to the question, and using this approach they are able to show dominance in a series of Stroop tasks. Whether dominance holds more broadly as we suspect remains timely and topical.

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.

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.* (Revision submitted 3/17)

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 sciencecollabra. *Collabra*, *2*, 6. Retrieved from `http://doi.org/10.1525/collabra.28`

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: rouderj@missouri.edu, Web: pcl.missouri.edu; Twitter: @JeffRouder.

Footnotes

[1] Power of a two-tailed test at the .05 level is given as follows: Let $f(t, \nu, \lambda)$ be the density of the noncentral T distribution evaluated at t with ν degrees of freedom and noncentrality λ. Let c be the critical t-value, i. e., $\int_c^\infty f(t, I-1, 0)dt = .025$. Then the power is given by $\int_c^\infty f(t, I-1, \lambda)dt + \int_{-\infty}^{-c} f(t, I-1, \lambda)dt$.

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.

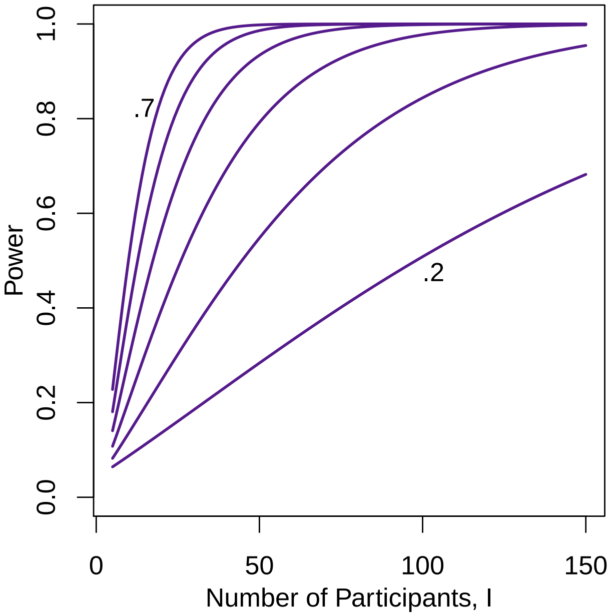
	d = 0.2	d = 0.5	d = 0.8
Two samples	394	64	26
Paired	199	34	15

Figure Captions

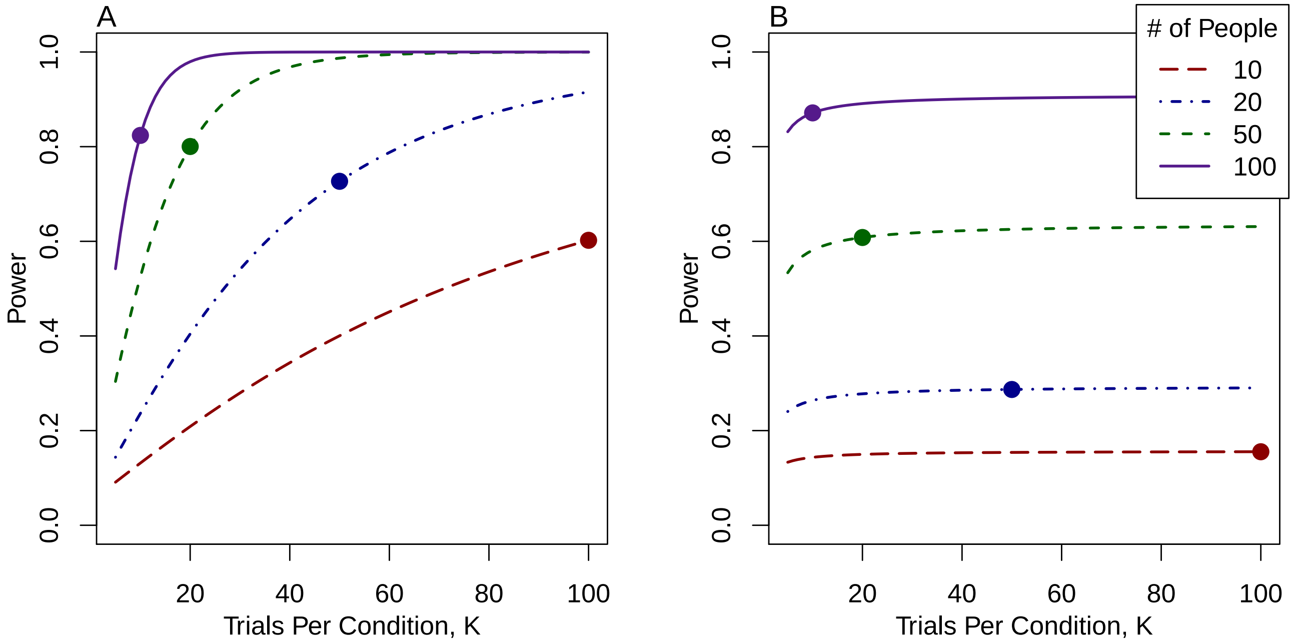
Figure 1. Power as a function of the number of participants for true effect sizes .2, .3, .4, .5, .6, and .7.

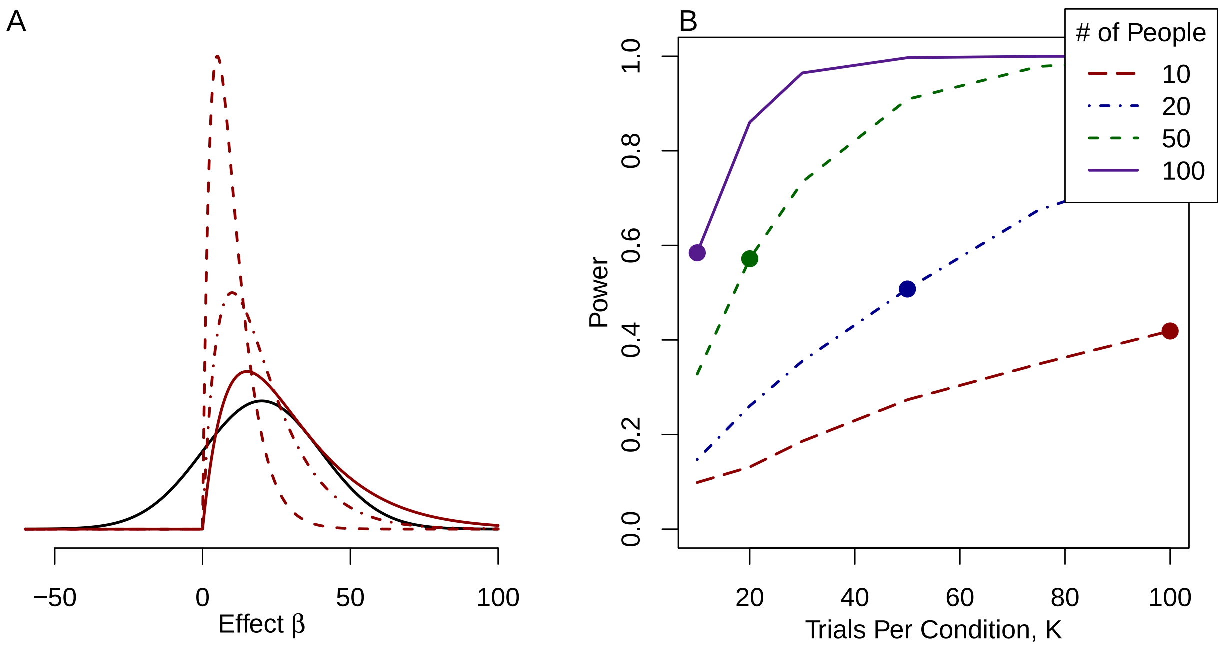
Figure 2. Power as a function of I, the number of participants, and K, the number of observations per condition per person. **A**. Population noise is small relative to trial noise ($\sigma_\beta = 28$ ms, $\sigma = 300$ ms). **B**. Population noise is large relative to trial noise ($\sigma_\beta = 120$ ms, $\sigma = 100$ ms).

Figure 3. 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 family of gamma distributions are all dominant as each individual has a positive effect. All of these gamma distributions have the same ratio of mean-to-standard-deviation, that is, the same effect size. **B**. The consequence of this dominance setup is that high power may be achieved in cognitive and psychophysical designs.


.7
.2
Power
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Number of Participants, I

Power, Dominance, and Constraint, Figure 2




A
−50
0
50
100
Effect β
B
# of People
10
20
50
100
Power
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100
Trials Per Condition, K