Journal of Experimental Psychology: General
2014, Vol. 143, No. 5, 2020–2045

© 2014 American Psychological Association
0096-3445/14/$12.00 http://dx.doi.org/10.1037/xge0000014

Statistical Power and Optimal Design in Experiments in Which Samples of Participants Respond to Samples of Stimuli

Jacob Westfall
University of Colorado Boulder

David A. Kenny
University of Connecticut

Charles M. Judd
University of Colorado Boulder

Researchers designing experiments in which a sample of participants responds to a sample of stimuli are faced with difficult questions about optimal study design. The conventional procedures of statistical power analysis fail to provide appropriate answers to these questions because they are based on statistical models in which stimuli are not assumed to be a source of random variation in the data, models that are inappropriate for experiments involving crossed random factors of participants and stimuli. In this article, we present new methods of power analysis for designs with crossed random factors, and we give detailed, practical guidance to psychology researchers planning experiments in which a sample of participants responds to a sample of stimuli. We extensively examine 5 commonly used experimental designs, describe how to estimate statistical power in each, and provide power analysis results based on a reasonable set of default parameter values. We then develop general conclusions and formulate rules of thumb concerning the optimal design of experiments in which a sample of participants responds to a sample of stimuli. We show that in crossed designs, statistical power typically does not approach unity as the number of participants goes to infinity but instead approaches a maximum attainable power value that is possibly small, depending on the stimulus sample. We also consider the statistical merits of designs involving multiple stimulus blocks. Finally, we provide a simple and flexible Web-based power application to aid researchers in planning studies with samples of stimuli.

Keywords: statistical power, experimental design, optimal design, stimulus sampling, mixed models

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Studies in which samples of participants respond to samples of stimuli in various experimental conditions are ubiquitous in experimental psychology. In studies of memory, participants often memorize lists of words that are drawn from a larger corpus of words. In studies of social cognition, participants often make judgments about sets of faces or read vignettes about hypothetical persons. In studies of emotion, participants are often exposed to photographs or film clips of emotion-provoking scenes. In all of these examples, the particular individual stimuli used are not of intrinsic theoretical interest, except insofar as they instantiate the general categories to which they belong (e.g., monosyllabic common words; African American male facial photos). What are typically of interest are any differences in the responses given in the different conditions to those stimuli (e.g., better word memory in one condition than another; more negative judgments of African American faces than of White faces).

In designing such studies, researchers are faced with a variety of design decisions concerning participants, stimuli, and the specifics of the design used. There is a series of decisions to be made about participants: Should one sample participants broadly to increase generalization, or should one restrict the sample of participants used to maximize power? Should one use a design where participants are only in one condition, or should a within-participant design be used? And finally, how many participants should be run? These are decisions that researchers realize have consequences for statistical power and the kinds of generalizations that are permitted from the subsequently collected data.

Similar decisions are made, although too often in a cursory manner, about stimuli: How variable should the stimuli be? How many stimuli should be used? And should stimuli be different in the different conditions or should the same stimuli be used across conditions? Unfortunately, researchers often do not realize that these decisions also have major consequences for statistical power and the kinds of generalizations that they permit. All too often, these decisions about stimuli are based more on the traditional practices within an experimental paradigm rather than well-thought-out principles about the optimal design of experiments.

This article focuses on issues of statistical power and optimal design in experiments in which samples of participants respond to

This article was published Online First August 11, 2014.

Jacob Westfall, Department of Psychology and Neuroscience, University of Colorado Boulder; David A. Kenny, Department of Psychology, University of Connecticut; Charles M. Judd, Department of Psychology, University of Colorado Boulder.

We thank Markus Brauer, Deborah Kashy, John Lynch, Dominique Muller, Gary McClelland, Thom Baguley, and members of the Stereotyping and Prejudice Lab at the University of Colorado for helpful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Jacob Westfall, Department of Psychology and Neuroscience, University of Colorado, Boulder, CO 80309-0345. E-mail: jake.westfall@colorado.edu

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

samples of stimuli in different conditions. Assuming that one seeks to generalize conclusions about condition differences to future studies that might be conducted, then one ought to be concerned about generalization not only to other samples of participants but also to other samples of stimuli that might have been used. Our goal is to provide the tools to enable researchers to think in an informed manner about design decisions involving not only samples of participants but also samples of stimuli. We consider a range of possible designs that might be used. We consider the power implications of each, and we consider the power implications of the number of stimuli as well as the number of participants.

Treating Participants and Stimuli as Random

In an earlier article (Judd, Westfall, & Kenny, 2012), we considered one particular design in which a sample of participants responds to a sample of stimuli. The hypothetical example used in that article involved participants giving responses to a series of facial photographs of White and African American male targets. Target ethnicity was the experimental factor of theoretical interest (i.e., were different responses given to the White targets on average compared to the African American targets?). Accordingly, in this design participant was crossed with ethnicity (condition), but stimuli were either of one ethnicity or the other.

Most typically when faced with data from such a design researchers conduct what we called a *by-participant* analysis, analyzing two means for each participant—one for the White targets and one for the African American targets—and testing whether the mean within-participant difference in these two means is significantly different from zero. Such an analysis implicitly treats participants as a random factor but does not treat stimuli as random, thus permitting generalization to other samples of participants but not to other samples of stimuli. Judd et al. (2012) and others have shown that this analysis gives rise to seriously inflated Type I error rates if in fact one seeks generalization not only to other samples of participants but also to other samples of stimuli (e.g., particular White and African American target photos) that might be used (Baayen, Davidson, & Bates, 2008; Clark, 1973; Coleman, 1964; Santa, Miller, & Shaw, 1979).

Traditionally, if both participants and stimuli were treated as random factors in the analysis, an analytic approach involving quasi-*F* ratios was the only practical solution, as outlined by Clark (1973) and many others. What Judd et al. (2012) made clear (as others had previously; see Baayen et al., 2008) was that an analysis based on linear mixed models could now be easily implemented, producing appropriate Type I error rates assuming generalization was sought across both random factors—participants and stimuli. Many of the shortcomings of the traditional quasi-*F* methods (e.g., restrictive assumptions about the study design, complicated design-specific derivations, lack of availability in major statistical software packages) are avoided by the more modern approach based on linear mixed models.

Although treating stimuli as random through the appropriate use of linear mixed models does bring the benefit of increased generalizability, it often does so at the cost of lower statistical power: More robust and general inferences must necessarily be supported by stronger statistical evidence. Thus, it is reasonable to expect that if we wish to purchase increased generalizability from our experiments, we must be prepared to pay at least some statistical power cost. However, in numerous studies appearing in the literature today, the power costs associated with treating stimuli as random are needlessly exacerbated by poor design choices. More specifically, in designs in which samples of participants respond to samples of stimuli, researchers typically ignore the power implications of the stimuli they use—both the variability of the stimuli and their number. This neglect of stimulus sampling (Wells & Windschitl, 1999) has led to many inefficiently designed studies. This unfortunate fact is particularly salient in light of the growing awareness in psychology and other experimental sciences of the problems created and maintained by a proliferation of chronically underpowered studies (Asendorpf et al., 2013; Bakker, van Dijk, & Wicherts, 2012; Button et al., 2013; Ioannidis, 2008; Schimmack, 2012).

Our goal in the rest of this article is to make researchers aware of the power implications of stimulus sampling when they appropriately treat stimuli as a random factor. We offer guidance about how one optimally designs studies with random factors of participants and stimuli to increase the probability that any condition differences that are found can in fact be replicated in future studies that employ different participant and different stimulus samples. We consider a broad range of designs in which samples of participants respond to samples of stimuli. For each of these, we develop appropriate linear mixed models that treat both factors as random. We further develop effect size and power estimation procedures for all of these designs, providing illustrative power results as a function of the numbers of participants and stimuli, and their variances. Additionally, we provide a Web-based application that permits generalization of these power results to any specific set of assumptions about designs, numbers of participants and stimuli, and the relevant variances. Finally, we consider more general issues of optimal design in research endeavors in which samples of participants respond to a sample of stimuli in different conditions.

Designs With Crossed Participant and Stimulus Random Factors

In considering the range of possible designs, we focus on designs in which participants are crossed with stimuli, as opposed to designs in which stimuli are nested in participants or vice versa. That is, we consider designs in which multiple participants respond to the same stimuli. We made the decision to limit the range of designs that we discuss in detail in this article because crossed designs are used more frequently than nested designs in experimental psychology. Additionally, there is an extensive literature in education and applied statistics on nested designs, commonly referred to as *hierarchically nested designs*, and this literature has considered issues of power and optimal design (Raudenbush, 1997; Raudenbush & Liu, 2000; Snijders, 2001; Snijders & Bosker, 1993). We know of no work that has addressed issues of statistical power in designs with crossed random factors.

In the following paragraphs, we define the five experimental designs considered in this article. For these designs, we assume a single experimental manipulation of interest having two levels. This manipulation, which we refer to as *Condition*, constitutes the only fixed factor in the design, differentiating it from the random

factors of stimuli and participants. A second assumption is that there is only a single replication in the design; that is, each participant responds to any individual stimulus within a condition only one time.

Table 1 provides schematics that define the five designs. Each design is described by a matrix in this table, with participants defining the rows and stimuli defining the columns. The cells of these matrices indicate the levels of condition (A and/or B) under which particular observations occur. If a cell has a dash, it means that that observation (a particular participant paired with a particular stimulus) is not collected.

Table 1
Schematics for Five Experimental Designs

	Stimuli					
Participants	1	2	3	4	5	6
Fully crossed design						
1	AB	AB	AB	AB	AB	AB
2	AB	AB	AB	AB	AB	AB
3	AB	AB	AB	AB	AB	AB
4	AB	AB	AB	AB	AB	AB
5	AB	AB	AB	AB	AB	AB
6	AB	AB	AB	AB	AB	AB
Counterbalanced design						
1	A	A	A	B	B	B
2	A	A	A	B	B	B
3	A	A	A	B	B	B
4	B	B	B	A	A	A
5	B	B	B	A	A	A
6	B	B	B	A	A	A
Stimuli-within-condition design						
1	A	A	A	B	B	B
2	A	A	A	B	B	B
3	A	A	A	B	B	B
4	A	A	A	B	B	B
5	A	A	A	B	B	B
6	A	A	A	B	B	B
Participants-within-condition design						
1	A	A	A	A	A	A
2	A	A	A	A	A	A
3	A	A	A	A	A	A
4	B	B	B	B	B	B
5	B	B	B	B	B	B
6	B	B	B	B	B	B
Both-within-condition design						
1	A	A	A	—	—	—
2	A	A	A	—	—	—
3	A	A	A	—	—	—
4	—	—	—	B	B	B
5	—	—	—	B	B	B
6	—	—	—	B	B	B

Note. Each schematic illustrates which participant (labeled 1 to 6) views which stimulus (labeled 1 to 6) under which condition (labeled A or B). AB means that this participant responds to this stimulus under both conditions; A means that this participant responds to this stimulus only under Condition A; B means that this participant responds to this stimulus only under Condition B. A dash means that this participant never responds to this stimulus.

The first design is the fully crossed design in which every participant responds to every stimulus twice, once in each condition. In the matrix for this design, every participant by stimulus cell contains both A and B, indicating that the participant–stimulus pair occurs in both conditions.

The second design in Table 1 is the counterbalanced design in which, for half of the participants, each stimulus is in either condition A or B, and for the other half, the stimulus is responded to in the other condition.

The third design is the stimuli-within-condition design in which each stimulus is responded to in only one of the two conditions, although participants are crossed with condition. This is the design that was the focus of Judd et al. (2012).

The fourth design, the participants-within-condition design, reverses the roles of stimuli and participants in that each participant responds in only one of the two conditions, but stimuli are responded to in both conditions, albeit by different participants.

Finally, the fifth design is the both-within-condition design. Here both participants and stimuli are nested under condition, and within each condition each participant responds to every stimulus.

Estimating Power in Designs With Crossed Participant and Stimulus Random Factors

The starting point for conducting a power analysis is to state explicitly the model underlying the observations in these designs, making explicit all the sources that contribute to variation in those observations. To do this, we assume that we have two random factors, one involving a sample of p participants and the other involving a sample of q stimuli. Additionally there is a single fixed condition factor, c, with two levels. In this article, we consider experiments where participants respond to each stimulus at most only once per condition; in other words, we assume only a single replication at the lowest level of observation.

Under these assumptions the fully specified mixed model for the response of the ith participant to the jth stimulus in the kth condition is

$$y_{ijk} = \beta_0 + \beta_1 c_k + \alpha_i^P + \alpha_i^{P\times C} c_k + \alpha_j^S + \alpha_j^{S\times C} c_k + \alpha_{ij}^{P\times S} + \varepsilon_{ijk},$$

$$\text{var}\left(\alpha_i^P\right) = \sigma_P^2, \quad \text{var}\left(\alpha_i^{P\times C}\right) = \sigma_{P\times C}^2, \quad \text{var}\left(\alpha_j^S\right) = \sigma_S^2,$$

$$\text{var}\left(\alpha_j^{S\times C}\right) = \sigma_{S\times C}^2, \quad \text{var}\left(\alpha_{ij}^{P\times S}\right) = \sigma_{P\times S}^2, \quad \text{var}(\varepsilon_{ijk}) = \sigma_E^2.$$

In this model, β_0 and $\beta_1 c_k$ represent the fixed effects and capture, respectively, the overall mean response and the condition difference in responses. We assume the values of c_k are contrast or deviation coded so that $c_1 + c_2 = 0$ and $c_1^2 = c_2^2 = c^2$. For example, the values of c_k might be $-1, +1$, or $-1/2, +1/2$. Importantly, the values of c_k are assumed not to be dummy or treatment coded (e.g., $c_1 = 0, c_2 = 1$), as this totally alters the meanings of the variance components (see also Barr, Levy, Scheepers, & Tily, 2013, footnote 2), and because we rely on the contrast-coding assumption to deal with potential covariances between the random effects (see the Appendix). The other components in this model are the random effect components and are defined in Table 2. These represent all the potential sources of variation that might be expected in responses.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 2
Definitions of Random Variance/Covariance Components in Mixed Model

Component	Definition	Interpretation
σ^2_P	Variance due to participant (Participant intercept variance)	To what extent do participants have different mean responses?
$\sigma^2_{P\times C}$	Variance due to participant by condition interaction (Participant slope variance)	To what extent does the mean difference between conditions vary across participants?
σ^2_S	Variance due to stimulus (Stimulus intercept variance)	To what extent do stimuli elicit different mean responses?
$\sigma^2_{S\times C}$	Variance due to stimulus by condition interaction (Stimulus slope variance)	To what extent does the mean difference between conditions vary across stimuli?
$\sigma^2_{P\times S}$	Variance due to participant by stimulus interaction (Participant by stimulus intercept variance)	To what extent do participant differences in responses vary with different stimuli?
σ^2_E	Residual error variation	To what extent is there residual variation in responses not due to the above sources?

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

The details of the power calculations are given in the Appendix. For these, we assume balanced designs with no missing data. Additionally we assume that all variance components, including residual error, are normally distributed. We follow there the general procedure specified by Cohen (1988) for simpler designs. This procedure involves first calculating an estimated effect size, analogous to Cohen's *d*: the expected mean difference divided by the expected variation of an individual observation, which in our case accrues from all the variance components specified above:

$$d = \frac{\mu_1 - \mu_2}{\sigma} = \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2_P + c^2\sigma^2_{P\times C} + \sigma^2_S + c^2\sigma^2_{S\times C} + \sigma^2_{P\times S} + \sigma^2_E}}.$$

In this equation, μ_1 and μ_2 are the two expected condition means. The denominator of the effect size is the square root of a pooled variance, albeit a complicated one. It represents the variation within each condition across both participants and stimuli. The Appendix shows how that variance can be determined for different designs. One then calculates an operative effect size (Cohen, 1988, p. 13) estimate that makes adjustments to *d* depending on the specific design used. Next one weights this design-specific effect size by the relevant sample sizes to obtain the noncentrality parameter for a noncentral *t* distribution and also then computes the appropriate degrees of freedom. Cohen (1988, p. 544) at this point relied on an approximation from Dixon and Massey (1957) to estimate power. We have adopted a more exact approach that directly computes areas under the appropriate noncentral *t* distribution.

When we consider power results for specific designs and when we compare results across designs, it is necessary for us to refer to some of the specific results that are in the Appendix. Happily, however, for most purposes the researcher can ignore the technical details that we present in the Appendix because we have implemented all of the steps in a user-friendly Web-based application (http://jakewestfall.org/power/), which performs the rather extensive computations. Figure 1 displays a screenshot of this power application. In addition to computing power estimates, users can also specify a desired power level and solve for the minimum number of participants, minimum number of stimuli, or minimum effect size that would lead to the desired power level. The power application also provides syntax for the R, SAS, and SPSS statistical software packages that one can use to specify the appropriate mixed model for each design and can be used to compute effect sizes from a set of unstandardized design estimates.

Accordingly, all one needs to estimate power from the app are the following:

1. The design to be used.
2. The anticipated effect size or the mean condition difference.
3. Estimates of the relevant variance components.
4. The anticipated numbers of participants and stimuli.

One potential difficulty in doing this is the determination of the estimates of the relevant variance components (Item 3 in the list above). In the Appendix, we develop formulas for the noncentrality parameters and degrees of freedom in terms of these unstandardized variances and also in terms of *variance partitioning coefficients* (VPCs; Goldstein, Browne, & Rasbash, 2002), which we denote using *V* in place of σ^2 (the subscripts remain the same). These in essence are standardized variances, expressing the proportion of the total variation in the observations that is due to a particular variance component. Thus, it is possible to estimate power (and use the application) based on estimates of the relative rather than absolute sizes of the variance components.

There are six possible variance components that must be considered (as laid out in Table 2), and the sum of the six VPCs for these components must equal one. We now move to a more extended discussion of the interpretations of these variance components.

Reasoning About Variance Components and VPCs

Reasoning about statistical power always requires that one think carefully about the factors that influence the variability of responses—that is, about the variance components or VPCs. For mixed models, this is more complicated because there are multiple sources of random variation. Table 2 provides substantive interpretations of all possible variance components in a general, abstract setting. In this section we try to make the meanings of the different VPCs more intuitive and concrete by walking through a hypothetical experimental scenario and describing all possible VPCs in terms of this scenario.

Consider an experiment in which heterosexual male participants consume an alcoholic drink or a nonalcoholic placebo drink and then are asked to judge the attractiveness of a stimulus set of

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

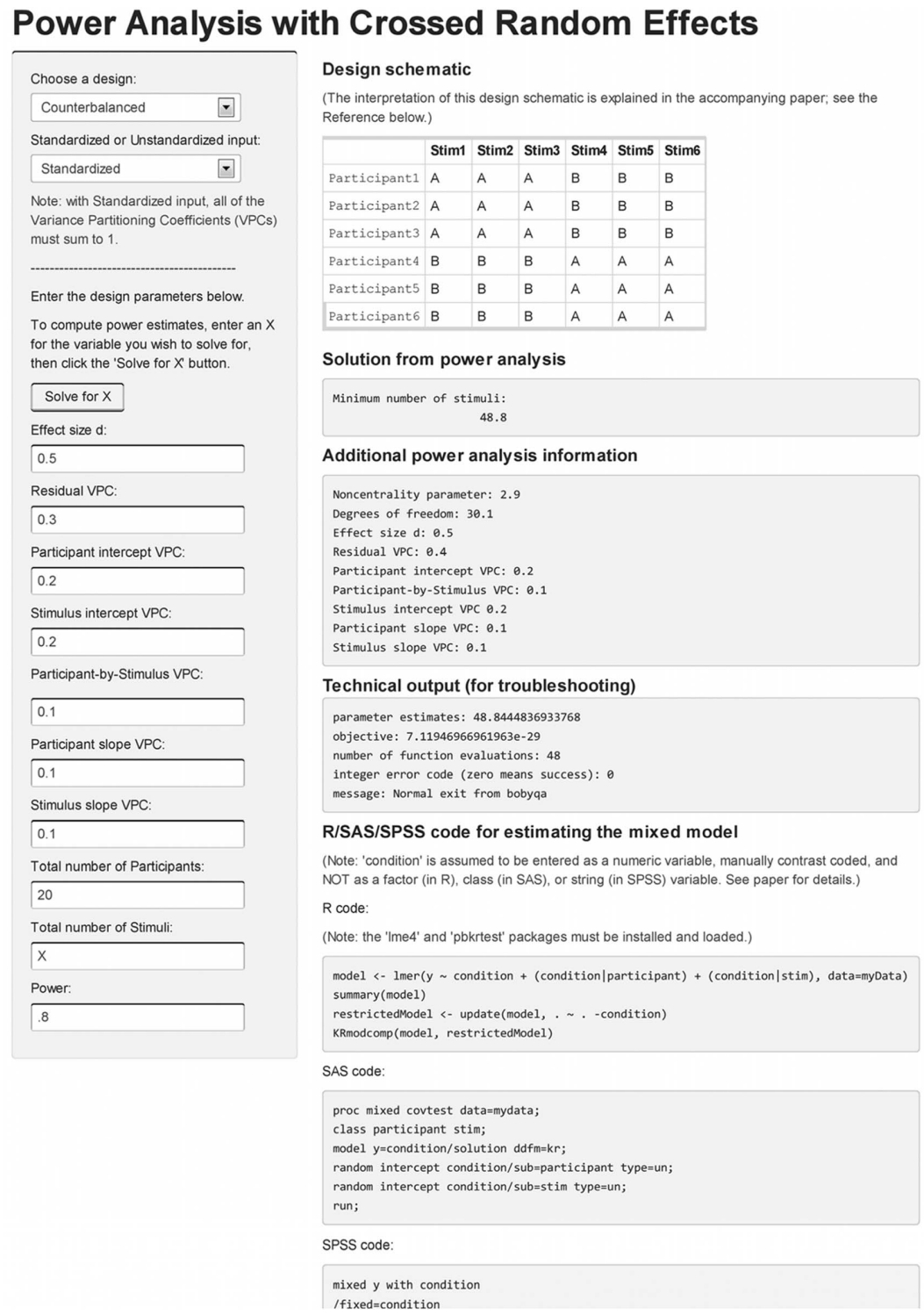

Figure 1. Screenshot from Web-based power application (http://jakewestfall.org/power/). The application allows users to compute statistical power for the five main designs discussed in this article; or to specify a desired power level and solve for the minimum effect size, minimum number of participants, or minimum number of stimuli that would lead to that power level. The application also prints useable syntax in the R, SAS, and SPSS software packages for fitting a linear mixed model in any of the designs and can be used to compute effect sizes and VPC values from a set of unstandardized design estimates. Stim = stimulus.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

photographs of female faces.[1] Thus, the response variable is some continuous rating of perceived attractiveness, the Condition predictor is alcoholic versus placebo drinks, the participants are the "perceivers," and the stimuli are the photographs or "targets" of perception. The predicted "beer goggle effect" is the tendency for attractiveness ratings to be higher on average following consumption of the alcoholic drink compared to the nonalcoholic placebo drink.

In this hypothetical experiment, V_p refers to the variance in the perceivers' average tendencies to view all targets as attractive or unattractive. That is, the perceiver Allen may be a high responder (Allen views most targets as being relatively attractive, compared to other perceivers) while the perceiver Bob may be a low responder (Bob is more picky and views most targets as being relatively unattractive), and V_p refers to the variation across perceivers in their average response tendencies. V_S refers to the variance in the targets' average tendencies to be perceived as attractive or unattractive. That is, the target Carol may tend to elicit high responses (most perceivers agree that Carol is attractive) while the target Diane may tend to elicit low responses (most perceivers agree that Diane is unattractive), and V_S refers to the variation across targets in their average perceived attractiveness. In mixed model terminology, V_p and V_S refer to the variance in the participant intercepts and stimulus intercepts, respectively.

Neither of the variance components described above were concerned with the beer goggle effect; they referred to attractiveness ratings on average across the alcohol and placebo conditions. The $V_{P\times C}$ component refers to the variance of the perceivers' beer goggle effects. That is, the perceiver Allen may exhibit a large beer goggle effect (Allen tends to give higher attractiveness ratings following consumption of the alcoholic drink compared to the placebo drink), whereas the perceiver Bob may exhibit a small or even a negative beer goggle effect (Bob's attractiveness ratings do not tend to be affected by alcohol consumption, or perhaps he even gives lower attractiveness ratings following alcohol consumption), and $V_{P\times C}$ refers to the variance across perceivers in the magnitudes of their beer goggle effects. $V_{S\times C}$ refers to variation in how much different targets tend to benefit from beer goggle effects. That is, the target Carol may benefit greatly from beer goggle effects (Carol tends to be rated as more attractive by perceivers who consumed the alcoholic compared to the placebo drink), whereas the target Diane may not benefit much from beer goggle effects, or may even tend to elicit negative beer goggle effects (Diane's attractiveness is judged to be about equal by perceivers who consumed alcoholic or placebo drinks, or perhaps is even viewed as less attractive following alcohol consumption). In mixed model terminology, $V_{P\times C}$ and $V_{S\times C}$ refer to variance in the participant slopes and stimulus slopes, respectively.

Next we have $V_{P\times S}$, which refers to the variance in the relationship effects for each pairing of perceiver and target (Kenny, 1994). Previously we supposed that the perceiver Allen tended to judge targets as attractive and that the target Carol tended to elicit high attractiveness ratings. On the basis of this, we certainly expect Allen to give Carol high attractiveness ratings. The extent to which Allen's ratings of Carol's attractiveness tend to systematically differ from this expected additive attractiveness rating is the perceiver–target interaction (or relationship effect) for Allen and Carol. If Allen systematically rates Carol as being even more attractive than expected based on their individual perceiver and target effects, then they have a high relationship effect. If Allen systematically rates Carol lower than expected, then they have a low relationship effect. $V_{P\times S}$ refers to the variance across all perceiver–target pairings in these interaction effects.

Finally, V_E refers to the residual or error variance. When the attractiveness ratings of a perceiver toward a target differs from what is expected based on both the intercept and slope for both the participant and stimulus, as well as their interaction effect, then this is considered unexplained variation and attributed to the error variance.[2]

Power Analyses for the Standard Case

In this section, we provide power estimates for each of the previously defined five designs.[3] For these estimates we made some relatively arbitrary but not unreasonable decisions about values of the VPCs and effect sizes that one might encounter. Specifically, we make the following assumptions about the relative magnitude of the various variance components: $V_E = .3$; $V_P = V_S = .2$; $V_{P\times C} = V_{S\times C} = V_{P\times S} = .1$. We refer to these values as the *standard case*. These VPCs reflect first our informal observation, from fitting mixed models to many different data sets in experimental psychology, that variance due to residual error tends often to account for the largest proportion of random variation; variance due to participant and stimulus mean response tendencies (i.e., random intercepts) tends to account for a noticeable but smaller fraction of random variation; variance due to random participant-by-condition and stimulus-by-condition interactions (i.e., random slopes) tends to account for still less of the total random variation; and finally variance due to participant-by-stimuli interactions in many contexts is typically small as well. These observations about the typical relative magnitudes of the variance components are also consistent with regularities that have been frequently remarked upon in the statistical literature on the design of experiments, where this phenomenon has been referred to as the "hierarchical ordering principle" (Li, Sudarsanam, & Frey, 2006; Wu & Hamada, 2000, p. 143) or the "sparsity-of-effects principle" (Montgomery, 2013, p. 290).

Of course, in particular cases, researchers who feel that these default VPC assumptions are unlikely to be a reasonable description of the data from their experiment can use the power application described in the previous section to get power estimates

[1] One could easily imagine an extended version of this design in which both genders serve as both perceivers and targets, but we wished to keep our example more consistent with the simpler single-fixed-factor designs that are the focus of this article, and hence we focus arbitrarily on males perceiving females.

[2] Note that in the designs that we consider in this article, V_E also implicitly contains variance due to the P × S × C interaction, as well as covariance between the P × S and P × S × C terms. However, it is only possible to uniquely estimate these two additional parameters if every participant receives every stimulus multiple times under both conditions—in other words, in the fully crossed design with multiple replicates. As mentioned in the main text, we do not consider the case of multiple replicates, but we mention these potential additions to the full mixed model here for the sake of completeness.

[3] All power results were derived as explained in the Appendix. Additionally, in the case of every design, selected results were empirically confirmed through simulations. The simulation code and results are posted in links at the bottom of the power application Web page (http://jakewestfall.org/power/).

tailored to their research domain. We certainly acknowledge that in many different domains our assumptions about these variance components, specifically that the variances due to participants and stimuli are approximately equal, do not hold.

In introducing the power results for each design in this standard case, we provide the algebraic expressions for the noncentrality parameter for each, pulled from the Appendix. This is done to make clear how the designs differ in their power estimates. Factors that increase the noncentrality parameter (i.e., move it further from 0) lead to greater power, while factors that decrease the noncentrality parameter (i.e., move it closer to 0) lead to lower power.

Fully Crossed Design

In this design, participants and stimuli are crossed with each other and both are crossed with condition. As an example, imagine that participants are asked to judge the attractiveness of a set of faces; they do this under two different context conditions, with the same faces in each condition. Because participants make two responses to each stimulus in this design, there is a possibility of observing order or carryover effects in participants' two responses to each stimulus; for the sake of simplicity, we assume here that no such effects are operative.

In this design, the fully specified mixed model can be estimated, with all the variance components defined as in that equation. In other words, there are no components of variance that are confounded given the mixed model specification. From the Appendix the noncentrality parameter for this design is

$$ncp = \frac{d}{2\sqrt{\frac{V_{P\times C}}{p} + \frac{V_{S\times C}}{q} + \frac{V_E}{2pq}}}.$$

Thus, the relevant variance components for calculating power in this design are the participant slope variance, the stimulus slope variance, and the residual error variance. Variations due to participant and stimulus intercepts, while estimable, do not contribute to the standard error of the treatment effect.

On the basis of the values of the VPCs given earlier, in Figure 2 we plot power results for this design, as a function of p (the number of participants), q (the number of stimuli), and small, medium, and large values of the standardized effect size ($d = 0.2$, 0.5, and 0.8, respectively). This figure contains two ways of looking at the power results. The graphs in the top row plot power (as a function of p and q) for the three effect sizes. Those in the bottom row plot minimum effect sizes needed to have power equal at least .5, .8, and .95, respectively.

These results show that power is dramatically affected by both the number of participants and the number of stimuli in this design, and, given the parallel magnitude of the relevant variance components, the power curves are perfectly symmetric as a function of the two sample sizes. Although most researchers are reasonably attuned to thinking about the need to gather data from a sufficiently large sample of participants to achieve acceptable power levels, it is rare for researchers to think in a parallel manner about the appropriate size of the stimulus samples they use. What is remarkable about our results here is that, given the assumptions we are making about the variance components, maximum achievable power with a medium effect size when using eight stimuli—a fairly typical value of q in many experimental studies—is only approximately .50, *even with an infinite number of participants*. Another way of saying this is that if one anticipates a medium effect size and one would like power to roughly equal .80, then the minimum number of stimuli that can be used, even with a very large number of participants, is about 16. We discuss the idea of maximum attainable power in more detail in the next major section on principles of optimal design.

Counterbalanced Design

In this design, each participant responds to every stimulus, one half of which are responded to in one condition and the other half in the other, and those halves are counterbalanced across participants. Again imagine that the attractiveness of a set of faces is judged, half in one condition and half in the second, but which set is judged in which condition varies between participants. Thus, in this design, each participant would see a given face only once. Technically, when analyzing data from the counterbalanced design, a second fixed predictor representing the counterbalancing factor (e.g., Set A vs. Set B for faces appearing in one of two sets) could be added to the model, and this would be sensible to do if, for example, the stimuli were assigned at all nonrandomly to the levels of the counterbalancing factor. Here, for the sake of simplicity, we assume there is no main effect of the counterbalancing factor and omit this second predictor from the model.

In this design, only five of the variance components are estimable. The variance due to the participant-by-stimulus interaction ($V_{P\times S}$), is not estimable and is confounded with V_E. From the Appendix the noncentrality parameter for this design is

$$ncp = \frac{d}{2\sqrt{\frac{V_{P\times C}}{p} + \frac{V_{S\times C}}{q} + \frac{[V_E + V_{P\times S}]}{pq}}}.$$

Here the confounding of the participant-by-stimulus interaction variance and the error variance is indicated by the brackets enclosing these two VPCs in the denominator. This noncentrality parameter differs from that of the fully crossed design, given above, as a function of this confounding. Additionally, in this design, compared to the fully crossed design, there is half the number of total observations here, given constant p and q. As in the fully crossed design, variation due to participant and stimulus intercepts does not affect power.

Because of the differences just noted, the power results for this design, given in Figure 3, are a bit below those given for the fully crossed design. Actually, for the values of the variance components that we are considering, the power difference is remarkably small. Now with eight stimuli and a medium effect size, the maximum achievable power, even with an extremely large number of participants, is still approximately .5. Put another way, to achieve power of .8 with a medium effect size and an extremely large number of participants, one would need at least 16 stimuli. At a later point, we discuss in greater detail the relative efficiency of this design compared to the fully crossed design.

Stimuli-Within-Condition Design

This is the design that Judd et al. (2012) considered in some detail, in which stimuli are in one condition or the other, but each participant responds to each stimulus and thus participant is crossed with condi-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

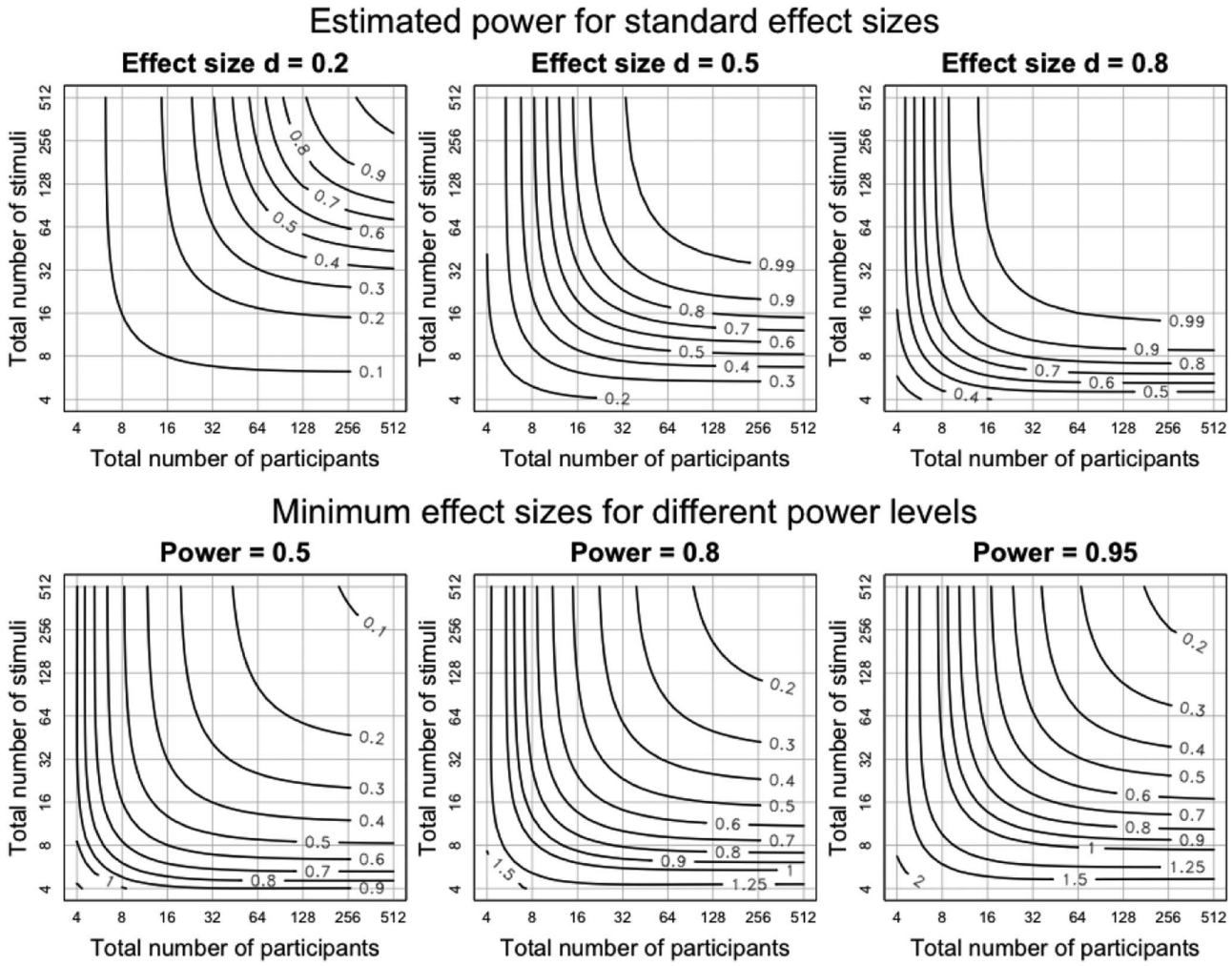

Figure 2. Contour plots for the fully crossed design. Top panel: Statistical power as a function of the effect size, the number of participants, and the number of stimuli. Bottom panel: Minimum effect sizes for different desired power levels as a function of the number of participants and number of stimuli. The VPCs are held constant at $V_p = V_S = .2$, $V_{P\times C} = V_{S\times C} = V_{P\times S} = .1$, and $V_E = .3$. The sample sizes are on log scales. VPCs = variance partitioning coefficients.

tion.[4] In terms of the beer goggle example given earlier, each participant is in both the alcohol and placebo conditions but different target persons are judged for attractiveness in the two conditions.

In this design, only four of the variance components are estimable. The variance due to the stimulus-by-condition interaction or $V_{S\times C}$, is not estimable because stimuli are not crossed with condition. Instead this variance is confounded with the stimulus mean variance V_S. And as in the counterbalanced design, the participant-by-stimulus interaction or $V_{P\times S}$ is confounded with the residual error variance.

From the Appendix, the noncentrality parameter for this design is

$$ncp = \frac{d}{2\sqrt{\frac{V_{P\times C}}{p} + \frac{[V_S + V_{S\times C}]}{q} + \frac{[V_E + V_{P\times S}]}{pq}}}.$$

Again, the confounding of variance components is indicated by the brackets around the VPCs in the denominator. The power of this design is less than the power of the first two designs considered largely as a function of the variation due to stimulus intercepts or means which now figures in the denominator of the noncentrality parameter or equivalently in the standard error for testing the conditions difference.

The power results for this design are given in Figure 4. Unlike the earlier two designs, the power results here are no longer symmetric with respect to the numbers of participants and stimuli. Under the assumptions that we have made, power in this design is more influenced by q, the number of stimuli, than by p, the number of participants. Here, with a moderate effect size and a very large number of participants, one achieves power of .80 only by using

[4] We caution the reader that in our treatment of this design in Judd et al. (2012), q was defined as the number of stimuli in each condition, not as the total number of stimuli, as it is here.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

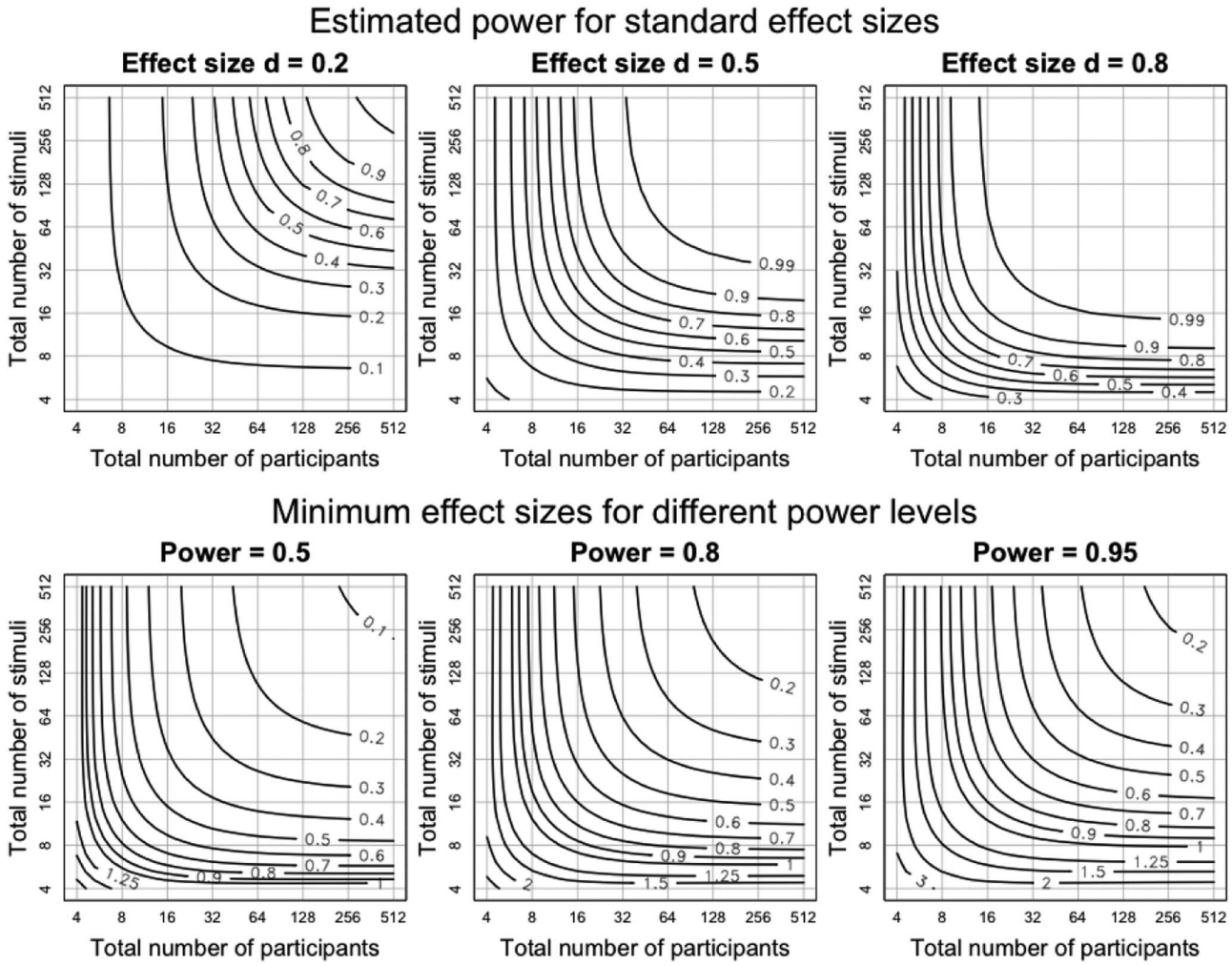

Figure 3. Contour plots for the counterbalanced design. Top panel: Statistical power as a function of the effect size, the number of participants, and the number of stimuli. Bottom panel: Minimum effect sizes for different desired power levels as a function of the number of participants and number of stimuli. The VPCs are held constant at $V_p = V_S = .2$, $V_{P\times C} = V_{S\times C} = V_{P\times S} = .1$, and $V_E = .3$. The sample sizes are on log scales. VPCs = variance partitioning coefficients.

more than 32 stimuli per condition. Another way of saying this is that power increases as a function of q begin to asymptote only at values of q greater than 64. On the other hand, power increases as a function of p begin to asymptote at values of p greater than 32.

Participants-Within-Condition Design

Statistically this is the same design as the previous one, except now the nesting or crossing of the two random factors with condition is reversed: Participants are now in only one condition or the other, but all stimuli are responded to in both. In terms of the beer goggle experiment, some participants receive only the alcoholic drink while other participants receive only the nonalcoholic placebo drink, but all participants respond to the same set of stimulus photographs.

In this design the variance attributable to the participant-by-condition interaction, $V_{P\times C}$, is not estimable. Instead, it is confounded with the participant mean variance, V_P. And as in the previous two designs, the participant-by-stimulus interaction or $V_{P\times S}$ is confounded with the residual error variance. The noncentrality parameter for this design, taken from the Appendix, is

$$ncp = \frac{d}{2\sqrt{\frac{\left[V_P + V_{P\times C}\right]}{p} + \frac{V_{S\times C}}{q} + \frac{\left[V_E + V_{P\times S}\right]}{pq}}},$$

with confounding again indicated by the brackets in the denominator. This time, to the extent there is large variation from participant to participant in their means, power would be reduced. In other words, in the previous design, variation in stimulus means increases the standard error in testing the condition difference; now what increases that standard error is variation in participant means.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

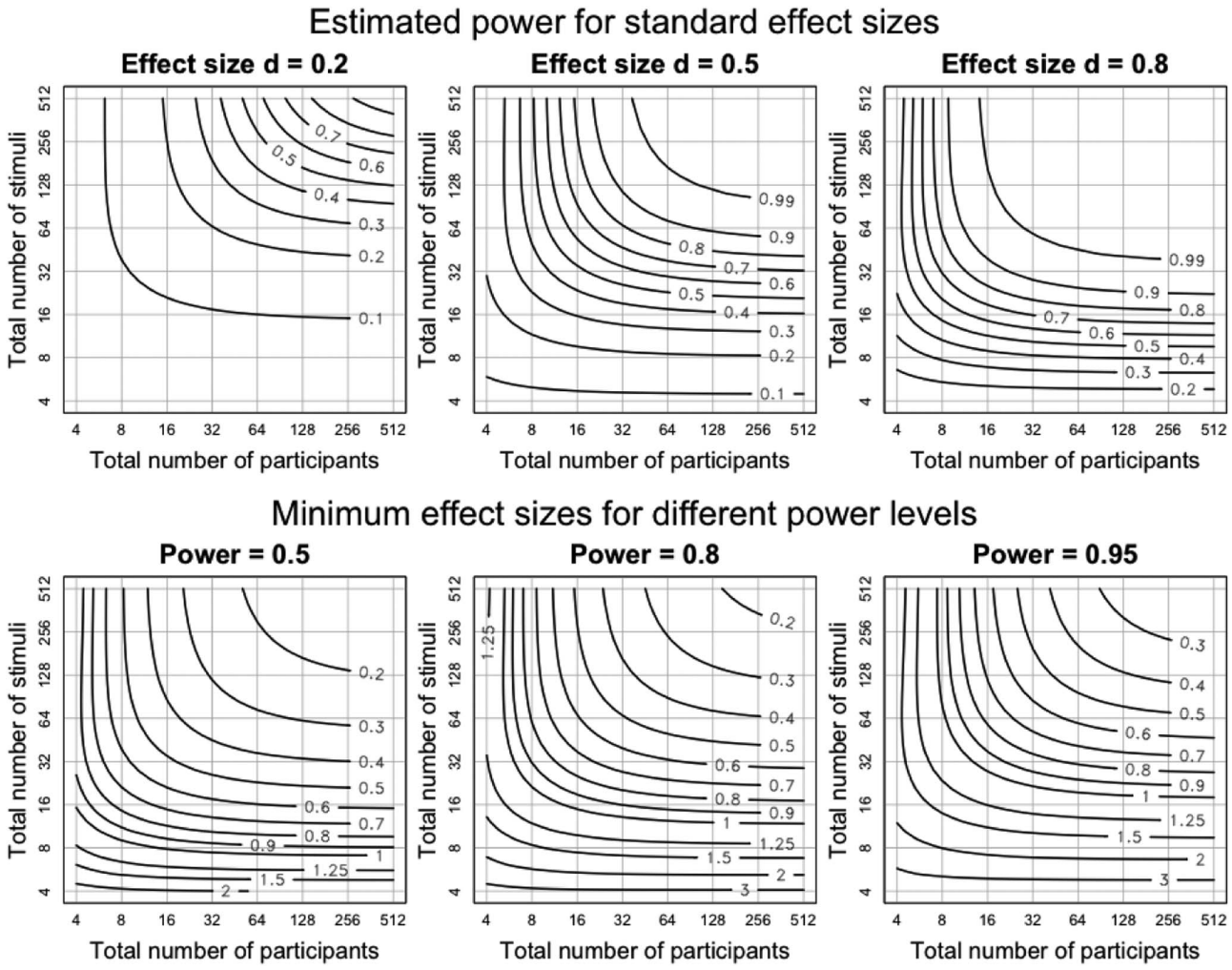

Figure 4. Contour plots for the stimuli-within-condition design. Top panel: Statistical power as a function of the effect size, the number of participants, and the number of stimuli. Bottom panel: Minimum effect sizes for different desired power levels as a function of the number of participants and number of stimuli. The VPCs are held constant at $V_p = V_S = .2$, $V_{P\times C} = V_{S\times C} = V_{P\times S} = .1$, and $V_E = .3$. The sample sizes are on log scales. VPCs = variance partitioning coefficients.

Under the assumptions that we have made, the power results for this design, given in Figure 5, are the same as those given for the previous design except they have been transposed so that now power is more dramatically affected by the number of participants than the number of stimuli. Even in the case of this design, however, it remains true that power of .80 is only achievable, given a moderate effect size, when the number of stimuli is greater than 16, a number which is larger than that typically used.

Both-Within-Condition Design

The final design has each participant and also each stimulus in only one of the two conditions, but within each condition all $p/2$ participants judge all $q/2$ stimuli. From the beer goggle example, each participant consumes either alcohol or the placebo, and each target is judged for attractiveness in only one of the two conditions.

In this design three different variance components are not estimable and are confounded with other components: the stimulus-by-condition component, $V_{S\times C}$, the participant-by-condition component, $V_{P\times C}$, and the participant-by-stimulus interaction component, $V_{P\times S}$. The first of these is confounded with the stimulus intercept variance, the second with the participant intercept variance, and the third with the residual error variance. From the Appendix, the noncentrality parameter for this design is

$$ncp = \frac{d}{2\sqrt{\frac{\left[V_P + V_{P\times C}\right]}{p} + \frac{\left[V_S + V_{S\times C}\right]}{q} + \frac{2\left[V_E + V_{P\times S}\right]}{pq}}}.$$

As a result, the standard error for testing the condition difference is inflated and power reduced to the extent that there exist both large participant and large stimulus differences in their means.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

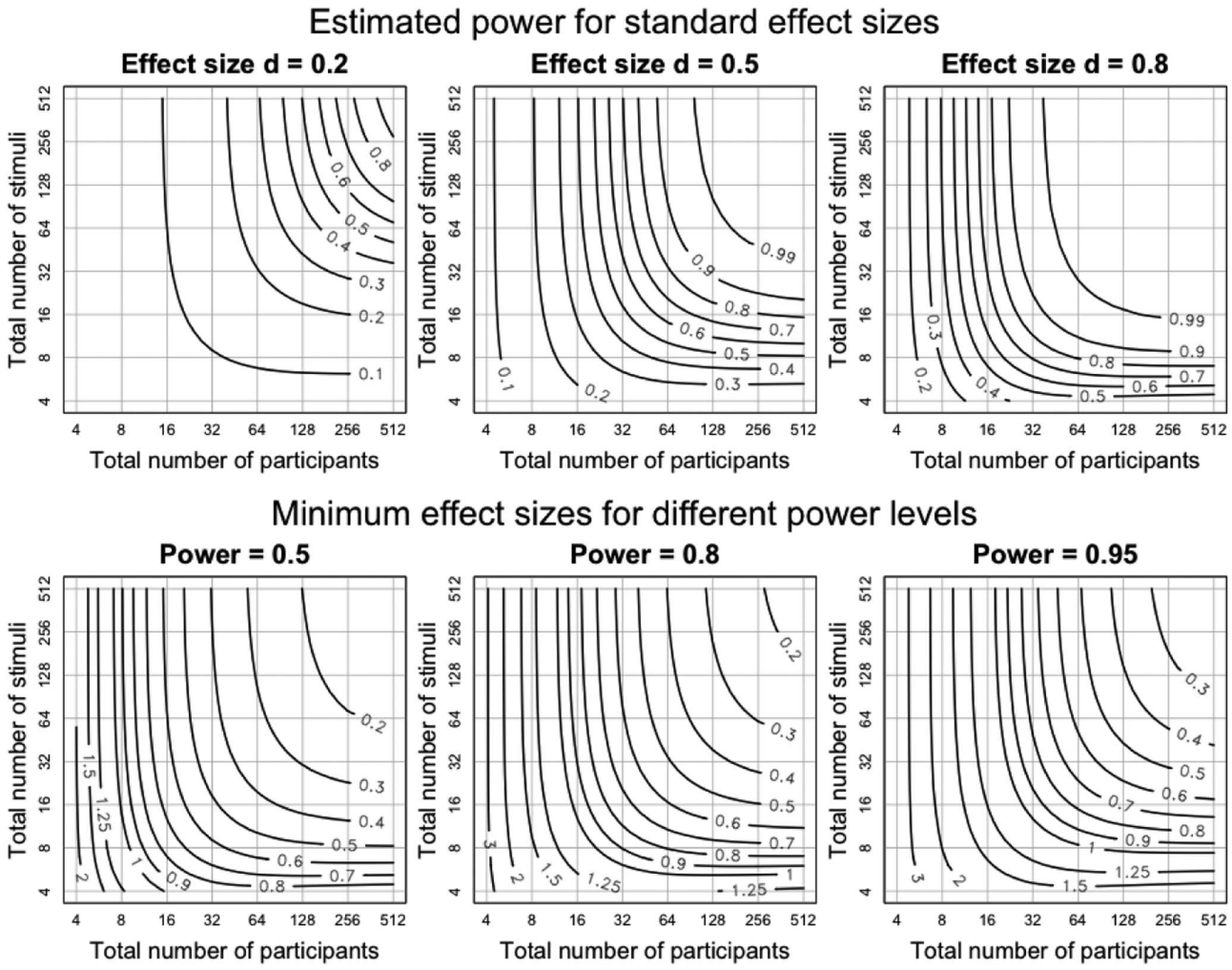

Figure 5. Contour plots for the participants-within-condition design. Top panel: Statistical power as a function of the effect size, the number of participants, and the number of stimuli. Bottom panel: Minimum effect sizes for different desired power levels as a function of the number of participants and number of stimuli. The VPCs are held constant at $V_p = V_S = .2$, $V_{P\times C} = V_{S\times C} = V_{P\times S} = .1$, and $V_E = .3$. The sample sizes are on log scales. VPCs = variance partitioning coefficients.

The power results given in Figure 6 for this design are again symmetric with respect to p and q under the assumptions we are making. And these power results are particularly poor, relative to the other designs we have considered. With a moderate effect size and as many as 25 participants responding to 25 stimuli in each condition, power only equals .5. Even with unlimited numbers of participants, power of .80 with a moderate effect size is achievable only if the number of stimuli per condition is greater than roughly 48.

Illustrative Example

To make our discussion of power analysis for these complex designs more concrete, here we work through an example for a hypothetical experiment. This example illustrates the process of coming up with some reasonable variance component and effect size estimates given what we know about our study, using the online power application to compute power and minimum numbers of participants and/or stimuli, and investigating how these answers would change under different assumptions about the variance components and effect size.

Consider a study where we are investigating the impact of a cognitive load manipulation (e.g., memorizing short lists of integers) on participants' performance on a series of items from a task that depends on working memory capacity. The dependent variable is some continuous measure of each participant's performance on each item. We would like to employ a within-subject design where each participant responds to items both under cognitive load and not under cognitive load, but we wish to avoid any potential carryover effects that might result from each participant responding to each item twice (once under each load condition). Therefore,

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

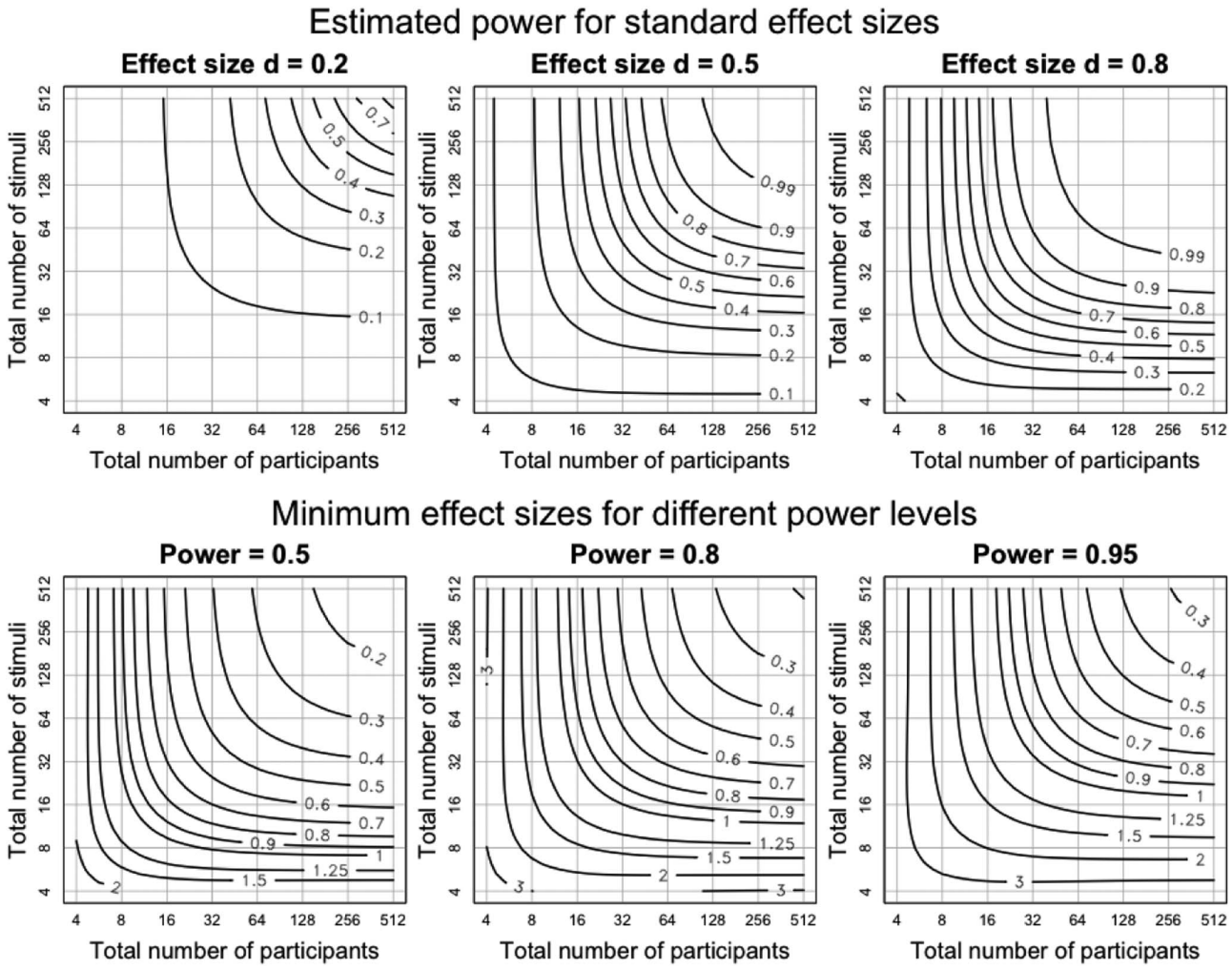

Figure 6. Contour plots for the both-within-condition design. Top panel: Statistical power as a function of the effect size, the number of participants, and the number of stimuli. Bottom panel: Minimum effect sizes for different desired power levels as a function of the number of participants and number of stimuli. The VPCs are held constant at $V_p = V_S = .2$, $V_{P\times C} = V_{S\times C} = V_{P\times S} = .1$, and $V_E = .3$. The sample sizes are on log scales. VPCs = variance partitioning coefficients.

the items are divided into two lists, List A and List B, and participants are randomly assigned to either respond to the List A items under cognitive load and the List B items not under load, or to respond to the List B items under cognitive load and the List A items not under load. Thus, we employ the counterbalanced design. We have two random factors, Participant and Item, and one fixed factor representing the overall difference in performance between the load and no-load conditions.

Suppose that we have developed a sample of 16 items (eight on each list) for this working memory task. How many participants must we recruit to achieve statistical power of 0.8? We first examine some answers to this question using the default set of VPCs that we proposed above ($V_E = .3$;$V_P = V_S = .2$; $V_{P\times C} = V_{S\times C} = V_{P\times S} = .1$), and a medium effect size of $d = 0.5$. The online power application (located at http://jakewestfall.org/power/) has these VPCs as the default. We specify the design as the counterbalanced design. According to some recent recommendations (Simmons, Nelson, & Simonsohn, 2011), a study should employ at least 20 participants per between-subjects condition, so we start with that value as the number of participants and enter 16 for the number of stimuli (items). Pressing the *Solve for X* button, we find that under the VPC assumptions above, power would be .571. If we want to know how many participants would be required to achieve statistical power of .80, we can enter the *x* symbol for number of participants, fill in .80 for power, and press *Solve for X*; the application informs us that we require 154 participants.

An examination of the power plots for this design (in Figure 3) makes clear that with only 16 items, the maximum power that can be achieved, under the assumptions made, even with a very large number of participants, is only slightly above .8. These power results also make clear that if more power is to be achieved, more items are necessary. Accordingly, as a result of reading what has been presented so far in this article, it is decided to increase the number of items to 30. Solving now for the number of participants

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

to achieve power of .80, the application tells us that 27 participants are needed. Clearly, increasing the number of stimuli makes a large power difference.

In the absence of any information about the study we are planning other than the possible sample sizes, the default set of VPCs that we used above and a medium effect size together represent a reasonable set of assumptions that we might expect to be approximately true for our study. But if we do know a little about the samples of participants and stimuli we are using and the effect which we are studying, we can do better by tailoring what we expect the VPCs to be for the research being conducted. For example, suppose that we know that the items we developed for the working memory task vary considerably in their average difficulty (some items tend to yield low performance scores, while others tend to yield high performances scores); as a consequence, the more difficult ones might be affected more by the cognitive load manipulation than the easier items. We also might know that our participants are likely to be rather homogeneous in their average working memory capacities and, we suspect, also in the degree to which cognitive load would interfere in task performance. To reflect this knowledge, it seems reasonable to increase the values of both V_S and $V_{S\times C}$ and to decrease the values of both V_P and $V_{P\times C}$. Accordingly, we decide to increase V_S and $V_{S\times C}$ to .25 and .15, respectively, and to decrease both V_P and $V_{P\times C}$ by the same amounts (to .15 and .05, respectively). We leave the other VPCs at their default values. These changes, tailored to our knowledge of the research domain, reveal that power of .80 is achievable with 25 participants, again assuming that we have increased the number of items to 30.

In the process illustrated above, we started with some default values of the variance components and effect size, and then we made adjustments to some of these default values based on substantive knowledge that we had about the details of the particular study we would be running. However, we note that all of these values, even the values we specifically adjusted, are ultimately educated guesses about the true parameter values, and so they should be seen as implicitly containing a degree of uncertainty. In acknowledgment of this, our recommendation is that researchers use these educated guesses as a starting point in computing power and sample size estimates for a range of plausible values of the parameters for the study at hand.

Optimal Design With Crossed Random Factors

On the basis of the above power analyses for our five designs, we can draw some general conclusions and formulate a variety of rules of thumb concerning the optimal design of experiments where a sample of participants respond to a sample of stimuli. In this section, we discuss the maximum attainable power in an experiment once the number of stimuli is fixed; when statistical power is best served by increasing the number of participants or by increasing the number of stimuli; the relative efficiency of the five designs; and finally the statistical merits of designs involving multiple blocks of stimuli as a way of dealing with the problem of time-consuming stimulus presentation.

Maximum Attainable Power

One potentially surprising fact that we learn from our power analyses about the use of designs with crossed random factors is that statistical power does not approach unity as the number of participants approaches infinity, whenever there is variation due to stimuli. Instead, statistical power asymptotes at a maximum theoretically attainable power value that depends on the effect size, the number of stimuli, and the variability of the stimuli. To see this in the fully crossed design, consider what happens to the noncentrality parameter (ncp_{FC}) as the number of participants (p) goes to infinity:

$$\lim_{p\to\infty} ncp_{FC} = \lim_{p\to\infty} \frac{d}{2\sqrt{\frac{V_{P\times C}}{p} + \frac{V_{S\times C}}{q} + \frac{V_E}{2pq}}} = \frac{d\sqrt{q}}{2\sqrt{V_{S\times C}}}.$$

Notice here that when p goes to infinity, the terms in the denominator involving V_E and $V_{P\times C}$ disappear, but the numerator and the term involving $V_{S\times C}$ are both unchanged, so that in the limit the noncentrality parameter converges to a finite (and possibly small) value. In Figure 7 we plot the maximum attainable power in the fully crossed design as a function of the effect size (d), the number of stimuli (q) and the variance in the stimulus slopes ($V_{S\times C}$). These are contour plots just like the plots displayed earlier; the difference is that instead of plotting the observed power for different combinations of parameters, they plot the upper bound for power at various combinations of parameters. The plots of maximum power for the other four designs all look nearly identical to those for the fully crossed design.

The primary lesson for experimenters following from the statistical fact that maximum attainable power does not approach unity with increasing numbers of participants is that it is very important to think carefully about the sample of stimuli in the experiment before the data collection begins. Once data collection begins, the effect size and the number and properties of the stimuli can no longer be changed, and thus the maximum power that would be attainable in the experiment has in essence already been decided. All that then remains is to recruit a certain number of participants, which determines how close to this maximum power level the actual power ends up being. Experimenters may believe that they can compensate for a suboptimal sample of stimuli by simply recruiting a larger number of participants, but in fact the degree to which this sort of compensation can take place is quite limited, a point which we discuss in more detail in the next section. The fact is that in many entirely realistic experimental situations, the maximum attainable power with an infinite number of participants can be quite low even for detecting true, large effects. For example, in an experiment employing the stimuli-within-condition design, under the standard case VPCs described above, where the true effect size is large at $d = 0.8$, and where there are a total of eight stimuli (four stimuli per condition)—a sample size which we suspect many experimenters would consider perfectly adequate for a stimulus sample—the maximum attainable power is only about .41. However, if we just double the sample size of stimuli to a still relatively modest 16 (eight per condition), then the maximum power to detect a large effect goes up to about .78.

Another implication of the maximum attainable power being less than one is that in studies involving crossed random factors, a direct replication with high statistical power is often theoretically impossible when the original study employed a relatively small number of stimuli. Recently, researchers have stressed the impor-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

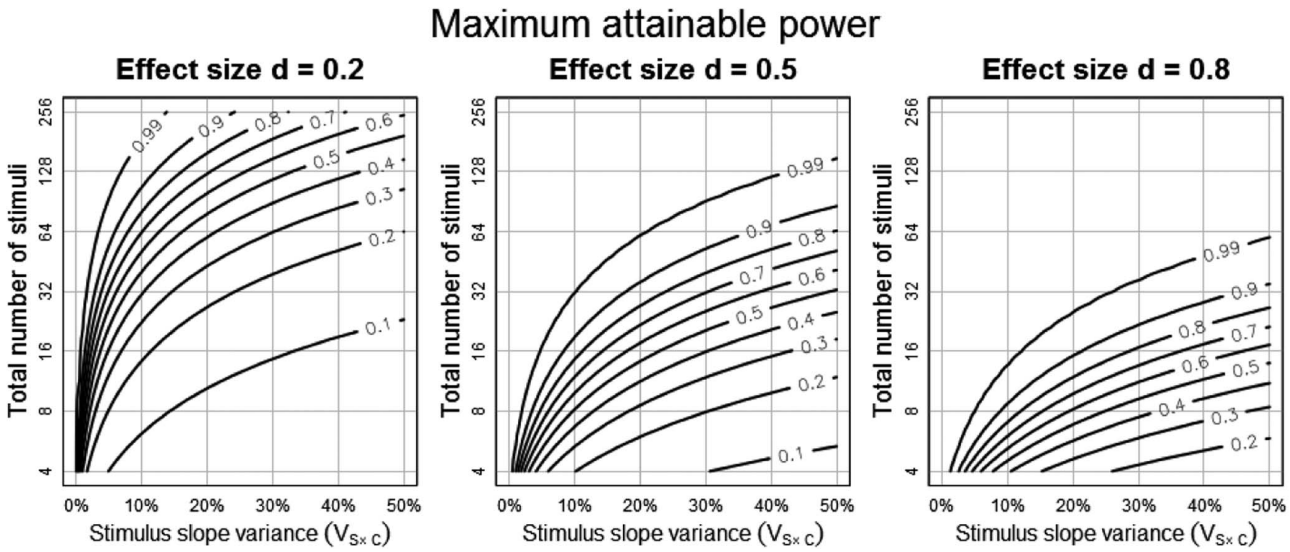

Figure 7. Contour plots of the maximum attainable power with an infinite number of participants in the fully crossed design as a function of the effect size, number of stimuli (on a log scale), and the proportion of stimulus slope variance. The contour plots of maximum power for the other four designs all look nearly identical to the plots shown here.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

tance of conducting direct replications (Francis, 2012; Ioannidis, 2012; Koole & Lakens, 2012; Nosek, Spies, & Motyl, 2012; Open Science Collaboration, 2012) and also emphasized that replication attempts should ideally have high statistical power, even when (perhaps especially when) the original study was underpowered (Brandt et al., 2014). Most typically, those who would replicate a study with higher power than the original might employ the same stimuli but increase the number of participants to a level they consider adequate. However, to the extent that there is substantial stimulus variability, replications with high power may in fact not be feasible using only the stimuli included in the original study. Sufficiently high power may be obtainable only by increasing the number of stimuli, as well as the numbers of participants.

We point out that a direct replication would virtually never use the same participants as the original study. Instead, it would probably just be assumed that the new participants were drawn from the same general population as the original participants. Why then should we require the stimuli to be exactly the same as in the original experiment, rather than just being drawn from the same population? We strongly suggest that, when conducting a replication of a study involving crossed random factors, it would be beneficial for the researchers conducting the replication to augment the stimulus set, drawing from the same population of stimuli as in the original study, in order to ensure that statistical power is acceptably high. We discuss these issues in more detail in a companion article (Westfall, Judd, & Kenny, 2014).

Increasing the Sample Sizes of Participants Versus Stimuli

Not surprisingly, experiments with larger sample sizes of both participants and stimuli always have greater power than experiments with smaller sample sizes. However, the question of whether one can expect a greater benefit to statistical power by increasing the sample size of participants or the sample size of stimuli turns out to depend on several different aspects of the experiment. Here we formulate two rules of thumb that researchers can use to identify situations where it would be better to increase the sample size of participants or of stimuli. Formally, these rules of thumb are based on an analysis of the conditions under which the rate of change in the noncentrality parameter with respect to the number of participants is greater than the rate of change in the noncentrality parameter with respect to the number of stimuli (or vice versa). For example, to examine when it is better to increase the sample size of stimuli rather than the sample size of participants in the fully crossed design, we first set

$$\frac{\partial ncp_{FC}}{\partial q} > \frac{\partial ncp_{FC}}{\partial p}.$$

With some work, we can simplify this inequality to obtain

$$pV_E + 2p^2V_{S\times C} > qV_E + 2q^2V_{P\times C}. \quad (1)$$

We refer to this inequality in discussing the rules of thumb below.

The first rule of thumb is that it is generally better to increase the sample size of whichever random factor is contributing more random variation to the data, considering the nature of the response variable and the relevant properties of the participants and stimuli. Statistically speaking, this is because the primary statistical consequence of adding participants or adding stimuli is that the corresponding participant or stimulus variance components, respectively, would become further diminished in their contribution to the denominator of the noncentrality parameter, and there is a greater advantage to diminishing the contribution of larger variance components compared to smaller variance components. An equivalent way of stating this is the following:

Different variance components contribute to the size of the noncentrality parameter in different designs, and the extent to which they contribute depends on the corresponding sample size on which they are based, with larger sample sizes leading to relatively less contribution. In terms of the example above concerning the noncentrality parameter for the fully crossed design, we can see that if the two sample sizes are equal (i.e., $p = q$) then Inequality 1 just reduces to $V_{S\times C} > V_{P\times C}$

As an example of applying this rule of thumb, consider a reaction time experiment where participants decide as quickly as possible whether each of a list of written stimuli is a word or a nonword. There is probably a great deal of variability between participants in both their mean reaction times and their differential reaction times to words versus nonwords. However, assuming the stimuli are strictly controlled in terms of their number of letters, ease of pronunciation, and so on, there is probably comparatively little variability in the mean reaction times elicited by each stimulus. In this situation there would tend to be a greater power benefit to adding participants compared to adding stimuli. Alternatively, consider an experiment where participants judge the attractiveness of a set of photographs of faces. It is probably the case that some stimulus faces tend to elicit higher or lower attractiveness ratings on average than do other stimulus faces, but there is probably not as much individual difference variance in how attractive a participant considers stimulus faces on average (Hönekopp, 2006). In this situation there would tend to be a greater benefit of adding more stimuli compared to adding more participants.

The second rule of thumb is that if one of the two sample sizes is considerably smaller than the other, there is generally a greater power benefit in increasing the smaller sample size compared to the larger sample size—for two reasons. First and most obviously, the impact of either sample size has a diminishing-returns relationship with statistical power, so that if one sample size is already large and the other is small, increases in the smaller sample size would have a greater impact on power. Second and more subtly, the two sample sizes together have a multiplicative effect on statistical power, so that the rate at which increasing one of the sample sizes leads to an increase in power depends in part on the magnitude of the other sample size. To make this multiplicative relationship clear in the case of the fully crossed design, consider again Inequality 1. When the number of participants is large relative to the number of stimuli this inequality is more likely to be true (i.e., the left-hand side is more likely to exceed the right-hand side), implying that there is greater incremental benefit to adding stimuli compared to adding participants. Likewise, when the number of participants is small relative to the number of stimuli, this inequality is more likely to be false, implying that there is greater incremental benefit of adding participants. More generally, this inequality shows that the relative rates of change of the noncentrality parameter with respect to either of the sample sizes depend in part on the other sample size.

A somewhat insidious consequence of this multiplicative sample size relationship is that the degree of trade-off that one can achieve with the sample sizes—that is, the degree to which a researcher can compensate for having a small number of levels for one random factor by adding many more levels of the other random factor—tends to be quite limited. Generally speaking, if one of the sample sizes of the study is badly deficient, there is little that one can realistically do with the size of the other sample to rescue statistical power, both because of the multiplicative nature of the sample size relationship and because of the low maximum attainable power imposed by the small sample size. Unfortunately, we note that in many areas of the psychological literature, experiments with very small samples of stimuli (e.g., fewer than 10) are quite common. Our power results indicate that such studies often are badly underpowered even for detecting true large effects.

A final point to consider is the relative costs of increasing the sample sizes of the two random factors, relative to the power benefits accruing from those increases. In many circumstances, it may be relatively more cost effective to increase the sample size of stimuli than the sample size of participants. Once participants are recruited for a study, increasing the time they spend in the lab by exposing them to additional stimuli is quite easy and therefore is likely more efficient from a logistical perspective than recruiting additional participants. Again, because researchers have typically not treated stimuli as a random factor in analyses, the power benefits, perhaps accruing at relatively little cost, from using larger samples of stimuli have mostly been ignored. If the goal is to produce replicable results that generalize across both participant and stimulus samples, then researchers clearly need to focus on both random factors in their power considerations, taking into account the power benefits of increasing both sample sizes relative to the costs of doing so.

A Comparison of Crossed Designs

If researchers feel they have a good understanding of what kind of patterns of participant and stimulus variation to expect in a study they are planning, and they also have some freedom to select among different design variations, the results of our power analyses can be used to help optimally select a design for studying the phenomenon of interest.

For a given total number of participants (p) and total number of stimuli (q), the fully crossed design is the most efficient whenever it is possible to run it. This is due to the facts that it provides the greatest number of observations, and that it eliminates the participant intercept, the stimulus intercept, and the participant by stimulus interaction variance terms (V_P, V_S, and $V_{P\times S}$) from the denominator of the noncentrality parameter. In practice the two intercept variances, due to participants and due to stimuli, are often relatively large.

The counterbalanced design, although not quite as powerful as the fully crossed design for a given p and q, also tends to be relatively powerful. Although it yields half as many total observations as the fully crossed design for a given total number of participants and total number of stimuli, it shares the advantage of eliminating the participant intercept and stimulus intercept variance terms (V_S and V_P) from the denominator of the noncentrality parameter. It does not, however, eliminate the participant by stimulus interaction component ($V_{P\times S}$), which is confounded with error in this design.

Note that the fully crossed design requires participants to make twice as many responses for the same number of stimuli as the counterbalanced design. For a fairer comparison of the two designs, we should allow the counterbalanced design to have twice the number of stimuli as the fully crossed design. If an experimenter only has enough time in an experimental session to elicit,

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

say, 100 responses from each participant, then the counterbalanced design would allow the employment of a sample of 100 stimuli, while the fully crossed design would only support a sample of 50 stimuli. It therefore seems that a more reasonable way to compare the two designs might be not to hold constant the number of stimuli but to hold constant the number of responses made by each participant. How do the fully crossed and counterbalanced designs stack up when compared in this way?

To compare the efficiency of the counterbalanced design and the fully crossed design with the number of responses per participant held constant, we first set the noncentrality parameter for the counterbalanced design (ncp_{CB}) greater than the noncentrality parameter for the fully crossed design (ncp_{FC}):

$$ncp_{CB} > ncp_{FC}.$$

Now we substitute in the definitions of these centrality parameters, but set the number of stimuli to $2q$ in the counterbalanced design and to q in the fully crossed design, which equates the two designs in the total number of responses made per participant:

$$\frac{d}{2\sqrt{\frac{V_{P\times C}}{p}+\frac{V_{S\times C}}{2q}+\frac{\left[V_E+V_{P\times S}\right]}{2pq}}} > \frac{d}{2\sqrt{\frac{V_{P\times C}}{p}+\frac{V_{S\times C}}{q}+\frac{V_E}{2pq}}}.$$

Finally, we simplify this inequality to obtain

$$p > \frac{V_{P\times S}}{V_{S\times C}}.$$

In words, the counterbalanced design has greater power than the fully crossed design when the ratio of participant-by-stimulus interaction variance ($V_{P\times S}$) to stimulus slope variance ($V_{S\times C}$) is less than the number of participants (p). We note that this condition is almost always true in any realistic data set. For example, if an experiment involves 50 participants, then we would need to have more than 50 times as much participant-by-stimulus variance as stimulus slope variance for this condition to fail. While there may be many contexts in which we do expect $V_{P\times S}$ to exceed $V_{S\times C}$, very rarely would we expect it to do so by a factor of more than 50. Thus, if the number of responses per participant is held constant rather than the number of stimuli, then the counterbalanced design is actually more efficient in most cases than the fully crossed design. The reason for this advantage is that the counterbalanced design can employ twice as many stimuli using the same number of total responses from each participant, leading to a doubling of one of the sample sizes of the study. It therefore seems that the choice of whether to use the fully crossed design or the counterbalanced design comes to down to whether the main constraint in the size of the experiment comes from the maximum number of responses that one can collect from each participant, or from the number of stimuli available to respond to. If one has only a small number of stimuli but is able to elicit more responses from each participant, then the fully crossed design is preferable. If one has a large stimulus pool but is limited in the number of responses that can be collected from each participant, then the counterbalanced design is preferable.

Moving now to the stimuli-within-condition and participants-within-condition designs, the relative efficiency of these two designs depends totally on the relative magnitudes of the participant and stimulus variance components and on the two sample sizes. If we set the noncentrality parameter for the Stimuli-within-condition design greater than the noncentrality parameter for the participants-within-condition design, we find that the stimuli-within-condition design is more powerful than the participants-within-condition design when

$$\frac{V_P}{p} > \frac{V_S}{q}.$$

In other words, if there is a great deal of variance in the mean responses of participants (V_P is high) or if the number of participants (p) is small, then the stimuli-within-condition design tends to be relatively more efficient. But if it is the stimuli that have relatively more variance in their mean responses elicited (V_S is high), or if the number of stimuli (q) is small, then the participants-within-condition design tends to be more efficient, all other things being equal.

Finally, the both-within-condition design generally has the lowest power of the five designs considered here. This design yields the lowest number of total observations for a given total number of participants and total number of stimuli, and the denominator of the noncentrality parameter in this design implicitly contains all possible sources of variation, whereas for the other four designs, one or more of the sources of variance are absent.

Dealing With Time-Consuming Stimulus Presentation: Stimuli-Within-Block Designs and Nested Designs

One of the major practical lessons in this article is that to ensure adequate statistical power, stimulus samples usually need to be at least moderately sized; in any case, larger than is very often seen in many areas of experimental psychology. However, there are many research paradigms in which it is infeasible to have each participant respond to more than a handful of stimuli. For example, if it is determined that a sample size of 60 stimuli is needed for statistical power reasons, but the stimuli are 5-minute film clips or songs that participants must view or listen to in their entirety before making each response, then the counterbalanced design for example would require each participant to spend over 5 hours viewing and responding to stimuli, which is likely impractical due to concerns of participant fatigue, time constraints on the duration of individual experimental sessions, and other reasons.

A solution to this dilemma is to use what we call a stimuli-within-block design (see Table 3). These are conceptually related to matrix sampling designs from survey research, an example of a planned missing data design (Graham, Taylor, Olchowski, & Cumsille, 2006; Shoemaker, 1973). In a stimuli-within-block design, the full stimulus sample is divided into a smaller number of comparable lists or blocks, and each participant is randomly assigned to receive only one of these blocks. In the previous example of an experiment involving a sample of 60 film clips or songs, these stimuli could be divided into, say, 10 blocks of six stimuli each. If each participant responds to only one of these blocks, then

Table 3
Schematic for a Particular Stimuli-Within-Block Design That Is an Extension of the Stimuli-Within-Condition Design

	Stimuli															
Participants	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	A	A	B	B	—	—	—	—	—	—	—	—	—	—	—	—
2	A	A	B	B	—	—	—	—	—	—	—	—	—	—	—	—
3	A	A	B	B	—	—	—	—	—	—	—	—	—	—	—	—
4	—	—	—	—	A	A	B	B	—	—	—	—	—	—	—	—
5	—	—	—	—	A	A	B	B	—	—	—	—	—	—	—	—
6	—	—	—	—	A	A	B	B	—	—	—	—	—	—	—	—
7	—	—	—	—	—	—	—	—	A	A	B	B	—	—	—	—
8	—	—	—	—	—	—	—	—	A	A	B	B	—	—	—	—
9	—	—	—	—	—	—	—	—	A	A	B	B	—	—	—	—
10	—	—	—	—	—	—	—	—	—	—	—	—	A	A	B	B
11	—	—	—	—	—	—	—	—	—	—	—	—	A	A	B	B
12	—	—	—	—	—	—	—	—	—	—	—	—	A	A	B	B

Note. Each participant is randomly assigned to respond to one of four stimulus blocks, each block containing four stimuli. A means that this participant responds to this stimulus only under Condition A; B means that this participant responds to this stimulus only under Condition B. A dash means that this participant never responds to this stimulus.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

each experimental session could take around 30 minutes. The promise of stimuli-within-block designs is that as long as the researcher is willing to spend some extra time assembling a larger stimulus sample, the study can enjoy the greater statistical power associated with larger stimulus samples while still avoiding placing heavy time demands on the participants.

To give a sense of the relative statistical efficiency of stimuli-within-block designs compared to the more commonly used crossed designs, we conducted a simulation examining the experimental design mentioned above, where 60 participants each respond to six stimuli, three in each condition. In this simulation we held constant the number of participants and the number of responses made by each participant, but varied the number of unique stimulus blocks from one to 60, so that the total number of stimuli ranged from six to 360 across the simulated experiments, with the assumption that stimuli are randomly assigned to blocks.[5] Note that when there is a single stimulus block, every participant responds to the same set of stimuli, and this particular stimuli-within-block design reduces to the crossed stimuli-within-condition design. At the other extreme, when the number of stimulus blocks equals the number of participants, then every participant responds to their own unique set of stimuli in what is commonly known as a *nested* design (Raudenbush, 1997; Raudenbush & Liu, 2000; Snijders, 2001; Snijders & Bosker, 1993). The relative statistical efficiency of these three types of designs, defined as the relative widths of the standard errors of the condition difference, is illustrated in Figure 8, which is based on the standard case variance components described earlier. In this example, increasing the number of stimulus blocks from one to 10 (so that the total number of stimuli increases from six to 60) increases statistical efficiency by a factor of about 2.3.

The major take-away point here is that even in experimental paradigms that involve time-consuming stimulus presentation, it is still possible to employ a reasonably large sample of stimuli by using a stimuli-within-block design; treating stimuli as random need not necessarily lead to a substantial decrease in statistical power even in these paradigms.

Sometimes it is even possible to assign a unique set of stimuli to every participant, leading to a nested design. From a logistical perspective, running an experiment with a fully nested design might be difficult, as it not only requires a very large number of stimuli but also requires that the experiment be conducted in such a way that no participant ever receives a stimulus that has been responded to by another participant. However, in those situations where it is feasible, there are potentially large statistical power benefits to employing this kind of nested design compared to the crossed designs that we have focused on in this article. Notably, in addition to the statistical power advantage illustrated in Figure 8, in fully nested designs statistical power does approach unity as the number of participants goes to infinity, unlike in the crossed and stimuli-within-block designs.

Concluding Remarks

For all too many years, until the groundbreaking work of Jacob Cohen and others, experimental psychologists paid little attention to issues of statistical power. Fortunately, that situation has begun to change. There are now discussions about how experiments should be optimally designed to maximize statistical power, although most of these discussions focus simply on making sure that adequate numbers of participants are recruited (for exceptions see Baguley, 2004; McClelland, 1997, 2000).

Hardly ever, however, in this increasing discussion of optimal design and statistical power does one hear mention of the sampling of stimuli to which participants are exposed. We all have rules of thumb in particular contexts about how stimuli should be constructed, about pretesting that ought to be done to ensure that stimuli represent what we hope they do, and about the approximate numbers of stimuli that should be employed. But by and large these rules of thumb are not informed by considerations of statistical power. Our goal in writing this article is to begin to change

[5] We note that this is not the only reasonable way that stimuli could be divided into blocks. For example, blocks could be deliberately constructed so that the variation across blocks in their overall mean responses or their condition differences was as small as possible, and these decisions would have power consequences. However, we consider these issues to be outside the scope of our limited discussion here, and so for simplicity we just assume that stimuli are randomly assigned to blocks.

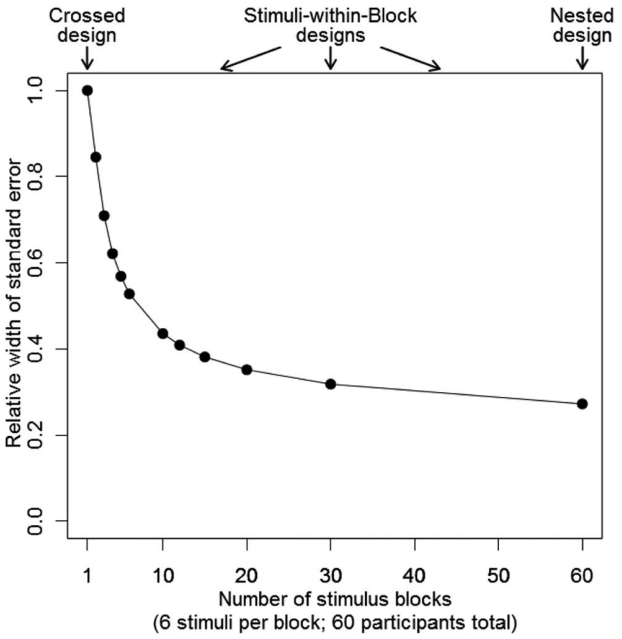

Figure 8. Simulation results illustrating the relative statistical efficiency of stimuli-within-block designs compared to a crossed design and a nested design. In the stimuli-within-block designs, 60 participants are randomly assigned to respond to just one of a series of stimulus blocks, where each stimulus block consists of three stimuli from Condition A and three stimuli from Condition B. If every participant receives a unique stimulus block, then this corresponds to a nested design. If every participant responds to the same stimulus block, then this corresponds to a crossed design. The numbers of stimulus blocks examined were 1, 2, 3, 4, 5, 6, 10, 12, 15, 20, 30, and 60, and the plotted points are the median simulation estimates at each number of blocks.

that. Our hope is that researchers will come to realize that stimulus sampling issues can have the same impact on statistical power considerations as participant sampling issues. Moreover, we have created a set of practical tools that can assist the researcher in making choices about design and the sample sizes of both participants and stimulus.

We are also convinced that because of the failure to take stimulus sampling into account in statistical models commonly used to analyze experimental data, many "significant" results reported in the literature are not replicable when different samples of stimuli are used in addition to different samples of participants. Of course, we do acknowledge that there are situations in which a sample of stimuli may in fact exhaust or nearly exhaust the population of stimuli that are of theoretical interest. For instance, imagine a study in which participants make responses as a function of whether a presented letter is a vowel or consonant. In such a study one might essentially exhaust the two categories of vowel and consonant with the specific letters used. In such a case, it may seem unreasonable to be concerned about generalization to other possible stimulus samples. Our experience, however, suggests that situations like these, where the stimulus samples that are used essentially exhaust the population, are rare indeed. More generally, we are wary of researchers trying to have it both ways—performing an analysis that does not permit any generalization beyond the specific stimuli used in the study, but then interpreting the results of that analysis as if they applied to the entire class of stimuli. Our goal, therefore, in addition to thinking about the impact of stimulus sampling for optimal design, has also been to alert researchers to the need to explicitly model stimulus variability in their analyses, treating stimuli as well as participants as random factors across which generalization is sought. We hope that one impact of this article will be in helping to promote the more widespread use in experimental psychology of linear mixed models with crossed random effects, which we believe remain underutilized in part because of their lower statistical power compared to the traditional procedures that ignore stimulus sampling (although this is not always the case; see e.g., Coleman, 1979).

Up to this point we have said nothing in the way of recommendations for exactly how stimuli should be sampled, or for that matter how participants should be sampled. Critical literature exists on both of these topics (e.g., Brunswik, 1955; Henrich, Heine, & Norenzayan, 2010), and in general we leave this issue to the scientific judgment of the experimenter. We do wish, however, to offer a way of thinking about the trade-offs that must often be made to satisfy the goals of (a) designing an experiment that is statistically efficient as well as logistically and financially feasible, and (b) using samples that are clearly representative in all relevant respects of the populations that we ultimately wish to speak about.

In a classic statistical article, Cornfield and Tukey (1956, p. 912) speak of "the two spans of the bridge of inference" as an analogy for the process of inference from the actual observations at hand to general conclusions about some wider populations of interest. In psychology, many experiments involve something like the following: a sample of students from an American university respond to a sample of photographs drawn from high school yearbooks. Strictly speaking, our inferential statistics cannot support generalizations beyond the restricted populations from which these can be considered random samples (e.g., all American college students responding to all yearbook-style photographs of high school students), and even then only if the appropriate statistical procedures have been used (Judd et al., 2012). Yet researchers often deem it acceptable to draw wider conclusions about, for example, all people responding to the appearance of another person. According to Cornfield and Tukey (1956),

> if we take the simile of the bridge crossing a river by way of an island, there is a statistical span from the near bank to the island, and a subject-matter span from the island to the far bank. Both are important. (p. 913)

By using different participant and stimulus samples or by analyzing the data in different ways, the "island" or restricted population can be moved closer to the near bank and further from the far bank, or vice versa. The consequences of this are that it may become easier to cross the statistical span from the sample to the restricted population but correspondingly more difficult to cross the further span to the wider population that we would ultimately like to speak about (or vice versa). For example, if an experiment uses a tightly controlled stimulus set that has been constructed to be as homogeneous as possible, the lack of stimulus variation will

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

often make it especially easy to obtain very precise estimates of experimental effects in this narrow population of stimuli, but it will often be less obvious that these precisely estimated effects are representative of those found in more natural settings. As another example, if one analyzes the data using a statistical model that incorporates uncertainty in the parameter estimates due to random stimulus variation (Judd et al., 2012), it will often be more difficult to obtain precise estimates of effects—a cost that we hope the lessons of this article will help to mitigate— but upon crossing this lengthened statistical span, it will then be easier to generalize the results to further, more interesting populations.

Though much has been covered here, the scope of our discussion is limited in several key ways. We have confined ourselves to only a few designs, involving only two levels of one fixed factor, when in practice many more complicated designs are used. In theory the present work could be extended to higher designs involving multiple categorical fixed factors, although these extensions are not without some complications, such as a larger number of random slopes to consider; we leave it to future research to work out these important issues. We have also confined ourselves to outcome measures that are nicely behaved, so that the underlying distributions of random effects in our models could be assumed to be normal. And we have confined ourselves to considering issues of optimal design and power only in relatively abstract contexts, divorced from the realities and constraints of any particular research paradigm and endeavor. We recognize that issues of design and power are necessarily more complicated than the relatively simple framework we have used here.

That said, we would be very gratified indeed if this article begins the discussion of how stimuli ought to be sampled in experimental research to achieve desirable levels of statistical power. We are convinced that the failure to recognize the random variation in data that results from stimulus variability has led us to give confidence to results that might routinely be nonreplicable except with the exact same stimuli. Moreover, we are convinced that the failure to appreciate the role of stimulus variability seriously affects the statistical power of much of our research. Hopefully, the tools we have provided can help us as a discipline to begin to overcome these problems.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108–119. doi:10.1002/per.1919

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. doi:10.1016/j.jml.2007.12.005

Baguley, T. (2004). Understanding statistical power in the context of applied research. *Applied Ergonomics, 35,* 73–80. doi:10.1016/j.apergo.2004.01.002

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7,* 543–554. doi:10.1177/1745691612459060

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278. doi:10.1016/j.jml.2012.11.001

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50,* 217–224. doi:10.1016/j.jesp.2013.10.005

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193–217. doi:10.1037/h0047470

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,* 365–376. doi:10.1038/nrn3475

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,* 335–359. doi:10.1016/S0022-5371(73)80014-3

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

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

Coleman, E. B. (1979). Generalization effects vs random effects: Is σTL2 a source of Type 1 or Type 2 error? *Journal of Verbal Learning & Verbal Behavior, 18,* 243–256. doi:10.1016/S0022-5371(79)90145-2

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

Dixon, W. J., & Massey, F. J. (1957). *Introduction to statistical analysis*. New York, NY: McGraw-Hill.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19,* 975–991. doi:10.3758/s13423-012-0322-y

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics, 1,* 223–231. doi:10.1207/S15328031US0104_02

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11,* 323–343. doi:10.1037/1082-989X.11.4.323

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33,* 61–83. doi:10.1017/S0140525X0999152X

Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 199–209. doi:10.1037/0096-1523.32.2.199

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19,* 640–648. doi:10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7,* 645–654. doi:10.1177/1745691612464056

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,* 54–69. doi:10.1037/a0028347

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York, NY: Guilford Press.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science, 7,* 608–614. doi:10.1177/1745691612462586

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

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods, 2,* 3–19. doi:10.1037/1082-989X.2.1.3

McClelland, G. H. (2000). Increasing statistical power without increasing sample size. *American Psychologist, 55,* 963–964. doi:10.1037/0003-066X.55.8.963

Montgomery, D. C. (2013). *Design and analysis of experiments* (Vol. 8). New York, NY: Wiley.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615–631. doi:10.1177/1745691612459058

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7,* 657–660. doi:10.1177/1745691612462588

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

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

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin, 92,* 500–504. doi:10.1037/0033-2909.92.2.500

Santa, J. L., Miller, J. J., & Shaw, M. L. (1979). Using quasi *F* to prevent alpha inflation due to stimulus variation. *Psychological Bulletin, 86,* 37–46. doi:10.1037/0033-2909.86.1.37

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

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Cambridge, MA: Ballinger.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Snijders, T. A. B. (2001). Sampling. In H. Goldstein & A. Leyland (Eds.), *Multilevel modeling of health statistics* (pp. 159–174). New York, NY: Wiley.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18,* 237–259. doi:10.3102/10769986018003237

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25,* 1115–1125. doi:10.1177/01461672992512005

Westfall, J., Judd, C. M., & Kenny, D. A. (2014). Replicating studies in which samples of participants respond to samples of stimuli.

Winer, B. J. (1971). *Statistical principles in experimental design*. New York, NY: McGraw–Hill.

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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix

Statistical Details of Power Analysis

In this Appendix, we provide the statistical details underlying the power results in the main article. We consider only experiments where participants respond to each stimulus no more than once per condition; in other words, we assume only a single replication at the lowest level of observation. Within this constraint, we consider all potential sources of random variation. Consequently, the fully specified mixed model for the response of the *i*th participant to the *j*th stimulus in the *k*th condition is

$$y_{ijk} = \beta_0 + \beta_1 c_k + \alpha_i^P + \alpha_i^{P\times C} c_k + \alpha_j^S + \alpha_j^{S\times C} c_k + \alpha_{ij}^{P\times S} + \varepsilon_{ijk},$$
$$\text{var}\left(\alpha_i^P\right) = \sigma_P^2, \quad \text{var}\left(\alpha_i^{P\times C}\right) = \sigma_{P\times C}^2, \quad \text{var}\left(\alpha_j^S\right) = \sigma_S^2,$$
$$\text{var}\left(\alpha_j^{S\times C}\right) = \sigma_{S\times C}^2, \quad \text{var}\left(\alpha_{ij}^{P\times S}\right) = \sigma_{P\times S}^2, \quad \text{var}\left(\varepsilon_{ijk}\right) = \sigma_E^2.$$

In this model, β_0 and $\beta_1 c_k$ represent the fixed effects and capture, respectively, the overall mean response and the condition difference in responses. We assume the values of c_k are contrast or deviation coded so that $c_1 + c_2 = 0$ and $c_1^2 = c_2^2 = c^2$. For example, the values of c_k might be -1, $+1$, or $-1/2$, $+1/2$. Importantly, the values of c_k are assumed not to be dummy or treatment coded (e.g., $c_1 = 0, c_2 = 1$), as this totally alters the meanings of the variance components (see also Barr, Levy, Scheepers, & Tily, 2013, footnote 2). There are also two covariance terms—that between the participant intercepts and slopes, $\sigma_{P,\ P\times C}$, and between the stimulus intercepts and slopes, $\sigma_{S,\ S\times C}$— but these can be omitted as unnecessary for reasons that are explained later in this Appendix. Finally, in what follows we assume that the experiments are balanced and have no missing data.

We follow Cohen's (1988) general approach to power analysis. In this approach, power estimation can be viewed as a three-stage process. We first select a suitable measure of effect size and define its value. In this article we rely on the standardized mean difference between conditions, Cohen's *d*. Next we rescale the effect size by an adjustment factor, sometimes called the *design effect* (Snijders, 2001), that depends on the particular experimental design chosen, yielding what Cohen (1988, p. 13) refers to as the *operative effect size*. Finally, we scale the operative effect size by a term that includes information about the number of observations in the study. This final quantity is known as the *noncentrality parameter* (*ncp*), and with it and the degrees of freedom (*df*) we can compute power. This last step is often hidden in power analyses, but it shows how the operative effect size is adjusted by the sample sizes to compute a power value. To go from the noncentrality parameter to a power estimate, Cohen (1988, p. 544) relied on an approximation put forward by Dixon and Massey (1957). We have adopted a more exact approach that directly computes areas under the appropriate noncentral *t* distribution.

(*Appendix continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Effect Size and Variance Partitioning Coefficients

With stimuli as well as participants as random factors in a design, the simple effect size measure d must be defined more generally to reflect all of these possible variation sources. That is, instead of defining the effect size in terms of the expected mean difference divided by the standard deviation across participants within condition, we define the effect size as the expected mean difference divided by the standard deviation across participants and stimuli within condition (i.e., the square root of the expected variance of the individual observations within a condition, pooled across both conditions). Let $\text{var}(y_1)$ and $\text{var}(y_2)$ be the variances of the response at the two levels of the contrast-coded predictors. Then the pooled variance is

$$\begin{aligned}\frac{\text{var}(y_1)}{2} + \frac{\text{var}(y_2)}{2} &= \frac{\left(\sigma_P^2 + c^2\sigma_{P\times C}^2 + \sigma_S^2 + c^2\sigma_{S\times C}^2 + \sigma_{P\times S}^2 + \sigma_E^2 + 2c_1\sigma_{P,P\times C} + 2c_1\sigma_{S,S\times C}\right)}{2} \\ &\quad + \frac{\left(\sigma_P^2 + c^2\sigma_{P\times C}^2 + \sigma_S^2 + c^2\sigma_{S\times C}^2 + \sigma_{P\times S}^2 + \sigma_E^2 + 2c_2\sigma_{P,P\times C} + 2c_2\sigma_{S,S\times C}\right)}{2} \\ &= \sigma_P^2 + c^2\sigma_{P\times C}^2 + \sigma_S^2 + c^2\sigma_{S\times C}^2 + \sigma_{P\times S}^2 + \sigma_E^2 + \quad (\sigma_{P,P\times C} + \sigma_{S,S\times C})\ (c_1 + c_2) \\ &= \sigma_P^2 + c^2\sigma_{P\times C}^2 + \sigma_S^2 + c^2\sigma_{S\times C}^2 + \sigma_{P\times S}^2 + \sigma_E^2.\end{aligned}$$

Notice that the covariance terms drop out, using the fact that $c_1 + c_2 = 0$ (i.e., that the predictor is contrast coded). This makes the effect size based on the pooled variance

$$d = \frac{\mu_1 - \mu_2}{\sigma} = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_P^2 + c^2\sigma_{P\times C}^2 + \sigma_S^2 + c^2\sigma_{S\times C}^2 + \sigma_{P\times S}^2 + \sigma_E^2}}.$$

Technically, the denominator of this effect size (i.e., the pooled variance) is a function of the predictor c, so that the variance of y is potentially different at different levels of the predictor, a fact pointed out by Goldstein et al. (2002). However, we note that under our assumption that the predictor is contrast coded, this denominator is identical at both levels of c.

We standardize the variance components by expressing them as variance partitioning coefficients (VPCs; Goldstein et al., 2002), specifically, as proportions of the total pooled variance. This leads to the following definitions of all standardized and unstandardized quantities in Table A1.

Table A1
Definitions of All Standardized and Unstandardized Quantities

Unstandardized	Standardized
$\mu_1 - \mu_2$ (Effect size)	$d = \frac{\mu_1 - \mu_2}{\sigma_P^2 + c^2\sigma_{P\times C}^2 + \sigma_S^2 + c^2\sigma_{S\times C}^2 + \sigma_E^2 + \sigma_{P\times S}^2}$
σ_P^2 (Participant intercept variance)	$V_P = \frac{\sigma_P^2}{\sigma_P^2 + c^2\sigma_{P\times C}^2 + \sigma_S^2 + c^2\sigma_{S\times C}^2 + \sigma_E^2 + \sigma_{P\times S}^2}$
σ_S^2 (Stimulus intercept variance)	$V_S = \frac{\sigma_S^2}{\sigma_P^2 + c^2\sigma_{P\times C}^2 + \sigma_S^2 + c^2\sigma_{S\times C}^2 + \sigma_E^2 + \sigma_{P\times S}^2}$
$\sigma_{P\times C}^2$ (Participant slope variance)	$V_{P\times c} = \frac{c^2\sigma_{P\times C}^2}{\sigma_P^2 + c^2\sigma_{P\times C}^2 + \sigma_S^2 + c^2\sigma_{S\times C}^2 + \sigma_E^2 + \sigma_{P\times S}^2}$
$\sigma_{S\times C}^2$ (Stimulus slope variance)	$V_{S\times c} = \frac{c^2\sigma_{S\times C}^2}{\sigma_P^2 + c^2\sigma_{P\times C}^2 + \sigma_S^2 + c^2\sigma_{S\times C}^2 + \sigma_E^2 + \sigma_{P\times S}^2}$
$\sigma_{P\times S}^2$ (Participant by Stimulus variance)	$V_{P\times S} = \frac{\sigma_{P\times S}^2}{\sigma_P^2 + c^2\sigma_{P\times C}^2 + \sigma_S^2 + c^2\sigma_{S\times C}^2 + \sigma_E^2 + \sigma_{P\times S}^2}$
σ_E^2 (Residual variance)	$V_E = 1 - (V_P + V_{P\times C} + V_S + V_{S\times C} + V_{P\times S})$

(*Appendix continues*)

The reason the term c^2 appears in the equations in Table A1 is that the variances of the random slope terms—which are multiplied by c in the mixed model regression equation—must be multiplied by some scaling of c for the value of the overall equation to be invariant to the scale of c (e.g., whether the values of c are -1 and $+1$ or $-1/2$ and $+1/2$). Note that with values of c of -1 and $+1$, $c^2 = 1$, in which case (and *only* in which case) the c^2 term can be disregarded.

Operative Effect Sizes and Mixed Model Equations

In this section, we give the mixed model equations for each of the five designs as they would be estimated from the data, and we give the operative or design-specific effect sizes for each design, denoted d_0. These are given both in terms of the standardized and unstandardized variance components. We feel these are helpful for very clearly showing which variance components contribute to the standard error in each design and which do not.

We also develop a notation to indicate which variance components from the fully specified mixed model are empirically confounded in each design, if any. Specifically, in the expressions given below, variance components that are joined together in brackets are empirically confounded in the given design. When variance components are estimated from data, the first term within a set of brackets implicitly includes the second term in those brackets. In the fully crossed design, all of the variance components are unconfounded and so are all uniquely estimable after the data have been collected. In the other designs some of these components are confounded, and thus not uniquely estimable, in ways that we indicate below. For instance, in the counterbalanced design, the estimated residual error variance also includes the P × S variance component, and so we denote it with $[\sigma^2_E + \sigma^2_{P\times S}]$. In this notation, the first terms can be estimated, but the remaining terms are confounded within that design. There are three types of confounding that occur for the five designs: P × C with P, S × C with S, and P × S with E. Nonetheless, for these designs the effect size can still be computed and power can be estimated.

For each design, we present the model equation, the effect size or d, and the operative effect size or d_0, both unstandardized and standardized.

Fully Crossed Design

$$y_{ijk} = \beta_0 + \beta_1 c_k + \alpha_i^P + \alpha_i^{P\times C} c_k + \alpha_j^S + \alpha_j^{S\times C} c_k + \alpha_{ij}^{P\times S} + \varepsilon_{ijk}$$

$$d = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_P^2 + c^2\sigma_{P\times C}^2 + \sigma_S^2 + c^2\sigma_{S\times C}^2 + \sigma_{P\times S}^2 + \sigma_E^2}}$$

$$d_0 = \frac{\mu_1 - \mu_2}{\sqrt{c^2\sigma_{P\times C}^2 + c^2\sigma_{S\times C}^2 + \sigma_E^2}}$$

$$d_0 = \frac{d}{\sqrt{1 - V_P - V_S - V_{P\times S}}} = \frac{d}{\sqrt{V_{P\times C} + V_{S\times C} + V_E}}$$

Counterbalanced Design

$$y_{ijk} = \beta_0 + \beta_1 c_k + \alpha_i^P + \alpha_i^{P\times C} c_k + \alpha_j^S + \alpha_j^{S\times C} c_k + \varepsilon_{ijk}$$

$$d = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_P^2 + c^2\sigma_{P\times C}^2 + \sigma_S^2 + c^2\sigma_{S\times C}^2 + \left[\sigma_E^2 + \sigma_{P\times S}^2\right]}}$$

$$d_0 = \frac{\mu_1 - \mu_2}{\sqrt{c^2\sigma_{P\times C}^2 + c^2\sigma_{S\times C}^2 + \left[\sigma_E^2 + \sigma_{P\times S}^2\right]}}$$

$$d_0 = \frac{d}{\sqrt{1 - V_P - V_S}} = \frac{d}{\sqrt{V_{P\times C} + V_{S\times C} + \left[V_E + V_{P\times S}\right]}}$$

(*Appendix continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Stimuli-Within-Condition Design

$$y_{ijk} = \beta_0 + \beta_1 c_k + \alpha_i^P + \alpha_i^{P\times C} c_k + \alpha_j^S + \varepsilon_{ijk}$$

$$d = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_P^2 + c^2\sigma_{P\times C}^2 + \left[\sigma_S^2 + c^2\sigma_{S\times C}^2\right] + \left[\sigma_E^2 + \sigma_{P\times S}^2\right]}}$$

$$d_0 = \frac{\mu_1 - \mu_2}{\sqrt{c^2\sigma_{P\times C}^2 + \left[\sigma_S^2 + c^2\sigma_{S\times C}^2\right] + \left[\sigma_E^2 + \sigma_{P\times S}^2\right]}}$$

$$d_0 = \frac{d}{\sqrt{1 - V_P}} = \frac{d}{\sqrt{V_{P\times C} + \left[V_S + V_{S\times C}\right] + \left[V_E + V_{P\times S}\right]}}$$

Participants-Within-Condition Design

$$y_{ijk} = \beta_0 + \beta_1 c_k + \alpha_i^P + \alpha_j^S + \alpha_j^{S\times C} c_k + \varepsilon_{ijk}$$

$$d = \frac{\mu_1 - \mu_2}{\sqrt{\left[\sigma_P^2 + c^2\sigma_{P\times C}^2\right] + \sigma_S^2 + c^2\sigma_{S\times C}^2 + \left[\sigma_E^2 + \sigma_{P\times S}^2\right]}}$$

$$d_0 = \frac{\mu_1 - \mu_2}{\sqrt{\left[\sigma_P^2 + c^2\sigma_{P\times C}^2\right] + c^2\sigma_{S\times C}^2 + \left[\sigma_E^2 + \sigma_{P\times S}^2\right]}}$$

$$d_0 = \frac{d}{\sqrt{1 - V_S}} = \frac{d}{\sqrt{\left[V_P + V_{P\times C}\right] + V_{S\times C} + \left[V_E + V_{P\times S}\right]}}$$

Both-Within-Condition Design

$$y_{ijk} = \beta_0 + \beta_1 c_k + \alpha_i^P + \alpha_j^S + \varepsilon_{ijk}$$

$$d = \frac{\mu_1 - \mu_2}{\sqrt{\left[\sigma_P^2 + c^2\sigma_{P\times C}^2\right] + \left[\sigma_S^2 + c^2\sigma_{S\times C}^2\right] + \left[\sigma_E^2 + \sigma_{P\times S}^2\right]}}$$

$$d_0 = \frac{\mu_1 - \mu_2}{\sqrt{\left[\sigma_P^2 + c^2\sigma_{P\times C}^2\right] + \left[\sigma_S^2 + c^2\sigma_{S\times C}^2\right] + \left[\sigma_E^2 + \sigma_{P\times S}^2\right]}}$$

$$d_0 = d$$

Noncentrality Parameters and Degrees of Freedom

In the last section of this Appendix we show the derivation of the noncentrality parameter for just one of the five designs, and we describe the construction of the approximate degrees of freedom for each design based on the Satterthwaite approximation. With the noncentrality parameter and degrees of freedom, one can compute power by evaluating areas under the curve of the corresponding noncentral *t* distribution. To determine the noncentrality parameter, we need to determine the standard error of the estimate of the condition effect.

The noncentrality parameter that we derive is for the stimuli-within-condition design. This design is selected for illustrative purposes just because it demonstrates how both nesting and crossing of random factors with the predictor variable are handled. The noncentrality parameter can be viewed as the population quantity estimated by the sample *t* statistic (see Rosenthal & Rubin, 1982, Appendix), and as

(*Appendix continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

such it has the form $E[\hat{\beta}_1]/\sqrt{\text{var}(\hat{\beta}_1)}$. Under our assumption of a balanced design, $\hat{\beta}_1$ is proportional to the simple mean difference between the two conditions, $\overline{y_{\cdot\cdot 1}} - \overline{y_{\cdot\cdot 2}}$. So we need only find the variance of the mean difference, $\text{var}(\overline{y_{\cdot\cdot 1}} - \overline{y_{\cdot\cdot 2}})$, to determine the noncentrality parameter. We find this as follows, letting $j = 1, \ldots, m$ index the stimuli separately within each of the two conditions (note that there are q stimuli in total so that $2m = q$):

$$\begin{aligned}
\text{var}(\overline{y_{\cdot\cdot 1}} - \overline{y_{\cdot\cdot 2}}) &= \text{var}\left[(pm)^{-1}\Sigma_{ij}\,(y_{ij1} - y_{ij2})\right] \\
&= (pm)^{-2}\text{var}\left[m\Sigma_i\left(\alpha_i^{P\times C}c_1 - \alpha_i^{P\times C}c_2\right) + p\Sigma_j\,(\alpha_{j1}^{S} - \alpha_{j2}^{S}) + \Sigma_{ij}\,(\varepsilon_{ij1} - \varepsilon_{ij2})\right] \\
&= (pm)^{-2}\Big[m^2\text{var}\left(\Sigma_i\alpha_i^{P\times C}c_1 - \alpha_i^{P\times C}c_2\right) + p^2\text{var}\big(\Sigma_j\alpha_{j1}^{S} - \alpha_{j2}^{S}\big) + \text{var}\big(\Sigma_{ij}\varepsilon_{ij1} - \varepsilon_{ij2}\big)\Big] \\
&= (pm)^{-2}\big(2pm^2\,(c_1 - c_2)^2\sigma_{P\times C}^2 + 2p^2m\sigma_S^2 + 2pm\sigma_E^2\big) \\
&= (pm)^{-2}\big(4c^2pm^2\sigma_{P\times C}^2 + 2p^2m\sigma_S^2 + 2pm\sigma_E^2\big) \\
&= 4\left(\frac{c^2\sigma_{P\times C}^2}{p} + \frac{\sigma_S^2}{q} + \frac{\sigma_E^2}{pq}\right).
\end{aligned}$$

The degrees of freedom (*df*) are based on the Satterthwaite approximation, which in turn is based on the mean squares for each design. Using the same stimuli-within-condition design as an example, the Satterthwaite degrees of freedom for the test of the condition difference in this design are

$$df = \frac{\left(MS_{P\times C} + MS_S - MS_E\right)^2}{\dfrac{MS_{P\times C}^2}{p-1} + \dfrac{MS_S^2}{q-2} + \dfrac{MS_E^2}{(p-1)\,(q-2)}},$$

where the mean squares have the following expectations:

$$E\left[MS_{P\times C}\right] = qc^2\sigma_{P\times C}^2 + \sigma_E^2, \quad E\left[MS_S\right] = p\sigma_S^2 + \sigma_E^2, \quad E\left[\,MS_E\right] = \sigma_E^2.$$

For more information on these degrees of freedom, see Winer (1971, pp. 375-378).

Notice that the previously mentioned covariance terms do not appear in the expressions of the noncentrality parameters or the degrees of freedom. Because none of the effect sizes, noncentrality parameters, or degrees of freedom depend on the covariance terms (under our assumptions of contrast-coded predictors and balanced designs), we can safely ignore the covariances in the power analyses reported in this article.

We used this same method for determining the noncentrality parameters and degrees of freedom for the five designs. Below we give for each design the noncentrality parameter and degrees of freedom (the latter with the expected values of the mean squares imputed), in terms of both the unstandardized variance components and the VPCs.

Fully Crossed Design

$$ncp = \frac{\mu_1 - \mu_2}{2\sqrt{\dfrac{c^2\sigma_{P\times C}^2}{p} + \dfrac{c^2\sigma_{S\times C}^2}{q} + \dfrac{\sigma_E^2}{2pq}}} = \frac{d}{2\sqrt{\dfrac{V_{P\times C}}{p} + \dfrac{V_{S\times C}}{q} + \dfrac{V_E}{2pq}}}$$

$$\begin{aligned}
df &= \frac{\left(qc^2\sigma_{P\times C}^2 + pc^2\sigma_{S\times C}^2 + \sigma_E^2\right)^2}{\dfrac{\left(qc^2\sigma_{P\times C}^2 + \sigma_E^2\right)^2}{p-1} + \dfrac{\left(pc^2\sigma_{S\times C}^2 + \sigma_E^2\right)^2}{q-1} + \dfrac{\left(\sigma_E^2\right)^2}{(p-1)(q-1)}} \\
&= \frac{\left(qV_{P\times C} + pV_{S\times C} + V_E\right)^2}{\dfrac{\left(qV_{P\times C} + V_E\right)^2}{p-1} + \dfrac{\left(pV_{S\times C} + V_E\right)^2}{q-1} + \dfrac{V_E^2}{(p-1)(q-1)}}
\end{aligned}$$

(*Appendix continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Counterbalanced Design

$$ncp = \frac{\mu_1 - \mu_2}{2\sqrt{\frac{c^2\sigma^2_{P\times C}}{p} + \frac{c^2\sigma^2_{S\times C}}{q} + \frac{\left[\sigma^2_E + \sigma^2_{P\times S}\right]}{pq}}} = \frac{d}{2\sqrt{\frac{V_{P\times C}}{p} + \frac{V_{S\times C}}{q} + \frac{\left[V_E + V_{P\times S}\right]}{pq}}}$$

$$df = \frac{\left(qc^2\sigma^2_{P\times C} + pc^2\sigma^2_{S\times C} + 2\left[\sigma^2_{P\times S} + \sigma^2_E\right]\right)^2}{\frac{\left(qc^2\sigma^2_{P\times C} + \left[\sigma^2_E + \sigma^2_{P\times S}\right]\right)^2}{p-2} + \frac{\left(pc^2\sigma^2_{S\times C} + \left[\sigma^2_E + \sigma^2_{P\times S}\right]\right)^2}{q-2} + \frac{4\left[\sigma^2_E + \sigma^2_{P\times S}\right]^2}{(p-2)(q-2)}}$$

$$= \frac{(qV_{P\times C} + pV_{S\times C} + 2V_E + V_{P\times S})^2}{\frac{\left(qV_{P\times C} + \left[V_E + V_{P\times S}\right]\right)^2}{p-2} + \frac{\left(pV_{S\times C} + \left[V_E + V_{P\times S}\right]\right)^2}{q-2} + \frac{4\left[V_E + V_{P\times S}\right]^2}{(p-2)(q-2)}}$$

Stimuli-Within-Condition Design

$$ncp = \frac{\mu_1 - \mu_2}{2\sqrt{\frac{c^2\sigma^2_{P\times C}}{p} + \frac{\left[\sigma^2_S + c^2\sigma^2_{S\times C}\right]}{q} + \frac{\left[\sigma^2_E + \sigma^2_{P\times S}\right]}{pq}}} = \frac{d}{2\sqrt{\frac{V_{P\times C}}{p} + \frac{\left[V_S + V_{S\times C}\right]}{q} + \frac{\left[V_E + V_{P\times S}\right]}{pq}}}$$

$$df = \frac{\left(qc^2\sigma^2_{P\times C} + 2p\left[\sigma^2_S + c^2\sigma^2_{S\times C}\right] + 2\left[\sigma^2_E + \sigma^2_{P\times S}\right]\right)^2}{\frac{\left(qc^2\sigma^2_{P\times C} + \left[\sigma^2_E + \sigma^2_{P\times S}\right]\right)^2}{p-1} + \frac{4\left(p\left[\sigma^2_S + c^2\sigma^2_{S\times C}\right] + \left[\sigma^2_E + \sigma^2_{P\times S}\right]\right)^2}{q-2} + \frac{4\left[\sigma^2_E + \sigma^2_{P\times S}\right]^2}{(p-1)(q-2)}}$$

$$= \frac{\left(qV_{P\times C} + 2p\left[V_S + V_{S\times C}\right] + 2\left[\ V_E + V_{P\times S}\right]\right)^2}{\frac{\left(qV_{P\times C} + \left[V_E + V_{P\times S}\right]\right)^2}{p-1} + \frac{4\left(p\left[V_S + V_{S\times C}\right] + \left[V_E + V_{P\times S}\right]\right)^2}{q-2} + \frac{4\left[V_E + V_{P\times S}\right]^2}{(p-1)(q-2)}}$$

Participants-Within-Condition Design

$$ncp = \frac{\mu_1 - \mu_2}{2\sqrt{\frac{\left[\sigma^2_P + c^2\sigma^2_{P\times C}\right]}{p} + \frac{c^2\sigma^2_{S\times C}}{q} + \frac{\left[\sigma^2_E + \sigma^2_{P\times S}\right]}{pq}}} = \frac{d}{2\sqrt{\frac{\left[V_P + V_{P\times C}\right]}{p} + \frac{V_{S\times C}}{q} + \frac{\left[V_E + V_{P\times S}\right]}{pq}}}$$

$$df = \frac{\left(2q\left[\sigma^2_P + c^2\sigma^2_{P\times C}\right] + pc^2\sigma^2_{S\times C} + 2\left[\sigma^2_E + \sigma^2_{P\times S}\right]\right)^2}{\frac{4\left(q\left[\sigma^2_P + c^2\sigma^2_{P\times C}\right] + \left[\sigma^2_E + \sigma^2_{P\times S}\right]\right)^2}{p-2} + \frac{\left(pc^2\sigma^2_{S\times C} + \left[\sigma^2_E + \sigma^2_{P\times S}\right]\right)^2}{q-1} + \frac{4\left[\sigma^2_E + \sigma^2_{P\times S}\right]^2}{(p-2)(q-1)}}$$

$$= \frac{\left(2q\left[V_P + V_{P\times C}\right] + pV_{S\times C} + 2\left[V_E + V_{P\times S}\right]\right)^2}{\frac{4\left(q\left[V_P + V_{P\times C}\right] + \left[V_E + V_{P\times S}\right]\right)^2}{p-2} + \frac{\left(pV_{S\times C} + \left[V_E + V_{P\times S}\right]\right)^2}{q-1} + \frac{4\left[V_E + V_{P\times S}\right]^2}{(p-2)(q-1)}}$$

(*Appendix continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Both-Within-Condition Design

$$ncp = \frac{\mu_1 - \mu_2}{2\sqrt{\frac{\left[\sigma_P^2 + c^2\sigma_{P\times C}^2\right]}{p} + \frac{\left[\sigma_S^2 + c^2\sigma_{S\times C}^2\right]}{q} + \frac{2\left[\sigma_E^2 + \sigma_{P\times S}^2\right]}{pq}}}$$

$$= \frac{d}{2\sqrt{\frac{\left[V_P + V_{P\times C}\right]}{p} + \frac{\left[V_S + V_{S\times C}\right]}{q} + \frac{2\left[V_E + V_{P\times S}\right]}{pq}}}$$

$$df = \frac{\left(q\left[\sigma_P^2 + c^2\sigma_{P\times C}^2\right] + p\left[\sigma_S^2 + c^2\sigma_{S\times C}^2\right] + 2\left[\sigma_E^2 + \sigma_{P\times S}^2\right]\right)^2}{\frac{\left(q\left[\sigma_P^2 + c^2\sigma_{P\times C}^2\right] + 2\left[\sigma_E^2 + \sigma_{P\times S}^2\right]\right)^2}{p-2} + \frac{\left(p\left[\sigma_S^2 + c^2\sigma_{S\times C}^2\right] + 2\left[\sigma_E^2 + \sigma_{P\times S}^2\right]\right)^2}{q-2} + \frac{4\left[\sigma_E^2 + \sigma_{P\times S}^2\right]^2}{(p-2)(q-2)}}$$

$$= \frac{\left(q\left[V_P + V_{P\times C}\right] + p\left[V_S + V_{S\times C}\right] + 2\left[V_E + V_{P\times S}\right]\right)^2}{\frac{\left(q\left[V_P + V_{P\times C}\right] + 2\left[V_E + V_{P\times S}\right]\right)^2}{p-2} + \frac{\left(p\left[V_S + V_{S\times C}\right] + 2\left[V_E + V_{P\times S}\right]\right)^2}{q-2} + \frac{4\left[V_E + V_{P\times S}\right]^2}{(p-2)(q-2)}}$$

Received February 11, 2014
Revision received May 6, 2014
Accepted May 30, 2014 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.