Experiments with More Than One Random Factor: Designs, Analytic Models, and Statistical Power

Charles M. Judd,[1] Jacob Westfall,[2] and David A. Kenny[3]

[1]Department of Psychology and Neuroscience, University of Colorado, Boulder, Colorado 80309; email: charles.judd@colorado.edu

[2]Department of Psychology, University of Texas at Austin, Austin, Texas 78712; email: jake.westfall@utexas.edu

[3]Department of Psychological Sciences, University of Connecticut, Storrs, Connecticut 06269; email: david.kenny@uconn.edu

Annu. Rev. Psychol. 2017. 68:601–25

First published online as a Review in Advance on September 28, 2016

The *Annual Review of Psychology* is online at psych.annualreviews.org

This article's doi: 10.1146/annurev-psych-122414-033702

Keywords

experimental design, mixed models, random factors, statistical power, sample size, effect size

Abstract

Traditional methods of analyzing data from psychological experiments are based on the assumption that there is a single random factor (normally participants) to which generalization is sought. However, many studies involve at least two random factors (e.g., participants and the targets to which they respond, such as words, pictures, or individuals). The application of traditional analytic methods to the data from such studies can result in serious bias in testing experimental effects. In this review, we develop a comprehensive typology of designs involving two random factors, which may be either crossed or nested, and one fixed factor, condition. We present appropriate linear mixed models for all designs and develop effect size measures. We provide the tools for power estimation for all designs. We then discuss issues of design choice, highlighting power and feasibility considerations. Our goal is to encourage appropriate analytic methods that produce replicable results for studies involving new samples of both participants and targets.

Contents

INTRODUCTION

Psychologists learn early in their statistical training to use analysis of variance procedures (*t*-tests and ANOVA) to analyze data from designs in which participants respond in various experimental conditions. In these designs and analyses, condition is a fixed factor, whereas participants are a random factor, meaning that the participants used in any particular study are thought to be a sample of participants that might have been used. In analyzing the data from such experiments, one obtains an estimate of the mean condition difference, as well as an estimate of the uncertainty surrounding that difference, by examining the variability across participants (i.e., across the random factor in the design). The goal is to determine whether the mean condition difference, given the variability of participants, is sufficiently large to permit the belief that it would continue to be found with other samples of participants.

However, many questions in psychology do not lend themselves easily to these well-learned analytic approaches. Frequently, research questions demand experiments that involve more than a single random factor across which generalization about condition differences should be sought. For instance, a memory researcher might be interested in memory for word lists under different conditions and wish to reach conclusions that generalize both to other samples of participants and to other samples of words that might have been used. Likewise, a social psychologist might ask participants to respond to faces of individuals coming from two different ethnic or racial categories. Here the goal would be to reach conclusions that generalize both to other participants and to other samples of faces that might have been used. Additionally, consider a clinical psychologist who is interested in showing that a new therapeutic approach for the treatment of depression is more effective than the standard approach. He or she might collect data from patients who are being treated by therapists under either the new or the standard approach. Again, generalization of any differences should reasonably be sought both across other patients and across other therapists that might have been studied.

Because psychological researchers are not routinely trained in the analysis of data from designs, such as those just illustrated, that have multiple random factors, all too often the data from such designs are inappropriately analyzed by collapsing across or ignoring one of the random factors so that the familiar *t*-tests and ANOVA procedures can be used. For instance, the memory researcher would typically compute a mean score for each participant for the word list as a whole; the social psychologist would compute, for each participant, means across faces within a racial category; and the depression researcher might simply ignore the therapists in the analysis. If the goal is to reach conclusions that generalize to both random factors, then these analyses are likely inappropriate because they have been shown to result in seriously inflated type I statistical errors, leading researchers to claim statistically significant effects that may be unlikely to replicate with different samples of words, faces, or therapists (Clark 1973, Judd et al. 2012). Many failures to replicate experimental results likely stem from this (Westfall et al. 2015).

To remedy these errors, in this review we provide a thorough treatment of the design and analysis of experiments in psychology that have more than one random factor. In such experiments, rather than having a single source of error variation in the data that arises from a single random factor (e.g., participants), there exist multiple sources of error variation arising from multiple random factors (e.g., words as well as participants, faces as well as participants, therapists as well as patients). Given this fact, a more general analytic approach is necessary, in which those multiple sources of random variation are explicitly modeled and estimated. This more general analytic approach relies on what are called generalized linear mixed models (Bolker et al. 2009, Stroup 2012). We provide a thorough treatment of this approach in the context of psychological experimental designs having two random factors.

We begin with the familiar designs involving only one random factor, participants, and a single fixed condition factor having two levels. These are the experimental designs for which the well-learned *t*-tests and ANOVA procedures are appropriate. We show how these procedures can be recast into the mixed-model framework so that the familiar analyses become special cases of mixed-model analyses.

We then turn to designs having two random factors (which we call participants and targets) and one fixed factor (which we call condition), and present the full mixed model that identifies all the sources of variation in the resulting data. We develop a comprehensive typology of all such designs, including designs in which the two random factors are crossed and designs in which one random factor is nested within the other. For each design, we give specifics about estimation and discuss an effect size estimate that is modeled on Cohen's *d* (i.e., the standardized mean difference; Cohen 1988) but generalized to the current designs involving two random factors.

Next we develop procedures for the estimation of statistical power in the context of the designs considered, including providing access to a web-based application for power estimation. In light of this, we discuss considerations relating to sample sizes, design choices, and the efficiency of alternative designs.

In the concluding sections of the review, we expand the design possibilities, discussing designs with more than two levels of condition, with multiple fixed factors, with more than two random factors, and with dyadic data.

MIXED MODELS FOR DESIGNS WITH ONE RANDOM AND ONE FIXED FACTOR

We begin with familiar designs in which there is one fixed factor, condition, having two levels, and only one random factor, participants. For instance, imagine that we are interested in task performance under stress. We are comparing the responses of participants under two conditions,

with and without stress. In this context, there are two possible designs: participants are in both conditions or participants are in only one condition. The former is typically called a within-participant design, whereas the latter is called a between-participant design (Smith 2014). We refer to the first design as the C design, meaning that participants are crossed with condition, and the second as the N design, meaning that participants are nested within the levels of condition. The standard least-squares analysis for data from the C design is the paired *t*-test or, equivalently, a repeated-measures ANOVA. For data from the N design, the standard analysis is the independent samples *t*-test or a between-subject ANOVA.

To recast the analysis of data from these designs into the mixed-model terminology, we first specify the possible sources of variation in the observations from these designs. We assume participants are measured on a single dependent variable, Y_{ik}, where i refers to the individual participant and k to the condition under which the observation is taken. The mixed-model specification of the individual observations can be written as follows:

$$Y_{ik} = \beta_0 + \beta_1 c_{ik} + \alpha_i^P + \alpha_i^{P\times C} c_{ik} + \varepsilon_{ik}. \tag{1}$$

The values of c_{ik} represent condition and are assumed to be contrast- or deviation-coded[1] (i.e., $c_{i1} = 1$ and $c_{i2} = -1$). The terms β_0 and β_1 represent the fixed effects and capture the overall mean response (β_0) and the condition difference in responses (β_1). In the mixed-model terminology, β_0 is the fixed intercept and β_1 is the fixed slope of condition. What makes this a mixed model is that, in addition to these fixed sources of variation in the data, there are multiple sources of variation that are random in the sense that they vary across the participants in the design. The following are the random components of variation in the observations:

$$\text{var}\left(\alpha_i^P\right) = \sigma_P^2, \qquad \text{var}\left(\alpha_i^{P\times C}\right) = \sigma_{P\times C}^2, \qquad \text{var}(\varepsilon_{ik}) = \sigma_E^2.$$

The variance attributable to participant mean differences is designated as σ_P^2. In the language of mixed models, this is the random variation across participants in their intercepts. The variance attributable to participant differences in the condition effects (i.e., participant-by-condition interaction effects) is $\sigma_{P\times C}^2$. In the language of mixed models, this is the random variation across participants in their condition slopes. And finally, σ_E^2 represents residual random error variation in the observations. These variances are also in the standard ANOVA approach to these designs; the mixed-model specification makes them explicit. Additionally, in the mixed-model specification, a possible covariance is explicitly considered between participant intercepts and their slopes,

$$\text{cov}\left(\alpha_i^P,\ \alpha_i^{P\times C}\right) = \sigma_{P,P\times C},$$

allowing those participants with higher mean responses to have smaller or larger condition differences. This covariance is typically ignored in the standard ANOVA approach.

The mixed model given in Equation 1 can be rewritten to make clear that the α_i^P and $\alpha_i^{P\times C}$ terms represent random variation in the intercepts and slopes across participants:

$$Y_{ik} = \underbrace{(\beta_0 + \alpha_i^P)}_{\text{intercepts}} + \underbrace{(\beta_1 + \alpha_i^{P\times C})}_{\text{slopes}} c_{ik} + \varepsilon_{ik}. \tag{2}$$

Cast this way, we have a linear model with a single predictor variable, c_{ik}, specifying varying intercepts and slopes over and above their fixed (or average) components.

[1]Other values could be used for the contrast codes (e.g., +0.5 and −0.5), but the key is that they are centered around zero. We assume +1 and −1 throughout. If dummy coding (+1, 0) is used, the meaning of the model's variance components dramatically changes: the intercept variance becomes the random variance of those scoring zero and not the variance for both groups.

As already specified, the condition effect in the above model is captured by β_1, which equals $(\mu_1 - \mu_2)/2$. Cohen (1988) defined the general standardized effect size d as the raw mean difference divided by the square root of the pooled variance of the observations within the conditions:[2]

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

This full model, with all the random components of variation, is estimable only when each participant is crossed with condition (as in the C design) and when there are multiple replicates (i.e., multiple independent observations from each participant in each condition). In the C design with only one replicate (i.e., one observation from each participant in each condition) and in the N design, one can still estimate the fixed effects, but there is a confounding of the three random variance components. Although we do not consider in detail designs with multiple replicates (although see the **Supplemental Appendix**; follow the **Supplemental Material** link in the online version of this article or at **http://www.annualreviews.org/**), we want to provide the details of how one would estimate the full model from such a design if one had such data available. The specifications for the C and N designs become a simple matter of trimming from the full model those variance components that cannot be estimated in those designs.

Supplemental Material

One important issue in estimating the mixed model is the structure of the data file. In the typical ANOVA approach to data, each participant has one row of data in the data file. For the mixed model estimation, each row of data consists of a single observation taken from a particular participant in a particular condition. For instance, if a given participant were to be observed in both conditions with three replicates in each, then that participant would have six lines of data in the data file.

The code for estimating the mixed model specified above for these data is as follows:

```
SAS:    proc mixed;
        class participant;
        model y = c;
        random intercept c/sub = participant type = un;
        run;
SPSS:   mixed y with c
        /fixed = c
        /print = solution testcov
        /random = intercept c | subject(participant) covtype(un).
R:      model <- lmer(y ~ c + (c|participant))
```

In each case, the fixed effects are specified in the mixed model, modeling the observations as a function of condition. Implicit in the model specification are the intercept and the residual at the level of the individual observation. The code specifies the random components of variance, indicating that both the intercept and the slope for condition are allowed to vary randomly across participants. In the *lme4* package in R, the random components are included by the specification "+ (c|participant)," which indicates that the slope for c (and implicitly the intercept) varies across participants. The "un" option in both SAS and SPSS specifies that the random intercept and slope are allowed to covary, which is implicit in the R code. The resulting output includes the intercept

[2]This formula assumes contrast codes of +1 and −1, an assumption that we make in defining effect sizes throughout. Given this coding, the intercept–slope covariance does not contribute to the expected within-condition variation (Westfall et al. 2014, appendix). Additionally, the c^2 term in the effect sizes given by Westfall et al. (2014) drops out because $c^2 = 1$.

and slope fixed estimates (along with standard errors) and the variances and covariance of the random intercepts and slopes. Assumptions are that the random effects are distributed normally and that the model residuals are independent across observations (e.g., no carryover or lagged effects).

We turn now to the C and N designs with a single replicate. In these designs, as we have said, the same underlying components of variance contribute to the individual observations, but not all of them are estimable.

Mixed-Model Specification of the C Design

In the C design, with each participant in both conditions, the same fixed effects can be estimated. However, not all of the random components are estimable. More specifically, σ_P^2 can be estimated, but $\sigma_{P\times C}^2$ cannot. Although $\sigma_{P\times C}^2$ still contributes to variation in the observations, it cannot be disentangled from the residual error term σ_E^2. Accordingly, in the mixed model code, one simply eliminates the random variance in participant slopes from the specification, as that source of variation is contained in the error variance. Thus, in this design, one estimates only two random variance components, participant intercepts and residual error.

The general effect size for this design is given as

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

The denominator of this effect size contains, as before, all three random sources of variation in the observations, but in this case two of these sources are placed in brackets together to indicate the confounding: What is estimated as residual error also includes the variance due to participant slopes.

The test of the condition effect is based on a *t*-statistic that divides the estimated mean difference between the conditions by its estimated standard error. In this design, the variance components that contribute to the standard error of the treatment difference include the participant slope variance and the residual error variance. The variance attributable to participant intercepts (or means) does not contribute to the standard error of the mean difference between conditions. Cohen (1988) defined what we call the operative effect size as the mean condition difference divided only by those variance components that contribute to its standard error. Accordingly, for the C design, the operative effect size is[3]

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

In any sample of data, the operative effect size is estimated as the mean observed condition difference divided by the square root of the estimated residual error (which contains in it the variance attributable to random participant slopes). The general, rather than operative, effect size is typically reported. We give both to clarify those variance components that do and do not contribute to the standard error of the condition difference.

Mixed-Model Specification of the N Design

In this design—the classic two-group between-subjects design—each participant is observed in only one condition. As a result, the error variance contains all three random components of variance

[3] This operative effect size can also be given as $d_O = d \times \frac{1}{\sqrt{1-\rho_{12}}}$, where ρ_{12} is the correlation between the participants' scores in the two conditions.

(participant intercepts, participant slopes, and residual error). Accordingly, in the mixed-model specification, no random components are estimable except for residual error. In the computer code to estimate and test the model, any reference to random slopes or intercepts due to participants is omitted.

The general effect size for this design is

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

As in the general effect size for the C design, the brackets indicate that the variance due to participant intercepts and participant slopes is now part of the residual error variance. This effect size is estimated as the mean observed condition difference divided by the square root of the estimated residual error variance.

Because variances due to both participant intercepts and participant slopes contribute to the estimated residual error in this design, all three components contribute to the standard error for testing the mean condition difference. Accordingly, in this design the operative effect size is identical to the general effect size.

The mixed-model specifications for the C and N designs yield tests of the condition difference that are identical to the comparable standard ANOVA approaches.[4] The difference lies in the structure of the data and the modeling of the sources of variation in the data. The standard ANOVA approach treats the individual participant as the unit of analysis and does not normally make explicit all sources of variation in the data. In contrast, the mixed-model specification treats the individual observations as the unit of analysis and allows multiple simultaneous sources of random variation in the data. For this reason, the mixed-model approach is appropriate for the analysis of data with multiple random factors.

DESIGNS WITH TWO RANDOM AND ONE FIXED FACTOR

With only one random factor, the design alternatives are limited. With two random factors, the design possibilities grow considerably. The random factors may be crossed with each other, or one may be nested within the levels of the other; each random factor may also be crossed with or nested within the levels of the fixed factor. In this section, we lay out all the design possibilities. We continue to refer to the fixed factor as condition, having two levels. We refer to the two random factors as participants and targets. We assume the goal is to estimate and test the condition difference so that inferences can be made to other samples of participants and targets that might have been used.

We start with the most general design, in which all factors are crossed with each other (every participant responds to every target in both conditions) and in which there are multiple replicates (multiple observations taken from each participant in response to each target in each condition). We refer to this as the most general design because it is only in the context of this design that we can define and estimate all the random components of variance contributing to the observations. Thus, only for this most general design can we give the full mixed-model specification and its associated code for estimation. We then provide a general effect size definition as the magnitude of the condition difference relative to all of the random sources of variation in the data.

We turn next to more widely used designs that do not include multiple replicates and in which, therefore, not all of the variance components are estimable. These include both designs

[4]Technically, this is true only if ρ_{12}, as defined in footnote 3, is nonnegative.

in which the two random factors are crossed with each other and designs in which one random factor is nested within the other. Accordingly, the designs that we consider, and their mixed-model specifications, bridge two rather disparate literatures devoted to linear mixed models. Designs with crossed random factors have been considered primarily by experimental researchers in psychology and linguistics (Baayen et al. 2008, Clark 1973, Judd et al. 2012); designs with nested random factors have a long history in educational psychology and applied statistics, fields in which they are commonly referred to as multilevel or hierarchical linear models (Hox 2010, Raudenbush & Bryk 2002, Snijders & Bosker 2011).

As was the case for the specific designs with participants as the only random factor that were considered in the previous section, these specific designs differ from the most general design in that only some of the variance components from the full set that contributes to the observations can be estimated. For each design, we give those variance components that are estimable and those that are not and then present the code modifications that are necessary for estimation. All designs permit the testing of condition differences with generalization across both participants and targets. For each design, we also give appropriate design-specific effect sizes.

Mixed-Model Specification and Effect Size for the Most General Design

In this section, we present the full mixed-model specification for designs with the two random factors of participants and targets and the fixed factor of condition having two levels. As discussed in the preceding paragraphs, this full specification requires a design in which all three factors are fully crossed and in which there are multiple replicates. This is the most general design in the sense that only in this design can all of the underlying variance components be estimated. All other designs involving these factors represent modifications of this design in which some of the observations are systematically missing and, accordingly, in which some of the variance components are confounded with each other.

We assume a single dependent variable with variation accruing from a condition difference; a series of random effects attributable to the underlying factors in the design; and, additionally, random error. The full mixed model for the response of the i^{th} participant to the j^{th} target in the k^{th} condition is

$$Y_{ijk} = \beta_0 + \beta_1 c_{ijk} + \alpha_i^P + \alpha_i^{P\times C} c_{ijk} + \alpha_j^T + \alpha_j^{T\times C} c_{ijk} + \alpha_{ij}^{P\times T} + \alpha_{ij}^{P\times T\times C} c_{ijk} + \varepsilon_{ijk} \qquad (3)$$

and the following are the sources of variation in Y_{ijk}:

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

$$\text{var}\left(\alpha_j^T\right) = \sigma_T^2, \quad \text{var}\left(\alpha_j^{T\times C}\right) = \sigma_{T\times C}^2, \quad \text{cov}\left(\alpha_j^T,\ \alpha_j^{T\times C}\right) = \sigma_{T,T\times C},$$

$$\text{var}\left(\alpha_{ij}^{P\times T}\right) = \sigma_{P\times T}^2, \quad \text{var}\left(\alpha_{ij}^{P\times T\times C}\right) = \sigma_{P\times T\times C}^2, \quad \text{cov}\left(\alpha_{ij}^{P\times T},\ \alpha_{ij}^{P\times T\times C}\right) = \sigma_{P\times T,P\times T\times C},$$

$$\text{var}(\varepsilon_{ijk}) = \sigma_E^2.$$

As above, β_0 and β_1 in this model represent the fixed effects and capture, respectively, the overall mean response and the condition difference in responses. The other elements in this model are the random effects, the variances and covariances of which are given intuitive interpretations in **Table 1**. To show more clearly the specification of some of these components as random intercept

Table 1 Definitions of random variance and covariance components in the designs considered in this review

Variance or covariance component	Definition
σ^2_P	Participant intercept variance: the extent to which participants have different mean responses
$\sigma^2_{P\times C}$	Participant slope variance: the extent to which the mean difference between conditions varies across participants
$\sigma_{P,P\times C}$	Participant intercept–slope covariance: the extent to which participants with larger mean responses also show larger condition differences
σ^2_T	Target intercept variance: the extent to which targets elicit different mean responses
$\sigma^2_{T\times C}$	Target slope variance: the extent to which the mean difference between conditions varies across targets
$\sigma_{T,T\times C}$	Target intercept–slope covariance: the extent to which targets that elicit larger mean responses also show larger condition differences
$\sigma^2_{P\times T}$	Participant-by-target intercept variance: the extent to which participants show stable, unique patterns of mean responses toward particular targets
$\sigma^2_{P\times T\times C}$	Participant-by-target slope variance: the extent to which the mean difference between conditions varies across different participant–target pairs
$\sigma_{P\times T,P\times T\times C}$	Participant-by-target intercept–slope covariance: the extent to which participant–target pairs with larger mean responses also show larger condition differences
σ^2_E	Residual error variance: the extent to which there is variation in responses not due to the above sources

Variables: *C*, fixed condition factor; *E*, error; *P*, random participant factor; *T*, random target factor.

components and others as random slope components, we rewrite the mixed model of Equation 3 as

$$Y_{ijk} = \underbrace{(\beta_0 + \alpha_i^P + \alpha_j^T + \alpha_{ij}^{P\times T})}_{\text{intercepts}} + \underbrace{(\beta_1 + \alpha_i^{P\times C} + \alpha_j^{T\times C} + \alpha_{ij}^{P\times T\times C})}_{\text{slopes}} c_{ijk} + \varepsilon_{ijk}. \quad (4)$$

On the basis of this model and again using Cohen's (1988) specification of the effect size, the following can be defined as the general effect size for this design:[5]

$$d = \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2_P + \sigma^2_{P\times C} + \sigma^2_T + \sigma^2_{T\times C} + \sigma^2_{P\times T} + \sigma^2_{P\times T\times C} + \sigma^2_E}}. \quad (5)$$

For mixed-model estimation, the data file is again structured so that each individual observation is a row of data. The code for estimating effects for data from this fully crossed design with multiple replicates is as follows:

```
SAS:    proc mixed;
        class participant target;
        model Y = c;
        random intercept c/sub = participant type = un;
        random intercept c/sub = target type = un;
        random intercept c/sub = participant*target type = un;
        run;
```

[5]Consistent with footnote 2, we assume contrast code values of +1 and −1. Other values require a slight modification of the effect size denominator. Under the contrast-coding convention, the intercept–slope covariances do not contribute to the expected within-condition variation (Westfall et al. 2014, appendix).

```
SPSS:   mixed y with c
        /fixed = c
        /print = solution testcov
        /random = intercept c | subject(participant) covtype(un)
        /random = intercept c | subject(target) covtype(un)
        /random = intercept c | subject(participant*target) covtype(un).
R:      model <- lmer(y ~ c + (c|participant) + (c|target)
         + (c|participant:target))
```

This code is an extension of the code given above for designs with one random factor (see Mixed Models for Designs with One Random and One Fixed Factor); in this case, it provides for the additional random components of variance in the design: random intercepts and slopes due to participants, those due to targets, and finally those due to the participant-by-target interaction.

As before, not all variance components contribute to the standard error used to test the condition difference in this design. Accordingly, the operative effect size for this fully crossed design with replicates, calculated by dividing the condition mean difference by those components that contribute to its standard error, is:

$$d_0 = \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2_{P\times C} + \sigma^2_{T\times C} + \sigma^2_{P\times T\times C} + \sigma^2_E}}. \quad (6)$$

In the following sections, we systematically define the possible designs that involve two random factors (participants and targets) and a single fixed factor (condition) but that have only a single replicate. All of the designs that we define can be seen as special cases of the most general design considered above but with systematically missing observations. Each design provides an estimate of the fixed effects of interest. In all the designs, the same variance components potentially contribute to the observations, but some of these components are confounded with each other, and thus model specification and effect sizes must be tailored to each particular design.

Design Possibilities

To define the full range of designs that have the three factors of condition, participants, and targets, we must consider the three possible pairs of these factors: condition and participants, condition and targets, and participants and targets. For each pair, the two factors may be crossed or nested. We use C and N to indicate whether the factors in each pair are crossed or nested, respectively. Each design is thus identified by three letters: the first C or N indicates whether participants are crossed with condition or are nested within condition; the second C or N indicates whether targets are crossed with condition or nested within condition; and, finally, the third letter defines whether the two random factors, participants and targets, are themselves crossed or nested. When the two random factors are nested, there are two possibilities: Either targets are nested within participants (meaning that each participant responds to a unique set of targets), or participants are nested within targets (meaning that each target is responded to by a unique set of participants). In the first case, the final letter in the definition of each design is N_P, meaning that participants are the higher-level factor within which targets are nested, and in the second case, in which participants are nested within targets, the final letter in the definition of each design is N_T.

The designs are listed in **Table 2**; each design is identified by the labels defined above. We now further define and illustrate each of these designs. We start with the first column of the table, in which the two random factors, participants and targets, are crossed with each other.

Table 2 Typology of designs with two random factors [participants (P) and targets (T)] and one fixed factor (condition)

How are P and T related to condition?	How are P and T related to each other?		
	P and T crossed	T nested in P	P nested in T
P and T crossed with condition	CCC	CCN_P	CCN_T
P crossed with condition; T nested in condition	CNC	CNN_P	Impossible
P nested in condition; T crossed with condition	NCC	Impossible	NCN_T
P and T nested in condition	NNC	NNN_P	NNN_T

Designs with crossed random factors. The four cells in the first column of **Table 2** define four designs with the final letter C. These designs are illustrated in **Figure 1***a*. For ease of depiction, in these matrices the numbers of levels of the random factors are considerably smaller than what they would likely be in any actual study.

The first design, CCC, is the fully crossed design in which every participant responds to every target twice, once in each condition. Imagine a design in which participants make speeded categorization judgments of two computer-altered versions of a set of male faces, one version morphed towards a prototypic White face and the other morphed towards a prototypic Black face. Faces constitute the random target factor. The condition variable refers to the two morphing conditions. Every participant categorizes every face in both its White-morphed version and its Black-morphed version (i.e., condition).

The CNC design is one in which every target is responded to by every participant, but each target is in only one condition. Imagine a variation of the previous example in which participants judge target faces of different races, but in this case the faces judged are actual faces of White and Black individuals rather than morphed versions of faces. Thus, each individual face is either White or Black, so targets are nested within condition. Participants make a speeded categorization judgment of every face, half of which are White and half Black.

In the NCC design, participants are nested within condition and targets are crossed with condition. Imagine that participants complete a series of target judgments either under cognitive load or without such load. Each participant is in only one of the two load conditions. However, every target is judged under both load conditions, albeit by different participants.

In the NNC design, both random factors are nested within condition. Imagine that participants make career likelihood judgments of faces (e.g., "How likely is it that this person is a scientist?"). There are two sets of faces, either male or female, and participants respond only to one set or the other. Gender of target is the condition variable of interest.

Designs with nested random factors. Designs in the second and third columns of **Table 2** have one of the two random factors nested within the other. In the second column, in which targets are nested within participants, each participant has a unique set of targets. In the third column, each target is responded to by its own unique set of participants. These designs are illustrated by the matrices in **Figure 1***b*.

The CCN_P and CCN_T designs have one random factor nested within the other, but both of these factors are crossed with condition. The classic nested design from educational research involves students who are nested in classrooms, each taught by a single instructor. In one version of this design, the CCN_P design, the instructors evaluate their students in two different conditions or subjects, math and language. Thus, the instructors are the participants and they evaluate their students (targets) on both subjects. The question is whether the evaluations depend on the subject matter. In the CCN_T design, the students are now the participants and they evaluate their instructor as both a math teacher and a language teacher. Thus, students continue to be nested within

a Designs with crossed P and T random factors

CCC

	T					
P	1	2	3	4	5	6
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

CNC

	T					
P	1	2	3	4	5	6
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

NCC

	T					
P	1	2	3	4	5	6
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

NNC

	T					
P	1	2	3	4	5	6
1	A	A	A			
2	A	A	A			
3	A	A	A			
4				B	B	B
5				B	B	B
6				B	B	B

b Designs with nested P and T random factors

CCN_P

	T					
P	1	2	3	4	5	6
1	AB	AB				
2			AB	AB		
3					AB	AB

CCN_T

	T		
P	1	2	3
1	AB		
2	AB		
3		AB	
4		AB	
5			AB
6			AB

CNN_P

	T							
P	1	2	3	4	5	6	7	8
1	A	A	B	B				
2					A	A	B	B

NCN_T

	T	
P	1	2
1	A	
2	A	
3	B	
4	B	
5		A
6		A
7		B
8		B

NNN_P

	T							
P	1	2	3	4	5	6	7	8
1	A	A						
2			A	A				
3					B	B		
4							B	B

NNN_T

	T			
P	1	2	3	4
1	A			
2	A			
3		A		
4		A		
5			B	
6			B	
7				B
8				B

c Other designs (also with crossed P and T random factors)

Counterbalanced

	T					
P	1	2	3	4	5	6
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

Replication designs

R(CCC)

Replication 1

	T			
P	1	2	3	4
1	AB	AB	AB	AB
2	AB	AB	AB	AB
3	AB	AB	AB	AB
4	AB	AB	AB	AB

Replication 2

	T			
P	5	6	7	8
5	AB	AB	AB	AB
6	AB	AB	AB	AB
7	AB	AB	AB	AB
8	AB	AB	AB	AB

R(NCC)

Replication 1

	T			
P	1	2	3	4
1	A	A	A	A
2	A	A	A	A
3	B	B	B	B
4	B	B	B	B

Replication 2

	T			
P	5	6	7	8
5	A	A	A	A
6	A	A	A	A
7	B	B	B	B
8	B	B	B	B

R(CNC)

Replication 1

	T			
P	1	2	3	4
1	A	A	B	B
2	A	A	B	B
3	A	A	B	B
4	A	A	B	B

Replication 2

	T			
P	5	6	7	8
5	A	A	B	B
6	A	A	B	B
7	A	A	B	B
8	A	A	B	B

R(NNC)

Replication 1

	T			
P	1	2	3	4
1	A	A		
2	A	A		
3			B	B
4			B	B

Replication 2

	T			
P	5	6	7	8
5	A	A		
6	A	A		
7			B	B
8			B	B

Figure 1

Illustrative matrices for all designs having two random factors, participants (P, *rows*) and targets (T, *columns*), and one fixed factor, condition, with two levels (A and B) under which particular observations occur. An empty or blank cell indicates that the observation for a specific combination of participant and target is not collected.

instructors, but these two groups have switched their roles in terms of the design: The instructors elicit responses, and we thus designate them as the targets, whereas the students produce responses, and we thus designate them as the participants. Both random factors are crossed with condition.

In the case of nested random factors where one of the factors is crossed with condition and the other is nested within condition, the higher-order random factor, rather than the lower-order one, must be the factor that is crossed with condition. Accordingly, two of the cells in the second and third columns of **Table 2** define impossible designs. In the CNN_P design, targets are nested within participants, participants are crossed with condition, and targets are nested within condition. Imagine that male participants are each asked to nominate and judge their two closest male and two closest female friends. Thus, each participant has a unique set of targets (friends) who are either male or female (the condition variable). The question is whether participants give systematically different ratings to their nominated male friends than their female friends.

In the NCN_T design, participants are nested within targets, targets are crossed with condition, and participants are nested within condition. In this case, the four friends nominated by each person are recruited as the participants, and they each rate their common (male) friend, the target. Participants (those who do the rating of their common nominating friend) are now nested within gender (their own), but target (the male nominating person who is rated) is crossed with the friends' genders.

The final two designs of **Table 2** are the fully nested designs, NNN_P and NNN_T, in which either targets are nested within participants or the other way around, and both random factors are nested within condition. As an example of the NNN_P design, imagine that male and female participants are recruited, and they nominate and rate as targets two friends of only their own gender. In this case, targets are nested within participants and both participants and targets are either male or female (condition being gender). For the NNN_T design, again imagine that people nominate their two friends, who are, again, the same gender as the nominating person. However, this time, the nominated friends serve as the participants, and they each rate the person who nominated them (target). Participants are now nested within targets and both are nested within condition (gender).

In addition to these designs, there are two final designs, used with some frequency, in which participants and targets are in fact confounded, with just a single target nested within each participant or, equivalently, a single participant nested within each target. Imagine research in which each participant thinks of a single friend and rates him or her, either in one condition only or in both. Thus participant and target are completely confounded, and both can be either crossed with condition or nested within condition. The analysis of this design is formally equivalent to those with one random factor that we considered above. However, in this case the random factor is not participants but the participant–target pair, and random variation in the data accrues from both sources, as well as their interaction.

Other designs. **Table 2** provides a coherent way of defining the possibilities with two random and one fixed, two-level factor. However, other possibilities deserve discussion. These designs are illustrated by the matrices in **Figure 1*c***.

First, there is a variation on the fully crossed CCC design that we call the counterbalanced design (Westfall et al. 2014). This is a fully crossed design in the sense that all participants are crossed with all targets and every participant and target occurs in both conditions. Unlike the CCC design, however, each participant responds to each target in just one condition. As shown in **Figure 1*c***, participants and targets are divided into two blocks that define the condition under which a specific participant–target pair is observed. In the CCC design, condition, participants, and targets are fully crossed, whereas in the counterbalanced design, condition is confounded with

the participant-by-target interaction. As an example, imagine that participants complete a set of math problems, some while under cognitive load and others without load. Every participant does all problems, but the division of the problems between the half that are done under load and the half that are not varies across participants.

Second, there are four designs that we refer to as replication designs in that they replicate some of the designs of **Table 2** with multiple sets of participants and targets. Above, we talked about designs with multiple replicates (meaning multiple observations from the same participant, target, and condition). We mean something entirely different by replication designs, i.e., that an entire previously defined design is replicated more than once with new sets of participants and targets. Consider, for instance, the first row of **Table 2**, in which both participants and targets are crossed with condition. Suppose that, rather than fully crossing participants and targets, we group participants and targets such that each group contains unique participants and unique targets. Within each group, participants and targets are fully crossed, but there are multiple such groups. This design essentially replicates the CCC design many times, with each group of participants and targets constituting one replication. We refer to this design as R(CCC). Again, a replication is defined as a specific group or subset of participants and targets. In **Figure 1*c*** we have illustrated the R(CCC) design with the number of replications equal to two (and four participants and targets in each replication). As a more extended example, suppose participants are put in groups of four and everyone in a particular group responds to the same four targets twice, once in one condition and once in the second. There might be a total number of 32 participants and 32 targets across the total of eight replications. The advantage of this design over the fully crossed design is that it potentially reduces participant load (i.e., participants do not need to make as many responses) while nevertheless using a large number of targets, which can be important for statistical efficiency reasons considered below (see Power Considerations and Research Design).

The R(CCC) design is the replication design from the first row of **Table 2**. The other three replication designs correspond to the remaining three rows of **Table 2**. These are illustrated in **Figure 1*c***, again with only two replications. The R(CNC) design is the CNC design replicated multiple times with different sets of participants and targets; each target occurs in only one condition or the other. The R(NCC) design is the NCC design replicated many times with different sets of participants and targets, in this case with participants nested within condition and targets crossed with condition. And finally, the R(NNC) design is the NNC design with multiple replications of different sets of participants and targets.

These replication designs, with participants crossed with targets in each replication, become the nested designs of the second and third columns of **Table 2** when the number of either participants or targets in each replication equals one. Thus, for instance, the R(CCC) design becomes the CCN_P design if each replication contains only a single participant, responding to the targets that are unique to that replication; it becomes the CCN_T design if each replication has only a single target responded to by the unique set of participants in that replication. The other replication designs also become the nested designs of the third and fourth columns of **Table 2** when the number of either participants or targets in each replication equals one.

Design-Specific Estimation and Effect Sizes

In this section, we discuss the mixed-model specification that estimates the condition difference given all of the random variance components that are estimable in each design. Recall that there are a total of six variance components (and three covariances) that, along with residual error, contribute to the total variation in observations. These are defined in **Table 1**. In the fully crossed design with multiple replicates (i.e., multiple observations from each participant, target, and condition combination) considered earlier, all of these variance components are estimable. Accordingly, we

gave the mixed-model code in SAS, SPSS, and R; this code specifies how one estimates the fixed effect of condition and the random variance components. Finally, we also gave the effect sizes for this design; the general effect size is defined as the mean condition differences divided by all six variance components plus the residual variance, and the operative effect size is defined as the mean condition difference divided by only those variance components that contribute to the standard error of the condition difference.

In the second column of **Table 3**, we present the general effect sizes for all of the designs that we have defined. (The third column of this table lists the noncentrality parameters, which are necessary for the computation of statistical power and are discussed in the section below devoted to that subject.) Consistent with our earlier treatment of designs that have participants as the only random factor, the confounding of the variance components in these designs is indicated by brackets in the effect sizes. The denominators of the general effect sizes include, for all designs, all six variance components defined in **Table 1** plus random error variance, but many of these are confounded and not separately estimable. The first variance component within a set of brackets indicates the component that is estimable in the mixed-model specification, and the components that follow within the brackets are those that are confounded with the estimable component. The operative effect sizes for all of the designs are given in the **Supplemental Appendix**. All the information necessary for specifying the appropriate mixed model for each design is implicit in its general effect size. In the **Supplemental Appendix**, we give the code (again in SAS, SPSS, and R) for each of the designs, but the specifics of the code follow from the denominators of each design's general effect size in **Table 3**. The rule is that one specifies as random effects those variance components that are not contained in brackets in the denominator of the effect size or those that appear first in a set of bracketed components. (Note that the residual variance is included in the model by default and does not need to be specified explicitly.)

Supplemental Material

In the following paragraphs we provide illustrations for a few of the designs of how one goes from the general effect sizes in **Table 3** to the mixed model code given in the **Supplemental Appendix** for each design. We also briefly discuss for each design the estimable components that do not contribute to the standard error of the condition difference, thus highlighting the differences between the general (in **Table 3**) and the operative effect sizes (which are listed in the **Supplemental Appendix**).

The first design we consider is the CCC design. The only use of brackets in the denominator of its general effect size in **Table 3** is $[\sigma^2_E + \sigma^2_{P\times T\times C}]$, which indicates that the variance attributable to the three-way interaction is confounded with the residual error variance. Accordingly, in modifying the code given in the section Mixed-Model Specification and Effect Size for the Most General Design (the most general design being the crossed design with multiple replicates), one includes all random components except random condition slopes for the participant-by-target interaction. Because participants and targets (and their interaction) are crossed with condition, the intercept variances attributable to these three do not contribute to the standard error of the condition difference, although those components are estimable and should be included in the model.

Second, the NCC design has two sets of brackets in the denominator of its effect size. Variance attributable to the participant by condition interaction is confounded with participant variance, and, additionally, both the participant-by-target interaction and the triple interaction are confounded with the residual error variance. Thus, the code must be modified to estimate only random participant intercepts and random target intercepts and slopes. All estimable components contribute to the standard error of the condition difference in this design except that due to target intercepts.

As a third example, the NCN_T design has three estimable components, those due to target variance, target by condition variance, and residual variance. Thus, the code specifies only those

Table 3 General effect sizes (d) and noncentrality parameters for the designs of Table 2 and Figure 1[a]

Designs	d	Noncentrality parameter
CCC	$\frac{\mu_1-\mu_2}{\sqrt{\sigma^2_P+\sigma^2_{P\times C}+\sigma^2_T+\sigma^2_{T\times C}+\sigma^2_{P\times T}+\left[\sigma^2_E+\sigma^2_{P\times T\times C}\right]}}$	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{\left[\sigma^2_E+\sigma^2_{P\times T\times C}\right]}{2pq}+\frac{\sigma^2_{P\times C}}{p}+\frac{\sigma^2_{T\times C}}{q}}}$
CNC	$\frac{\mu_1-\mu_2}{\sqrt{\sigma^2_P+\sigma^2_{P\times C}+\left[\sigma^2_T+\sigma^2_{T\times C}\right]+\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}\right]}}$	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}\right]}{pq}+\frac{\sigma^2_{P\times C}}{p}+\frac{\left[\sigma^2_T+\sigma^2_{T\times C}\right]}{q}}}$
NCC	$\frac{\mu_1-\mu_2}{\sqrt{\left[\sigma^2_P+\sigma^2_{P\times C}\right]+\sigma^2_T+\sigma^2_{T\times C}+\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}\right]}}$	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}\right]}{pq}+\frac{\left[\sigma^2_P+\sigma^2_{P\times C}\right]}{p}+\frac{\sigma^2_{T\times C}}{q}}}$
NNC	$\frac{\mu_1-\mu_2}{\sqrt{\left[\sigma^2_P+\sigma^2_{P\times C}\right]+\left[\sigma^2_T+\sigma^2_{T\times C}\right]+\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}\right]}}$	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{2\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}\right]}{pq}+\frac{\left[\sigma^2_P+\sigma^2_{P\times C}\right]}{p}+\frac{\left[\sigma^2_T+\sigma^2_{T\times C}\right]}{q}}}$
CCN_P	$\frac{\mu_1-\mu_2}{\sqrt{\sigma^2_P+\sigma^2_{P\times C}+\left[\sigma^2_T+\sigma^2_{P\times T}\right]+\left[\sigma^2_E+\sigma^2_{T\times C}+\sigma^2_{P\times T\times C}\right]}}$	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{\left[\sigma^2_E+\sigma^2_{P\times T\times C}+\sigma^2_{T\times C}\right]}{2q}+\frac{\sigma^2_{P\times C}}{p}}}$
CCN_T	$\frac{\mu_1-\mu_2}{\sqrt{\left[\sigma^2_P+\sigma^2_{P\times T}\right]+\sigma^2_T+\sigma^2_{T\times C}+\left[\sigma^2_E+\sigma^2_{P\times C}+\sigma^2_{P\times T\times C}\right]}}$	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{\left[\sigma^2_E+\sigma^2_{P\times T\times C}+\sigma^2_{P\times C}\right]}{2p}+\frac{\sigma^2_{T\times C}}{q}}}$
CNN_P	$\frac{\mu_1-\mu_2}{\sqrt{\sigma^2_P+\sigma^2_{P\times C}+\left[\sigma^2_E+\sigma^2_T+\sigma^2_{P\times T}+\sigma^2_{T\times C}+\sigma^2_{P\times T\times C}\right]}}$	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}+\sigma^2_T+\sigma^2_{T\times C}\right]}{q}+\frac{\sigma^2_{P\times C}}{p}}}$
NCN_T	$\frac{\mu_1-\mu_2}{\sqrt{\sigma^2_T+\sigma^2_{T\times C}+\left[\sigma^2_E+\sigma^2_P+\sigma^2_{P\times T}+\sigma^2_{P\times C}+\sigma^2_{P\times T\times C}\right]}}$	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}+\sigma^2_P+\sigma^2_{P\times C}\right]}{p}+\frac{\sigma^2_{T\times C}}{q}}}$
NNN_P	$\frac{\mu_1-\mu_2}{\sqrt{\left[\sigma^2_P+\sigma^2_{P\times C}\right]+\left[\sigma^2_E+\sigma^2_T+\sigma^2_{P\times T}+\sigma^2_{T\times C}+\sigma^2_{P\times T\times C}\right]}}$	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}+\sigma^2_T+\sigma^2_{T\times C}\right]}{q}+\frac{\left[\sigma^2_P+\sigma^2_{P\times C}\right]}{p}}}$
NNN_T	$\frac{\mu_1-\mu_2}{\sqrt{\left[\sigma^2_T+\sigma^2_{T\times C}\right]+\left[\sigma^2_E+\sigma^2_P+\sigma^2_{P\times T}+\sigma^2_{P\times C}+\sigma^2_{P\times T\times C}\right]}}$	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}+\sigma^2_P+\sigma^2_{P\times C}\right]}{p}+\frac{\left[\sigma^2_T+\sigma^2_{T\times C}\right]}{q}}}$
Counterbalanced	$\frac{\mu_1-\mu_2}{\sqrt{\sigma^2_P+\sigma^2_{P\times C}+\sigma^2_T+\sigma^2_{T\times C}+\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}\right]}}$	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}\right]}{pq}+\frac{\sigma^2_{P\times C}}{p}+\frac{\sigma^2_{T\times C}}{q}}}$
R(CCC)	Same as CCC	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{r\left[\sigma^2_E+\sigma^2_{P\times T\times C}\right]}{2pq}+\frac{\sigma^2_{P\times C}}{p}+\frac{\sigma^2_{T\times C}}{q}}}$
R(CNC)	Same as CNC	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{r\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}\right]}{pq}+\frac{\sigma^2_{P\times C}}{p}+\frac{\left[\sigma^2_T+\sigma^2_{T\times C}\right]}{q}}}$
R(NCC)	Same as NCC	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{r\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}\right]}{pq}+\frac{\left[\sigma^2_P+\sigma^2_{P\times C}\right]}{p}+\frac{\sigma^2_{T\times C}}{q}}}$
R(NNC)	Same as NNC	$\frac{\mu_1-\mu_2}{2\sqrt{\frac{2r\left[\sigma^2_E+\sigma^2_{P\times T}+\sigma^2_{P\times T\times C}\right]}{pq}+\frac{\left[\sigma^2_P+\sigma^2_{P\times C}\right]}{p}+\frac{\left[\sigma^2_T+\sigma^2_{T\times C}\right]}{q}}}$

[a]Brackets indicate the confounding of variance components. All variance components in the noncentrality parameters are defined in **Table 1**. The number of participants is p and the number of targets is q. In replication designs, the number of replications is r.

random components due to target intercepts and slopes, in addition to the implicit residual error term. Target intercept variance, although estimable, does not contribute to the standard error of the condition difference.

At the bottom of the first column of **Table 3**, we indicate that the effect sizes for the four replication designs are identical to those of the parallel designs in which participants and targets are crossed but the design is replicated only once (i.e., all participants get all targets). Thus, for instance, the effect size for the R(CCC) design is identical to that given for the CCC design. The syntax for these designs is also the same as that for the parallel designs with a single replication, although it probably makes sense to include replications in the model as an additional fixed factor (along with the fixed condition-by-replications interaction).

We end this section with a final warning on model specification. We have seen published analyses of designs with crossed random factors of participants and targets that use a mixed model specified as if it were a nested design in which targets are nested within participants (e.g., Toma et al. 2012). Typically in diary studies, for instance, days of measurement are crossed with participants but are treated as nested within them. We suspect this happens because so-called hierarchical or multilevel models have been used in the literature for some time, whereas models for crossed random factors are a more recent development. The misspecification of a crossed design as a nested one essentially amounts to ignoring the random variation in the target factor and thus risks serious inflation of type I errors if in fact there is nonzero target variance (Judd et al. 2012). The lesson is that model specification should follow from the design.

STATISTICAL POWER FOR DESIGNS WITH TWO RANDOM FACTORS

Supplemental Material

In this section and in the **Supplemental Appendix**, we provide the tools necessary to estimate statistical power for the test of the condition difference for each of the designs that we have covered. We discuss the general approach to power estimation and then provide a web-based application that computes power for all the designs. In earlier work (Westfall et al. 2014), we developed a power application for those designs that involve two crossed random effects. The current application (located online at **http://jakewestfall.org/two_factor_power/**) extends the range of designs treated to all those defined in this review, having both crossed and nested random effects, as well as the replication designs.

Our approach to statistical power estimation is consistent with the general approach laid out by Cohen (1988). One begins by specifying both a null hypothesis of no condition difference and an alternative hypothesis given an anticipated condition difference of some magnitude. Power is defined as the probability of correctly rejecting the null hypothesis when the alternative hypothesis is correct. To compute power, one must specify those variance components that contribute to the standard error of the condition difference for each design; these are given in the denominators of the design-specific operative effect sizes (see the **Supplemental Appendix**). These variance components are then weighted appropriately by the sample sizes involved in the prospective study [total numbers of participants (p) and targets (q) in the design and number of replications (r) in the replication designs] to give what is called the noncentrality parameter for the hypothesized true effect, which is presented in the third column of **Table 3** for each design. One can think of the noncentrality parameter as approximating the average t-statistic that one would obtain if the expected condition difference were found. The denominator of the noncentrality parameter can be thought of as the expected standard error of the condition mean difference. When squared and multiplied by the total number of observations, the denominator is equal to what the ANOVA literature refers to as the expected mean square (EMS) for the condition factor (Winer 1971).

Given degrees of freedom for this noncentrality parameter, power can be computed by examining areas under the noncentral t-distribution, given the multivariate normality assumption of the underlying random effects' distributions. Because the noncentrality parameters pool or combine various relevant variance components, the degrees of freedom of the noncentral t must be approximated. We use the Satterthwaite approximation to estimate the relevant degrees of freedom (Satterthwaite 1946, Welch 1947). Expressions for the approximate degrees of freedom for each design are given in the **Supplemental Appendix**.

We provide a web-based application (**http://jakewestfall.org/two_factor_power/**) that computes power for these designs. The user must identify the specific design used, the numbers

of participants and targets,[6] the hypothesized mean difference or effect size, and the relevant variance components. In the application, the user has a choice between two different ways of thinking about the variance components and the effect size. Under the first option, the user inputs the mean difference expected and estimated values for all of the estimable variance components. An often-simpler option is to input what might be thought of as standardized versions of these, including the anticipated effect size (in terms of d) and the relative magnitude of the estimable variance components for each design (the proportion of the total variance in the observations attributable to a particular component of variance). We refer to these relative estimates of the variance components as Variance Partitioning Coefficients (Goldstein et al. 2002), and designate them as V (e.g., V_P for participant intercept variance, $V_{T\times C}$ for target slope variance). By definition, the sum of all the Vs (including residual error) must equal 1.0.

POWER CONSIDERATIONS AND RESEARCH DESIGN

All designs permit an estimate of the condition difference. Therefore, in making a decision about which design to use, the most important considerations are feasibility and statistical power. We discuss the feasibility issues in the section Design Choices: Power and Feasibility Considerations. In this section, we consider those factors influencing the power to detect the anticipated condition difference.

In general, the smaller the variance components that contribute to the noncentrality parameter (or operational effect size) and the larger the relevant sample sizes, the greater the power. In designs with participants as the only random factor, power is determined by the participant variance components and the participant sample size. In the designs that we are now considering, power is determined by the variance components and sample sizes of both participants and targets, although the extent to which these matter varies from design to design. The important point, however, is that we must think in terms of multiple relevant variance components and multiple sample sizes.

To increase power in designs with participants as the only random factor, researchers can either decrease the error variability in the data or increase the number of participants. Both strategies involve costs. The costs associated with increasing the number of participants are obvious. Those associated with decreasing participant variability are less obvious. Selecting participants who are relatively homogeneous on relevant variables related to the outcome should decrease the relevant variance components. Doing so, however, restricts one's ability to generalize observed results to other samples that are not so restricted.

The same considerations hold in thinking about designs with multiple random factors, in which variance components due to targets and their sample size, in addition to those due to participants, figure prominently in determining power. Power often dramatically increases as the number of targets in a design increases. Additionally, if we restrict the variance attributable to targets through pretesting, removing extraneous factors, and other strategies, then power should increase. For instance, it is common in research on face perception to edit target faces to eliminate facial hair and other idiosyncrasies. However, restricting target variance imposes a cost in that one is unable to generalize to samples of targets that are not so restricted. The important point is that the same power considerations apply to the sampling of targets as to the sampling of participants. Larger and more homogeneous samples of both increase power, but these strategies come with costs.

[6]In nested and replication designs, care must be exercised in defining the values of p and q for designs. They represent the total numbers of participants and targets in the entire design.

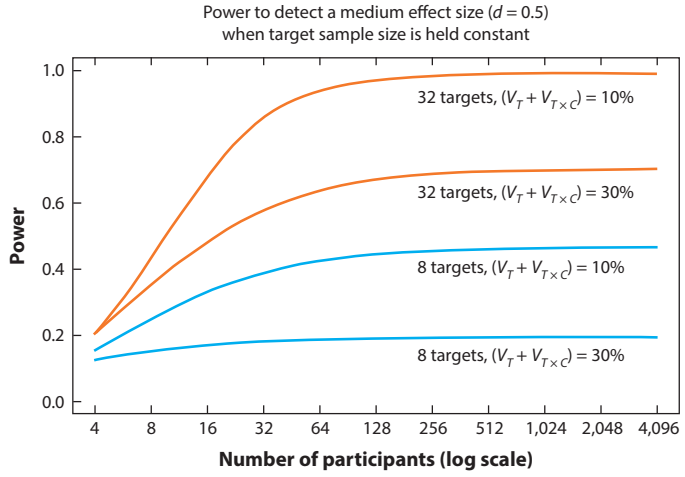

Figure 2

Plot of statistical power as a function of the total number of participants for the CNC (P and T crossed, P crossed with C, and T nested in C) design. The number of targets has been set to either 8 or 32. The other variance components are set to the following values: $V_E = 0.3$, $V_P \times T = 0.1$, $V_P \times T \times C = 0$, $V_P = 0.2$, and $V_P \times C = 0.1$. Note that these other variance components affect only the rate at which the power functions converge to their asymptotes; they do not affect the maximum attainable power values, which depend only on the effect size, number of targets, and the target variance components. Abbreviations: C, fixed condition factor; E, error; P, random participant factor; T, random target factor.

Increasing the Sample Sizes

In designs with just one random factor, as the participant sample size increases, power eventually approaches one. However, in many of the designs considered in this review, if the sample size of one of the two random factors is fixed, then increasing the sample size of the other random factor increases power, but generally to a limit of less than one. This is a surprising result; in our experience, many researchers naturally assume that increasing the number of participants in a design will eventually ensure adequate power. When the number of targets in a design is small, power will increase as the participant sample size increases but may asymptote at levels that are quite a bit lower than one. In many designs that involve both participants and targets, the number of targets used is typically substantially smaller than the number of participants [e.g., a meta-analysis by Bond & DePaulo (2008) in one research domain reports on average 80 participants per study but only 12 targets], as researchers may mistakenly think that power is determined only (or primarily) by the sample size of participants.

To illustrate, in **Figure 2** we plot the power to detect a medium effect size of $d = 0.5$ in the CNC design as a function of number of participants under different assumptions about the sample of targets. Varied in the figure are number of targets, 8 or 32, and how much variance is due to targets and their interaction with condition, 10% or 30% of the total variance. When the number of targets is small or when the targets are highly variable, the maximum attainable power in the study can be far less than one.

A similar phenomenon occurs in nested designs (targets within participants or participants within targets), but only for the lower-level factor. For example, if targets are nested within participants so that each participant responds to a unique set of targets, then the maximum attainable power is less than one if the participant sample size is held constant and the target sample size is increased, but power does approach one if the participant sample is increased. The reason for this

asymmetry is that increasing the sample size of the higher-level factor necessarily entails increasing the sample size of the lower-level factor, but the reverse is not true.

Assuming that one is able to vary either the participant sample size or the target sample size (or both), which is expected to have a greater effect on statistical power? The answer depends on several factors, including the initial sample sizes, the relative sizes of the participant and target variance components, and the design of the experiment. The definitive answer is contained implicitly in the noncentrality parameters we have given. In the following paragraphs, we offer a few rules of thumb.

First, assuming crossed random factors, the larger the variance components associated with one random factor (relative to the other random factor), the more beneficial it is to increase the sample size of that factor. Thus, if participant mean and slope variances are larger than those due to targets, then increasing the sample size of participants reaps greater benefits than increasing that of targets.

Second, if the sample sizes of targets and participants are substantially different, then there will generally be a greater power benefit to increasing the size of whichever sample is smaller, assuming that both participants and targets have approximately equal associated variance components. For instance, if one sample size is 300 units and the other is 10, then adding an additional 10 units to the larger sample size (for a new total of 310) is unlikely to have as big an effect on statistical power as adding an additional 10 units to the smaller sample size (for a new total of 20).

Third, all else being equal, it is better to increase the sample size of a random factor that is nested within condition than one that is crossed with condition. The reason for this is that when a random factor is nested within condition, the standard error of the condition difference depends on both the intercept and slope variance components of that factor, whereas when the random factor is crossed with condition, only the random slope variance is relevant.

Fourth, in a design in which one random factor is nested within the other (e.g., targets within participants), it is usually more effective to increase the sample size of the higher-level factor (e.g., participants) than that of the factor nested within it (e.g., targets). As discussed above, the maximum attainable power level when increasing the lower-level sample size in a nested design is, in general, less than 1.0. According to this rule, even in smaller studies that do not approach the theoretical maximum power level, power increases more quickly by increasing the higher-level sample size than by increasing the lower-level sample size.

Design Choices: Power and Feasibility Considerations

Power is not the only consideration guiding the choice of design; feasibility issues also figure prominently. We discuss some of those issues in this section.

Although power will often increase dramatically as the target sample size increases, sometimes it is not feasible for participants to respond to a large number of targets. In this case, the best strategy may be to use a design with nested, rather than crossed, random factors. Researchers often assume that a crossed design is more powerful, but in fact it can be shown that for any crossed design, a nested version of the same design—that is, one with the same number of responses per participant and the same relationships between the random factors and condition, but in which every participant receives different targets—is always more powerful. This difference derives simply from the fact that as we move to the nested design there is a dramatic increase in the number of targets even as the number of responses per participant remains constant.

However, nested designs may in some contexts require unreasonable numbers of targets (the number of participants times the number of responses per participant). Consider the CCC design and the CCN_P design. If each of 30 participants gives 15 responses, then the CCC design involves

responses to only 15 targets, whereas the CCN_P design involves responses to 450 targets, resulting in a potentially dramatic increase in power if random variance due to targets is present. However, it may simply not be feasible to find so many targets. A reasonable alternative is to consider the R(CCC) design, containing, for instance, three replications of the CCC design, with 10 participants and 15 targets in each, for a total of 45 targets. Each participant still responds only 15 times, but the total number of targets has gone up threefold over the number in the CCC design. Generally speaking, in cases in which each response imposes a considerable burden on participants, it makes sense to increase power by increasing the number of targets across different replications of a design, rather than to limit the number of targets by the use of a design in which all participants respond to the same set of targets (Westfall et al. 2014).

If one has a choice between crossing a random factor with condition and nesting that random factor within condition, then one should always choose to cross the random factor with condition to maximize statistical power. This is because intercept variance due to a random factor contributes to the noncentrality parameter (making it smaller) when that factor is nested within condition, but not when it is crossed with condition. It follows that, if only one random factor is to be crossed with condition, then it is generally better to cross whichever factor has the larger anticipated variance components. Of course, there are feasibility issues that arise in considering whether a random factor can be crossed with condition. Crossing participants with condition raises issues of order and carryover effects, as well as the potential suspicion that participants may develop about the study's purpose. These issues do not arise if the crossed factor is targets.

Finally, if one is using a design with nested random factors and one has the choice of which is the higher-order and which the lower-order factor, then it is always better to choose as the higher-order factor the one with less variance. So, for instance, if participants have larger associated variance components than targets, then a design that nests participants within targets is preferable to one that nests targets within participants. This is true so long as it is feasible to have a reasonable sample size of the higher-level factor.

COMPLICATIONS AND EXTENSIONS

Our designs have assumed only two random factors and one fixed factor having only two levels. We have also assumed that when one factor is nested within another, the nesting is randomly determined. What happens when we go beyond these assumptions? We first discuss the issue of nonrandom nesting and then turn to design extensions.

Nonrandom Nesting

When a random factor is nested within condition, differences attributable to that random factor are confounded with condition differences. With random assignment of levels of the random factor to condition, that confounding can be estimated and dealt with, which is not possible with nonrandom assignment. This is as true of targets as it is of participants. Hence, nonrandom assignment of either participants or targets to condition generally results in bias in estimating condition effects.

When one of the two random factors is nested within the other, either targets within participants or participants within targets, we have assumed random assignment of the nested factor to the levels of the higher-order factor. What happens when this is not the case? For simplicity, we rely on the situation in which targets are nested within participants, but the following considerations apply under the reverse nesting as well. With nonrandom nesting of targets within participants, target differences are confounded with participant differences, resulting in covariances between participant and target intercepts and (perhaps) condition slopes. In many situations, it is likely that

such covariances are positive (participants with higher means respond to targets that, on average, have associated higher means). This positive covariance augments the variance components of participants, which generally results in less efficient tests of condition differences. Thus, a power consequence of nonrandom assignment of targets to participants is likely. However, nonrandom assignment of targets to participants does not result in bias in the estimate of the condition difference, so long as participants are crossed with condition or, when they are nested within condition, they have been randomly assigned to condition. In other words, in a fully nested design, nonrandom assignment of the lower-order random factor to levels of the higher-order one (resulting in a nonzero covariance between the random participant and target effects) does not bias the condition difference estimate so long as the higher-order random factor is still randomly assigned to condition levels.

In the replication designs, participants and targets are nested within each replication. We have assumed random assignment of both factors to each replication. If this is not the case, for instance when unique sets of participants and targets come to the experiment intact and constitute the replications, then the replications often should be treated as an additional random factor in the design specification rather than as a fixed factor, as we suggested when we discussed the model specification for these replication designs. In consequence, one must have sufficient numbers of replications, because the sample size of this factor now becomes relevant in determining power.

Design Extensions

As discussed in the previous section, replications in the replication designs should be considered a random factor in the case of nonrandom nesting within replications. Other contexts also require more than two random factors. For instance, priming studies present participants with primes and ask them to respond to subsequently presented targets. In most cases, one should treat both primes and targets, in addition to participants, as random factors. Using principles extracted from what we have covered in this review, we can extend the designs and models to cover these scenarios. Additional random factors lead to additional complexities in specifying the random components of the underlying model, as the number of variance components can increase exponentially if the random factors are all crossed with each other (as in the priming example). Condition slope variance components must be specified in the model for those random factors that are crossed with condition (and their interactions). Accordingly, though the complete model specification is possible, it may be necessary to specify a large number of variance components, leading to possible convergence problems in estimating the parameters of the linear mixed model. The recent literature has recommended specifying the complete underlying model (including all random variances and covariances that are estimable; see Barr et al. 2013), a recommendation with which we generally concur. At the same time, if convergence cannot be achieved in estimating the model, respecification may help by dropping some of the variance components that represent higher-order interactions that might reasonably be expected to be nonexistent (see also Bates et al. 2015).

Our designs also have only one fixed factor, condition, with only two levels. In many experiments there are more fixed factors, generally crossed and often with more than two levels. So long as those fixed factors (and their interactions) are contrast-coded and the effects of those contrasts tested as single-degree-of-freedom tests, additional fixed factors present no further problems other than, again, the complexity of the model to be specified. Intercept and slope variance components must be specified for any random factors that are crossed with those additional fixed factors (and with the interactions of fixed factors).

There may also be continuous covariates that one would like to include in the fixed part of the model. We strongly recommend centering such covariates (Judd et al. 2008). Doing this

is necessary to preserve the meaning of the random variance components and fixed parameter estimates (including the condition difference).

Finally, we have assumed completely balanced research designs with no missing data. Mixed-model estimation can generally be accomplished with missing data. However, to have unbiased estimates, it must be assumed that such missing data are randomly lost, a highly dubious assumption. More detail on missing data is contained in the **Supplemental Appendix**.

Supplemental Material

FROM OUR DESIGNS TO DYADIC DESIGNS

There is extensive literature on what are called dyadic designs (Kenny et al. 2006), in which participants interact with other participants. Dyadic designs are also designs with two random factors. In fact, all of our designs can be viewed as dyadic. This seems most natural when the targets are people; nonetheless, even when the targets are inanimate, each observation involves a dyad or a pair.

One advantage of viewing studies with participants and targets as dyadic designs is that there is an established tradition of quantifying the random sources of variances in the observations from such designs. As we have discussed, having some idea of the relative amount of variance of the different components can be very helpful in planning a study. Designs using the social relations model (SRM) are understood in great detail; the rest of this section describes these designs.

The SRM examines observations taken from actors about partners (Kenny et al. 2006). In the parlance of this review, an actor is a participant and a partner is a target. In the SRM, the variance in the observations is partitioned into actor, partner, and relationship, i.e., actor × partner interaction. In most applications that use the SRM, there is no fixed variable such as condition, so variances due to condition slopes are not present. The traditional focus in an SRM analysis is on the partitioning of variance into actor, partner, and relationship. For instance, Hönekopp (2006) had participants in three studies judge the physical attractiveness of targets' faces using photographs. In the second study, which had 31 actors and 60 targets, he found that 15% of the variance was due to the participant or actor, 26% due to the target or partner, 33% due to the relationship or participant × target interaction, and the remaining 26% due to error. Quite clearly in this study, the two key systematic sources of variance were relationships and partners, and knowledge of these variances would prove useful in planning studies on interpersonal attraction.

Although all the designs in this review can be viewed as dyadic designs, only one of the designs is a SRM design: the CCC design, which is called the half-block design in SRM parlance. In this design, there are actors (i.e., participants), each of whom is paired with the same set of partners (i.e., targets). The prototypical SRM design, however, involves a case in which the actors and partners are the same group of people, and each person is paired with every other person. Such a design is called a round robin design and is not considered in this review (see, however, the extensive discussion in Kenny et al. 2006).

Many dyadic designs, like the round robin design, are reciprocal in the sense that observations accrue from both participant A responding to target B and from participant B responding to target A. In other words, each person serves as both a participant and a target. The CCC design would be a reciprocal design in the following situation: Consider a study in which a sample of men interact with a sample of women; in each interaction, each person states how much he or she likes his or her interaction partner. If we were interested in the effect of gender, then we could view the study as a fully crossed design in which, for the male judge condition, men are the participants and women are the targets, and for the female judge condition, the same women are now the participants and the same men are the targets. This SRM design is referred to as the asymmetric block design.

Other designs that we have considered can be made into reciprocal designs by combining them. For instance, in the example we used for the CCN_P design, the instructors evaluate their students in two different conditions or subjects, math and language; in the CCN_T design, the students evaluate their instructor in math and language. If we obtained data from both students and teachers in the same study, we would have a reciprocal design. Using the parlance developed by Kenny et al. (2006), such a design is a reciprocal one-with-many study. Other designs can also be combined to form reciprocal designs, as is discussed in the **Supplemental Appendix**.

Supplemental Material

CONCLUSION

In psychology experiments, we frequently ask participants to respond to targets (e.g., faces, words, other people) in various experimental conditions. In such experiments, the interest is in reaching conclusions about condition differences that generalize to other samples of participants and, typically, other samples of targets that might have been used. To permit this, it is essential that the analysis of the resulting data treat both participants and targets as random factors. The failure to do so, collapsing across targets or ignoring the variation they induce, leads to serious bias and, we suggest, failures to replicate experimental effects when other samples of targets are used.

In this review, we provide an exhaustive typology of designs based on the nesting or crossing of three factors (participants, targets, and condition) in such experiments. For each of these designs, we discuss the mixed-model specification that permits unbiased estimates of condition effects while treating both participants and targets as random factors. Additionally, we provide tools to estimate the effect size and statistical power of the condition difference in each design.

We conclude by emphasizing the importance of considering targets as well as participants in determining statistical power. We also discuss considerations that ought to drive the choice among alternative designs. Our hope is that researchers will adopt appropriate designs and analytic models that incorporate target variation and thus permit conclusions that are more likely to replicate with other samples of both participants and targets.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Markus Brauer, Leonard Katz, Reinhold Kliegl, Gary McClelland, Dominique Muller, and Vincent Yzerbyt for their helpful comments on drafts of this review.

LITERATURE CITED

Baayen RH, Davidson DJ, Bates DM. 2008. Mixed-effects modeling with crossed-random effects for subjects and items. *J. Mem. Lang.* 59:390–412

Barr DJ, Levy R, Scheepers C, Tily HJ. 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68:255–78

Bates D, Kliegl R, Vasishth S, Baayen H. 2015. Parsimonious mixed models. arxiv:1506.04967v1 [stat.ME]

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, et al. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.* 24(3):127–35

Bond CF, DePaulo B. 2008. Individual differences in judging deception: accuracy and bias. *Psychol. Bull.* 134:477–92

Clark HH. 1973. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. *J. Verb. Learn. Verb. Behav.* 12:335–59
Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum
Goldstein H, Browne W, Rasbash J. 2002. Partitioning variation in multilevel models. *Underst. Stat.* 1:223–31
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. *J. Exp. Psychol. Hum. Percept. Perfor.* 32:199–209
Hox JJ. 2010. *Multilevel Analysis: Techniques and Applications*. New York: Routledge
Judd CM, McClelland RG, Ryan CS. 2008. *Data Analysis: A Model Comparison Approach*. New York: Routledge
Judd CM, Westfall J, Kenny DA. 2012. Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *J. Pers. Soc. Psychol.* 103:54–69
Kenny DA, Kashy DA, Cook WL. 2006. *Dyadic Data Analysis*. New York: Guilford Press
Raudenbush SW, Bryk AS. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage
Satterthwaite FE. 1946. An approximate distribution of estimates of variance components. *Biom. Bull.* 2(6):110–14
Smith ER. 2014. Research design. In *Handbook of Research Methods in Social and Personality Psychology*, ed. HT Reis, CM Judd, pp. 27–48. Cambridge, UK: Cambridge Univ. Press
Snijders T, Bosker R. 2011. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks, CA: Sage
Stroup WW. 2012. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. New York: CRC Press
Toma C, Corneille O, Yzerbyt V. 2012. Holding a mirror up to the self: egocentric similarity beliefs underlie social projection in cooperation. *Pers. Soc. Psychol. Bull.* 38:1259–71
Welch BL. 1947. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika* 34:28–35
Westfall J, Judd CM, Kenny DA. 2015. Replicating studies in which samples of participants respond to samples of stimuli. *Pers. Psychol. Sci.* 10:390–99
Westfall J, Kenny DA, Judd CM. 2014. Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *J. Exp. Psychol. Gen.* 143:220–45
Winer BJ. 1971. *Statistical Principles in Experimental Design*. New York: McGraw-Hill

Annual Review of Psychology

Volume 68, 2017

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Psychology* articles may be found at http://www.annualreviews.org/errata/psych