Journal of Personality and Social Psychology

0022-3514/12/$12.00 DOI: 10.1037/a0028347

Treating Stimuli as a Random Factor in Social Psychology: A New and Comprehensive Solution to a Pervasive but Largely Ignored Problem

Charles M. Judd and Jacob Westfall
University of Colorado Boulder

David A. Kenny
University of Connecticut

Throughout social and cognitive psychology, participants are routinely asked to respond in some way to experimental stimuli that are thought to represent categories of theoretical interest. For instance, in measures of implicit attitudes, participants are primed with pictures of specific African American and White stimulus persons sampled in some way from possible stimuli that might have been used. Yet seldom is the sampling of stimuli taken into account in the analysis of the resulting data, in spite of numerous warnings about the perils of ignoring stimulus variation (Clark, 1973; Kenny, 1985; Wells & Windschitl, 1999). Part of this failure to attend to stimulus variation is due to the demands imposed by traditional analysis of variance procedures for the analysis of data when both participants and stimuli are treated as random factors. In this article, we present a comprehensive solution using *mixed models* for the analysis of data with crossed random factors (e.g., participants and stimuli). We show the substantial biases inherent in analyses that ignore one or the other of the random factors, and we illustrate the substantial advantages of the mixed models approach with both hypothetical and actual, well-known data sets in social psychology (Bem, 2011; Blair, Chapleau, & Judd, 2005; Correll, Park, Judd, & Wittenbrink, 2002).

Keywords: stimulus sampling, random effects, mixed models

The issue of stimulus sampling has a long but somewhat neglected history in social cognitive psychology (Brunswik, 1955; Kenny, 1985; Wells & Windschitl, 1999). In fact, some of the harshest research critiques come when conclusions are reached, for instance, about reactions to African Americans when an experimental participant has encountered a single African American as a stimulus in an experimental interaction. Most researchers do not make this mistake in its most extreme form, as they generally attempt to include some reasonable sample of stimuli in order to suggest generalization.

Seldom, however, is the sampling of stimuli explicitly taken into account in the analysis of social cognitive data. Instead, the standard practice is to average the responses of individual participants across stimuli and analyze these resulting averages, with participant as the sole random factor in the design.

This practice persists in spite of the now-classic warning given by Clark (1973) nearly 40 years ago, who argued at the time that many "statistically significant" effects in cognitive psychology do not permit robust generalization across stimuli, when stimuli are in fact treated as a random factor in the analysis. At that time, the standard analytic fix, assuming one wanted to treat both participants and stimuli as random effects in an analysis, was to calculate quasi-F statistics, following the exposition in Green and Tukey (1960), Winer (1971), and elsewhere, based on the full variance partitioning of experimental designs with orthogonal experimental factors (that include participants and stimuli).

Since that time, the use of experimental paradigms with samples of stimuli to which participants react has only increased in prevalence, especially with the abundant recent interest in examining response latency measures as indicators of implicit responses, which are generally quite unreliable unless one employs a large number of trials, involving different stimuli, to which responses are given (e.g., implicit association test, affective priming, lexical decision task, affect misattribution task, shooter task, and so on).

Part of the reason why Clark's warnings have gone unheeded is because of the inherent limitations of using designs in which traditional analysis of variance (ANOVA) decomposition can proceed and associated quasi-F statistics computed. Such designs generally require complete data with orthogonal effects of the various crossed and nested factors (i.e., equal n) and no continuous covariates. And for each new design, relatively complex calculations are necessary to derive the expected mean squares (Cornfield & Tukey, 1956; Winer, 1971) and from these, the appropriate quasi-F statistics.

This failure to adopt analyses that treat participants and stimuli as simultaneous random effects comes at a substantial cost. Analyses that rely on averaging across stimuli to obtain within-participant cell means ignore systematic variation between experimental stimuli, and this variation may contribute to statistically significant mean differences that may not replicate in studies with different stimulus samples. It has been shown that the failure to

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

We are grateful to Daryl Bem, Irene Blair, and Joshua Correll for making data available to us. We also thank John Lynch, Gary McClelland, Dominique Muller, Vincent Yzerbyt, and members of the CUSP lab at the University of Colorado for helpful comments on an earlier version of this article.

Correspondence should be addressed to Charles M. Judd, Department of Psychology and Neuroscience, University of Colorado, Boulder, CO 80309-0345. E-mail: Charles.Judd@colorado.edu

treat stimuli as a random effect can cause the empirical Type 1 error rate for an analysis to exceed the nominal alpha level by more than an order of magnitude, depending on the details of the experimental design (Forster & Dickinson, 1976; Rietveld & van Hout, 2007; Wickens & Keppel, 1983).

In recent years, alternative statistical techniques have been developed that easily and effectively address the problems for which quasi-*F* ratios were proposed (Baayen, Davidson, & Bates, 2008). Further, these *mixed effects models* offer many additional advantages over both traditional repeated-measures ANOVA and quasi-*F* statistics. These include the ability to handle incomplete and unbalanced data, the ability to easily accommodate continuous as well as categorical predictors, avoidance of information loss due to prior averaging over stimuli or participants, principled unbiased handling of incomplete and/or outlying cases, a simple empirical solution to the problem of when effects ought to be considered fixed or random, and their widespread availability and relative ease of use (see Baayen, 2008, chapter 7).

We have three major purposes in writing this article. The first is to underline once again the pitfalls of not treating both participants and stimuli as random effects. We discuss these pitfalls in the context of a design in which multiple participants are each exposed to multiple stimuli and these stimuli are in one of two experimental conditions for all participants. Thus, participants are crossed with treatment condition, and in each treatment condition, there are multiple stimuli. We discuss analyses that ignore one or the other of the random effects in this design and show through a series of simulations the substantial costs of doing so.

Second, still in the context of this design, we briefly review the classic quasi-*F* solution to testing the treatment effect while treating both participants and stimuli as random. We then introduce the newer and more general approach based on mixed models with multiple crossed random effects. In addition to detailing this approach and its specification, we explore both its Type I error rates, given no treatment effects, and its associated statistical power.

Finally, we turn to the benefits of the mixed effects modeling approach, illustrating with some well-known social cognition data sets the ability of mixed effects models to easily handle missing data, to accommodate treatment variables that vary continuously rather than discretely, and to provide meaningful estimates of variance components that are often of great theoretical interest.

Illustrative Nested-Stimuli Design and Typical Analyses

As just discussed, we start with a ubiquitous but simple design in which each participant responds to several different stimuli that are nested under two levels of an independent variable of interest. For example, imagine that participants are asked to make judgments about African American and White males, with each stimulus person described with a name and photo. In this design, the independent variable of interest, race of the stimulus person, is crossed with each participant and the individual stimulus persons are nested under it.

In the context of this design, there are two random factors (participants and stimuli) and one fixed factor (stimulus race). Although there is ambiguity in the literature on how best to define the random–fixed distinction (Gelman & Hill, 2007, p. 245), for now we define these terms as they have been conventionally in the ANOVA literature (Green & Tukey, 1960; Winer, 1971). According to this definition, random factors are factors whose levels are sampled from some larger population of levels across which the researcher wishes to generalize, whereas fixed factors are those whose levels are exhaustive. To say that both participants and stimuli are random factors is to say that we have two different samples in this study, one of participants and one of stimuli, drawn in theory from populations of interest. Our participants are sampled from some population to which we would presumably like to make inferences. Likewise, our stimulus photographs are also sampled in theory from some population to which we would also like to make inferences. In both cases, although the populations may be only imprecisely specified, it is clearly the case that our two samples do not exhaust the universe of potential participants and stimuli that might be used.[1]

Data from such a design are typically analyzed ignoring one or the other random effects. The most typical analysis treats participants as random but ignores the random effects due to stimuli. This analysis involves computing two mean scores for each participant, one averaging that participant's judgments across all African American stimuli and one averaging that participant's judgments across all White stimuli. Then one conducts a repeated-measures ANOVA, treating stimulus person race as a within-participant factor, to estimate and test the significance of the mean difference in judgments as a function of race. We refer to this most typical analysis as the *by-participant* analysis.

A much less typical analysis is one that ignores the random effects due to participants. This analysis involves computing a mean for each stimulus, averaging across all participants. Then one conducts a between-stimuli analysis of variance on these means to ask whether the mean judgments given to the African American stimuli differ from the mean judgments given to the White stimuli, treating stimuli as the unit of analysis. We refer to this analysis as the *by-stimulus* analysis.

In order to understand the statistical pitfalls of these two analyses that ignore one or the other of the random effects in the design, it is helpful to examine the expected values of the mean squares for all sources of variation in data from this design. These expected mean squares are given in Table 1. We will not present the mathematical derivation of these expected mean squares (see Cornfield & Tukey, 1956; Green & Tukey, 1960; Winer, 1971). Instead we will focus on developing an intuitive understanding of what the terms of these expected mean squares are telling us and of the serious consequences that result from ignoring certain variance components underlying them by analyzing aggregated within-participant means (the by-participant analysis) or within-stimulus means (the by-stimulus analysis).

[1] As we will develop at a later point in this article (see Footnote 9), a more complete definition of fixed versus random factors specifies not only the sampling of its levels, as in the analysis of variance literature, but also whether the effects of that factor vary. For instance, to say stimuli is a random factor means both that we select only a sample of stimuli and that different stimuli have different effects (i.e., they make a difference in the response measured). When the effects of a factor do not vary, then ultimately it can be treated as fixed even if in fact its levels were sampled. This then permits tests of whether factors ultimately should be treated as fixed or random, which mixed models allow, as we previously suggested.

Table 1
Expected Mean Squares from a Design in which Stimuli are Nested Under Condition and Participants are Crossed with Condition, with Stimulus as Fixed and Participants Random, Participants as Fixed and Stimuli Random, or Participants and Stimuli Both as Random

Label	Source of variance	Degrees of freedom	Stimulus factor fixed/ participants random	Participant factor fixed/stimuli random	Participant and stimulus factors both random
C	Condition (r)	$r - 1$	$\sigma_e^2 + q\sigma_{P\times C}^2 + pq\sigma_C^2$	$\sigma_e^2 + p\sigma_{S(C)}^2 + pq\sigma_C^2$	$\sigma_e^2 + \sigma_{P\times S(C)}^2 + q\sigma_{P\times C}^2 + p\sigma_{S(C)}^2 + pq\sigma_C^2$
S(C)	Stimuli within Condition (q)	$r(q - 1)$	$\sigma_e^2 + \sigma_{P\times S(C)}^2 + p\sigma_{S(C)}^2$	$\sigma_e^2 + p\sigma_{S(C)}^2$	$\sigma_e^2 + \sigma_{P\times S(C)}^2 + p\sigma_{S(C)}^2$
P	Participants (p)	$p - 1$	$\sigma_e^2 + rq\sigma_p^2$	$\sigma_e^2 + \sigma_{P\times S(C)}^2 + rq\sigma_P^2$	$\sigma_e^2 + \sigma_{P\times S(C)}^2 + rq\sigma_P^2$
$P\times C$	Participants by Condition	$(r - 1)(p - 1)$	$\sigma_e^2 + q\sigma_{P\times C}^2$	$\sigma_e^2 + \sigma_{P\times S(C)}^2 + q\sigma_{P\times C}^2$	$\sigma_e^2 + \sigma_{P\times S(C)}^2 + q\sigma_{P\times C}^2$
$P \times S(C)$	Participants by Stimuli within Condition	$r(p - 1)(q - 1)$	$\sigma_e^2 + \sigma_{P\times S(C)}^2$	$\sigma_e^2 + \sigma_{P\times S(C)}^2$	$\sigma_e^2 + \sigma_{P\times S(C)}^2$

We assume that we have p Participants and q Stimuli within each level of Condition, and r Conditions (in the race design, $r = 2$). In addition to these three sources of variance, we also have the Participant × Condition interaction (i.e., different participants may show different magnitudes of the condition difference), the Participant × Stimulus interaction (i.e., participants may respond idiosyncratically to particular stimuli), and residual error variance. With each participant responding to each stimulus only once, in this design the Participant × Stimulus interaction is confounded with the residual error variance. Each source of variance has a mean square associated with it, and these mean squares, in turn, have expected values that are functions of the six underlying variance components. These are the residual error variance σ_e^2 as well as a variance component uniquely due to each source of variation in the data, with the source denoted in subscript.[2] Notice that the variance components expected to underlie each mean square depend on whether we consider Participants and/or Stimuli to be random factors. The fourth and fifth columns give the variance components if either Stimuli *or* Participants are not random, respectively. The sixth column gives the variance components if they are both random. It is this last column on which we focus, because we assume that both factors are in fact random, and we want to examine the biases that ensue from the by-participant analysis or the by-stimulus analysis.

An unbiased F test of the Condition effect involves choosing a denominator, or error term, that includes all of the variance components that underlie the Condition mean square *except* for the variance component due to Condition, σ_C^2. The resulting F ratio then tests the null hypothesis that σ_C^2 equals 0. The by-participant analysis, based on participant means and ignoring the variation due to stimuli, involves a repeated-measures ANOVA, analyzing the mean difference between conditions within participants. The resulting F ratio is the ratio of the mean square (MS) due to Condition and the mean square due to Participants × Condition:

$$F_{1,(p-1)} = \frac{MS_C}{MS_{P\times C}}$$

If stimuli are in fact a fixed factor (column 4 of the table), then the expected values for the numerator and denominator of this F ratio are

$$F_{1,(p-1)} = \frac{\sigma_e^2 + q\sigma_{P\times C}^2 + pq\sigma_C^2}{\sigma_e^2 + q\sigma_{P\times C}^2}$$

As a result, the numerator and denominator have different expected values only if the variance due to condition is not zero. Accordingly, if stimuli are fixed, this analysis is appropriate, permitting generalization to other studies with different participants but the same stimuli.

On the other hand, if stimuli are in fact a random effect (column 6), then the expected values for the numerator and denominator of this F ratio are

$$F_{1,(p-1)} = \frac{\sigma_e^2 + \sigma_{P\times S(C)}^2 + q\sigma_{P\times C}^2 + p\sigma_{S(C)}^2 + pq\sigma_C^2}{\sigma_e^2 + \sigma_{P\times S(C)}^2 + q\sigma_{P\times C}^2} \quad (1)$$

In the absence of any condition effect (i.e., $\sigma_C^2 = 0$), the numerator and denominator do not have the same expected values. The expected value of the numerator will be larger than the expected value of the denominator as a function of the number of participants (p) and the magnitude of the variation between Stimuli within Condition, $\sigma_{S(C)}^2$, leading to alpha inflation.

In a practical sense, what this means is that the by-participant analysis in any one study will yield a positively biased F for the condition effect, assuming stimuli are random, to the extent that the variation between stimuli within conditions is large, among other factors. To provide a more intuitive understanding of this bias, let us assume that there is no condition effect. In a particular study, however, only a sample of stimuli are included, and because of this stimulus sampling, it is extremely likely that those in one condition may elicit somewhat different scores on average across the participants than those in the other condition. Hence, even if there were no true condition difference, the sampling of stimuli in any one study would undoubtedly yield some small condition difference in that particular study. As the number of participants

[2] As we have said, in this design where each participant responds only once to each stimulus, the variance due to the Participant × Stimulus interaction is confounded with the residual error variance. Nevertheless, in Table 1 and the formulas that follow, we keep these terms separate in the interest of generality to situations where there are replications of each Participant × Stimulus observation.

increases, this small condition difference, due to stimulus sampling, will loom larger and larger. Additionally, the bias that results from stimulus sampling in any one study will of course be greater when the number of stimuli is relatively small.[3] Hence, the by-participant analysis will yield *F*s that are too large if in fact stimuli are random. This bias increases as the variation between stimuli, $\sigma^2_{S(C)}$, increases, as this makes larger random condition differences more likely. Additionally it increases as the number of participants increases and as the number of stimuli decreases.

Although not routinely employed by social psychologist, the by-stimulus analysis is also biased, albeit by different factors. As already discussed, in this analysis one collapses across participants, computing a mean judgment for each stimulus. Then one conducts a between-stimulus analysis of variance to test whether the mean judgments of stimuli in one condition differ from the mean judgments of stimuli in the other. The resulting *F* is the ratio of the mean square due to Condition and the mean square due to Stimuli within condition:

$$F_{1,2(q-1)} = \frac{MS_C}{MS_{S(C)}}$$

Parallel to the analysis that treats stimuli as a fixed factor, this analysis implicitly assumes participants to be a fixed factor (column 5 of Table 1), permitting generalization to future studies involving different samples of stimuli but the same participants. The assumption of participants as a fixed factor is clearly an assumption with which most social psychologists would not be happy.

From the sixth column of Table 1, we can see that the numerator and denominator of this *F* ratio will have the following expected values, given that both participants and stimuli are random

$$F_{1,2(q-1)} = \frac{\sigma^2_e + \sigma^2_{P\times S(C)} + q\sigma^2_{P\times C} + p\sigma^2_{S(C)} + pq\sigma^2_C}{\sigma^2_e + \sigma^2_{P\times S(C)} + p\sigma^2_{S(C)}} \qquad (2)$$

And again the numerator and denominator of this *F* ratio do not have the same expected values in the absence of a condition effect. The expected value of the numerator will exceed the expected value of the denominator as a function of the number of Stimuli (q) and the magnitude of the Participant × Condition variance component ($\sigma^2_{P\times C}$), again leading to alpha inflation.

In a practical sense, what this means is that the by-stimulus analysis in any one study will yield a positively biased *F* for the condition effect to the extent that there is variation in the magnitude of the condition difference from participant to participant, again assuming participants to be random. To provide a more intuitive understanding of this bias, let us again assume that there really is no condition effect (i.e., across all possible participants, the mean condition difference is zero). In a particular study, however, only a sample of participants are included and because of this participant sampling, it is extremely likely that the average condition difference across participants will not be exactly zero in that particular study. Some participants will show a condition difference in one direction and some in the other. And when one collapses across them, the average condition difference will not be zero simply because only a sample of participants has been used. As the number of stimuli increases, this small condition difference, due to participant sampling, will loom larger and larger. Additionally, the bias that results from participant sampling in any one study will of course be greater when the number of participants is relatively small.[4] Hence, the by-stimulus analysis will yield *F*s that are too large if in fact participants are random. This bias will increase as the variation in the condition difference between participants ($\sigma^2_{P\times C}$) increases, again because this makes larger random condition differences more likely. Additionally it will increase as the number of stimuli increases and as the number of participants decreases.

Type 1 Error Rates for By-Participant and By-Stimuli Analyses

To demonstrate these conclusions and illustrate the approximate magnitude of alpha inflation that an experimenter might expect to introduce in his or her analysis by inappropriately treating participants or stimuli as fixed rather than random effects, we present the results of a Monte Carlo simulation of the experimental design discussed in the previous section. In this design, as in the previous examples, stimuli are nested under two treatment conditions while participants are crossed with stimuli and treatment conditions. This represents a typical "within participants" design that one would encounter in the social cognitive literature.

We varied two aspects of the experimental design orthogonally across all simulated experiments: the number of participants in the experiment and the number of experimental stimuli used in total across the two conditions,[5] both of which ranged independently from 10 to 90 in steps of 20, resulting in a total of 25 experimental designs. Each cell of this simulation matrix consisted of 10,000 simulations, for a total of 250,000 simulated experiments. The variance components were held constant across the initial simulations at $\sigma^2_e = 16$, $\sigma^2_{P\times S(C)} = 0$, $\sigma^2_P = \sigma^2_{P\times C} = \sigma^2_{S(C)} = 4$, and the true Condition difference was set at zero.[6] The data from these simulated experiments were analyzed with both traditional by-participant and by-stimulus analyses as described in the previous section, with alpha = .05. We note that our choice of the magnitude of the variance components is arbitrary and that the relevant feature of the variance components in any analysis is their *relative* magnitudes. The purpose of these simulations is to give a general illustration of the degree of bias associated with ignoring random effects for a reasonable set of variance components and in particular to show how this bias varies as a function of the two sample sizes.

The results of the simulation are given in Table 2. In these simulated experiments, the standard by-participant analysis of variance showed a remarkable degree of positive bias. Empirical Type 1 error rates for this analysis ranged from .086 in the best

[3] Note that $q\sigma^2_{P\times C}$ is found in both the numerator and denominator of Equation 1. Hence with smaller q, the bias due to the presence of $\sigma^2_{S(C)}$ in the numerator increases.

[4] Parallel to the explanation in Footnote 3, $p\sigma^2_{S(C)}$ is found in both the numerator and denominator of Equation 2. Hence, with smaller p, the bias due to the presence of $\sigma^2_{P\times C}$ in the numerator increases.

[5] We assume that this total number of stimuli is divided equally between the two conditions. Accordingly, the total number of stimuli is equal to $2q$, using the definition of q from the earlier variance decomposition.

[6] Again, the variance component due to the Participant × Stimulus interaction is set to zero because it is confounded with the residual error variance.

Table 2
Empirical Type 1 Error Rates for By-Participant and By-Stimulus Analyses

No. of participants/ Type of analysis	No. of stimuli				
	10	30	50	70	90
By-participant					
10	.187	.133	.105	.095	.086
30	.381	.288	.233	.194	.170
50	.494	.394	.315	.279	.241
70	.560	.451	.392	.335	.296
90	.616	.506	.442	.385	.351
By-stimulus					
10	.070	.108	.150	.182	.221
30	.053	.074	.093	.105	.130
50	.055	.065	.078	.088	.100
70	.049	.058	.072	.078	.085
90	.055	.058	.061	.070	.077

case to .616 in the worst case, with an average error rate of .317, over six times the nominal alpha level. Consistent with the expected mean square formula for this analysis (Equation 1), increasing the number of participants led to greater positive bias in the error rate. In the worst case, with 90 participants, the average error rate was .460; while in the best case, with only 10 participants, the average error rate was still substantially inflated at .121. Increasing the number of stimuli while holding constant the number of participants did improve the Type 1 error rates somewhat. However, this improvement was never enough to counterweigh the strong positive bias due to participants. Even in the best case with 90 stimuli, Type 1 error rates ranged from .086 (with 10 participants) to .351 (with 90 participants), with an average of .229.

The by-stimulus analysis performed better than the by-participant analysis in these simulated experiments, but was still considerably biased. Type 1 error rates for this analysis ranged from .049 to .221, with an average error rate at .089. Patterns for the Type 1 error rates generally mirrored those in the by-participant analysis. That is, increasing the number of stimuli led to increasingly inflated Type 1 error rates, while increasing the number of participants only partially counteracted this positive bias. Only in the experiments with 10 stimuli did the Type 1 error rates even begin to approach an acceptable average level at .056. Of course, a between-stimulus analysis with only 10 stimuli will generally be severely deficient in statistical power.

Because these simulations were based on rather arbitrary specifications of the variance components, we redid them four different times, each time reducing one of the four variance components by half, but leaving the other three components as previously specified. When $\sigma^2_{P\times C}$ was set at 2 (rather than 4 as in the previous simulations) and all other components specified as previously, all Type 1 error rates for the by-participant analysis increased (the average was .382 instead of .317) while the Type 1 error rates for the by-stimulus analysis slightly decreased (the average was .070 instead of .089). When $\sigma^2_{S(C)}$ was reduced to 2 (rather than 4 as in the previous simulations) and again all other components were specified as in the original simulations, all Type 1 error rates for the by-stimulus analysis increased (the average was .115 instead of .089) while the Type 1 error rates for the by-participant analysis decreased (the average was .221 instead of .317). When σ^2_P was reduced to 2, there were minimal effects on Type 1 error rates for either the by-participant or the by-stimulus analyses. Finally, a reduction of σ^2_e, from 16 to 8, resulted in somewhat greater inflation of Type 1 error rates for both the by-participant and the by-stimulus analyses. In sum, what these further simulations show is that the Type 1 error rates remain inflated for these tests regardless of variation in the magnitude of these components of variance.

In some areas of psychology, and especially in psycholinguistics, it has become common practice to report the results of both the by-participant and by-stimulus analyses of the data, and to accept a result as significant only when both individual analyses indicate a significant result (Raaijmakers, Schrijnemakers, & Gremen, 1999). The reasoning seems to be that if a significant result for the by-participant analysis permits one to generalize across participants, and a significant result for the by-stimulus analysis permits one to generalize across stimuli, then it must be that having significant results from both analyses permits one to generalize across both participants and stimuli.

As others have pointed out, this reasoning, though intuitively appealing, is in fact flawed (Raaijmakers et al., 1999; Raaijmakers, 2003). Conceptually, a significant by-participant result suggests that experimental results would be likely to replicate for a new set of participants, but only using the same sample of stimuli. A significant by-stimulus result, on the other hand, suggests that experimental results would be likely to replicate for a new set of stimuli, but only using the same sample of participants. However, it is a fallacy to assume that the conjunction of these two results implies that a result would be likely to replicate with *simultaneously* new samples of both participants and stimuli.

Treating Both Participants and Stimuli as Random: Quasi-*F*s and Mixed Models

The classic solution for testing the effects of a fixed treatment variable in the presence of two crossed random factors (participants and stimuli) involves the computation of a quasi-*F* ratio (Clark, 1973; Winer, 1971). This quasi-*F* statistic derives from the variance decomposition of the full design and the computation of expected mean squares in terms of the variance components that contribute to the overall variation in the data.

Given the design we have been using, the recommended quasi-*F* for an analysis that collapses across neither participants nor stimuli is given by

$$F_{df_n,df_d} = \frac{MS_C + MS_{P\times S(C)}}{MS_{P\times C} + MS_{S(C)}}$$

In terms of expected mean squares from Table 1, this quasi-*F* has an expected value of

$$F_{df_n,df_d} = \frac{(\sigma^2_e + \sigma^2_{P\times S(C)} + q\sigma^2_{P\times C} + p\sigma^2_{S(C)} + pq\sigma^2_C) + (\sigma^2_e + \sigma^2_{P\times S(C)})}{(\sigma^2_e + \sigma^2_{P\times S(C)} + q\sigma^2_{P\times C}) + (\sigma^2_e + \sigma^2_{P\times S(C)} + p\sigma^2_{S(C)})}$$

The rationale underlying this quasi-*F* is that its numerator will exceed its denominator only to the extent that that the variance due to condition (σ^2_C) is greater than zero. The degrees of freedom associated with this quasi-*F* are approximate, due to the fact that it is not truly an *F* ratio; that is, it is not the ratio of two chi-square-distributed random variables. Expressions for calculating approx-

imate df_n and df_d can be found in the relevant literature we have cited.

As we have said before, the variance decomposition and expressions for the expected mean squares of Table 1 rely on assumptions of complete data and accordingly independence of the fixed and random factors in the design. Additionally, in designs more complicated than the simple one we have so far explored, the computation of the expected mean squares and the derivation of the appropriate quasi-*F* ratio become more complex.

A more general and tractable solution to the analysis of data with multiple random effects is provided by the literature and software devoted to what is called *mixed effects* modeling of data. Social psychologists are increasingly familiar with a subset of these models that are variously known as *hierarchical linear models* or *multilevel models* (Hox, 2002; Kenny, Kashy, & Bolger, 1998; Raudenbush & Bryk, 2002; Singer & Willett, 2003; Snijders & Bosker, 1999). Data structures for such models involve hierarchically nested random factors. For instance, a common situation in which multilevel modeling is used involves students who are nested in classrooms. There may be one or more fixed effects included in such designs, and these may vary either at the higher level of the design (between classrooms) or at the lower level (between students within classrooms). The estimation of these models typically involves restricted maximum likelihood estimation whereby different random error variances are estimated at the different levels of the model.

Less well known is that these models can accommodate data structures with crossed, rather than nested, random effects. Indeed some of the best known sources that explicate multilevel models contain chapters that discuss models with crossed random effects (Chapter 12 in Raudenbush and Bryk, 2002; Chapter 11 in Snijders and Bosker, 1999). Most recently, Baayen, Davidson, and Bates (2008) have shown how such models can be used to analyze designs such as the one that we have been considering, where participants and stimuli are both random and crossed with each other, and one or more treatment variables varies either within or between participants and/or stimuli.

Let us consider an example data set of the design that we have been considering with two crossed random effects, participants and stimuli, and a treatment varying within participants but between stimuli.[7] In this example data set, there are 30 participants who each give responses to 30 stimuli. These stimuli are nested within one of two treatment conditions, with the first 15 stimuli in one condition ($C = -0.5$) and the second fifteen in a second condition ($C = +0.5$).[8]

We can specify a basic linear regression model for these data simply as

$$Y_{ij} = \alpha_0 + \alpha_1 C_{ij} + \varepsilon_{ij}$$

where i refers to participant, j refers to stimulus within condition, Y is the outcome variable, and C is treatment or condition (as defined previously). This model contains a parameter for the intercept, α_0, a parameter for the slope of the condition effect, α_1, and the residual error term ε_{ij}. For these data, this basic regression equation almost certainly violates one of the major assumptions of the general linear model—the assumption that the residuals ε_{ij} are independent—because it ignores the natural groupings in the data due to both participants and stimuli. Indeed, the original motivation to compute within-participant mean scores was precisely to correct this nonindependence by reducing each subject's data to two means and a single difference between these, giving us the simplified model

$$Y_{Di} = \alpha_0 + \varepsilon_i$$

where the outcome variable Y_{Di} is now the difference between the two within-participant means computed over the stimuli nested under each level of the treatment and where the diminished number of residuals ε_i (one for each participant) are independent of one another. Although this has long been the standard fix among psychologists, it fails to deal completely with the nonindependence in the data.

Mixed effects models avoid the problems associated with analyzing within-participant or within-stimulus mean scores by instead explicitly modeling the dependencies in the data. They do this by partitioning the error term in the classical regression model into several different "errors." These include the usual residual error term plus a number of *random effect* terms. These additional terms account for the dependencies in the data by adjusting the predicted values of the model separately for each level of the grouping factors (e.g., for each participant and/or stimulus). Note that in the mixed effects framework, it becomes an empirical question whether random effects are warranted for a given factor. That is, although our initial designation of a particular factor as either fixed or random may be guided by conceptual concerns regarding the intended universe of generalization, it is ultimately the empirically estimated random variance components from our mixed model that inform us whether and to what extent experimental effects vary randomly with respect to each grouping factor. Mixed effects models therefore offer a natural solution to the issues that we outlined above concerning treating factors as fixed versus random.[9]

Let us reformulate this model in terms of a mixed model. In this model, the intercept has two random components, one varying from participant to participant and one varying from stimulus to stimulus within condition. These allow for the fact that some participants have higher scores on average than others, and some stimuli elicit higher scores on average than others. The slope for the Condition variable has only one random error component, varying across participants (because participants are crossed with Condition but stimuli are not). Thus, we have a fixed effect of Condition that is estimated along with four different random error components: variation in the intercept due to stimuli, variation in the intercept due to participants, variation in the Condition slope

[7] For readers who wish to run analyses on these data, parallel to those we report, they are available at http://psych.colorado.edu/~cjudd/mixedexample.csv, with one row for each Stimulus × Participant observation, for a total of 900 rows of data.

[8] We chose these codes (contrast or effect codes) over other coding methods (e.g., dummy), so that the intercepts in the models that follow estimate the mean across conditions and the condition slope estimates the mean difference.

[9] If the true variance of the effects of a random factor is zero, then it ultimately does not matter if the included levels of the factor are exhaustive or are only sampled from a larger population of levels. In this sense, there are two necessary conditions in our opinion for defining fixed versus random factors: whether or not their levels are sampled and whether or not the effects of the factor (on means, slopes, and so on) vary.

due to participants, and finally random error variation at the level of the individual observation. In other words, the model is more accurately and completely given as

$$Y_{ij} = \alpha_0 + \alpha_1 C_{ij} + \mu_{0j} + \mu_{0i} + \mu_{1i} C_{ij} + \varepsilon_{ij}$$

Estimation involves estimating the fixed effects in this model (the grand intercept α_0 and the condition effect α_1) and the variances of the random effects ($\sigma^2_{\mu_{0j}}$, $\sigma^2_{\mu_{0i}}$, $\sigma^2_{\mu_{1i}}$, and $\sigma^2_{\varepsilon_{ij}}$). Additionally, there is potential covariance between the random participant intercept effect and the participant slope effect (e.g., participants with high intercepts might tend to have low slopes). Because of the fact that there are now multiple random error components with different variances, least-squares estimation has difficulties (Kenny et al., 1998). Instead, estimation typically proceeds iteratively using a restricted maximum likelihood loss function.

Let us now turn to the actual analysis of our example data using mixed effects models. In the Appendix, we provide three different sets of code for conducting this analysis, using the *lme4* package in R, the PROC MIXED procedure in SAS, and the MIXED procedure in SPSS. There we also give the output from each of these analyses, using the specified commands on the illustrative data set. We summarize these results in Table 3.

In general, fixed effects are typically associated with factors of substantive interest that motivated data collection. Random effects are included in models to account for the patterns of nonindependence present in the data, but they may also be of substantive interest, as we illustrate later. In general, we would recommend estimating and testing all fixed effects of theoretical interest and all random variance components dictated by the particular design used. Then models may well be trimmed, if fixed effects turn out to be nonsignificant and if random effects turn out to have zero variances.

There are three possible random effects (in addition to the residual error) in the dataset we are using to illustrate these analyses: (a) intercepts may vary across participants ($\sigma^2_{\mu_{0i}}$; that is, there may be variation in the mean response as a function of participant; (b) Condition slopes may vary across participants, ($\sigma^2_{\mu_{1i}}$); that is, there may be variation in the magnitude of the condition effect as a function of participant; and (c) intercepts may vary across stimuli ($\sigma^2_{\mu_{0j}}$). Additionally, because both the intercepts and slopes vary randomly across participants, it may be that they also covary.[10]

We can test the null hypothesis of zero variance in a random effect by conducting a likelihood ratio test comparing a more general model that estimates the variance in the effect against a nested model that has the same fixed effects structure, but that sets the particular random effect variance to 0. The likelihood ratio test statistic is equal to two times the difference of the log-likelihoods of the two models (sometimes called the model *deviances*). This test statistic is asymptotically distributed as a chi-square distribution with degrees of freedom equal to the difference in the numbers of parameters between the two models. It is generally recommended that a liberal criterion be used for rejecting the nested model when performing hypothesis tests for random effects using likelihood-ratio tests for two reasons. First, as discussed by Pinheiro & Bates (2000), the *p* values for the likelihood-ratio test using this reference chi-square distribution are often conservative; that is, the reported *p* values are often greater than the actual *p* value for the test. It is, therefore, misleading to use the conventional alpha level of .05. Second, failing to include a random effect in a model when in fact the random variance for that effect is not zero can bias the tests of the fixed effects. In general, our recommendation is that one should explicitly model the random effects that may be present in the data when it is feasible to do so.

In this case, the model includes seven parameters (two fixed effect estimates, three random effect variances, one covariance parameter, and the residual variance) and has log-likelihood −2590.54. The parameter estimates are listed in Table 3, including the results of likelihood-ratio tests on each of the random effect variances, as well as the *t* statistics and *p* values for the fixed effects using the Kenward–Roger approximation for degrees of freedom.[11] The random effect parameters are clearly warranted by the data, with the exception of the covariance between participant intercepts and slopes. Additionally, we see a significant fixed effect of Condition, $F_{1,\ 38.52} = 9.11$, $p = .005$.

For illustration's sake, we also present the results of traditional by-participant and by-stimulus analyses on these illustrative data. The by-participant analysis, which in this case would be equivalent to a paired *t* test on the within-subject differences in means, gives us $F_{1,\ 29} = 30.48$, $p < .0001$. The by-stimulus analysis, which in this case would be equivalent to a two-sample *t* test comparing the stimuli nested under each level of Condition, gives us $F_{1,\ 28} = 11.38$, $p = .0022$. Both of these *F* ratios are greater than the *F* we obtained from the mixed model with crossed random effects for both participants and stimuli.

Estimating a mixed model that includes random intercepts and slopes for participants, but no random effects for stimuli, gives us exactly the same *F* ratio that we obtained from the by-participant analysis, which implicitly assumed participants to be the only random effect. Likewise, estimating a mixed model that includes random intercepts for stimuli, but no random effects for participants, gives us the same *F* that we obtained from the by-stimulus analysis. It turns out that for certain experimental designs with complete, balanced data, these two pairs of (inappropriate) analyses are formally equivalent.

Empirical Type 1 and Power Estimates for the Mixed Models Analyses

Our goals in this section are twofold. First, we want to demonstrate through further simulations that the mixed models approach in which both participants and stimuli are treated as random yields acceptable Type 1 error rates in exactly the same situations where our earlier simulations demonstrated substantial inflation in Type 1 error rates when either the by-participant or by-stimulus analyses were used. Second, we provide some further simulations that begin to explore the relative statistical power of the mixed model analysis, given the presence of a true Condition difference, under

[10] The presence of a covariance between a random slope and a random intercept in a design with a categorical condition variable, such as the one we are using, will be affected in part by the coding convention used to code that variable. It is partly for this reason that we have chosen to use a contrast coding convention which builds in no such covariance by design. This is not true for other coding conventions (e.g., using a dummy code).

[11] See the Appendix for a discussion of how degrees of freedom are estimated in mixed models.

Table 3
Mixed-Models Results of the Illustrative Data Set

Effect	Variance	χ_1^2	Estimate	*SE*	*t*	*df*	*p*
Random effects							
Participants							
Intercept	4.2940	145.9					<.0001
Condition	4.1823	24.3					<.0001
Covariance (Intercept, Condition)	1.1465	1.3					.2539
Stimuli							
Intercept	3.6706	117.7					<.0001
Residual	15.557						
Fixed effects							
Intercept			−0.180	0.532	−0.339	50.47	.736
Condition			2.521	0.834	3.018	38.52	.005

Note. $-2 \times$ log-likelihood = 5180. See Appendix for the SAS, SPSS, and R input code and output results.

varying numbers of participants and stimuli, again given the classic design that has been the basis for all we have done so far.

For the simulations that explored Type 1 error rates, as before we varied the number of participants in the experiment and the number of experimental stimuli used in total across the two conditions, with both ranging from 10 to 90 in steps of 20, resulting in a total of 25 unique experimental designs. Each cell of this simulation matrix consisted of 10,000 simulations. And as before, the variance components were held constant across the simulations at $\sigma_e^2 = 16$, $\sigma_P^2 = \sigma_{P\times C}^2 = \sigma_{S(C)}^2 = 4$, and the true Condition difference was zero. Using the Kenward–Roger approximation for the degrees of freedom for tests of the fixed effects, we analyzed the data from each experiment using the specification laid out in the previous section, treating both participant and stimulus as random factors and estimating variance components due to participant intercepts, stimulus intercepts, participant slopes, and residual variation.

The resulting empirical Type 1 error rates for these simulations are given in Table 4. On average, across these 250,000 generated data sets, with the number of participants and the number of stimuli varying from 10 to 90, the average Type 1 error rate was .0494. Although there was some variation in these Type 1 error rates across the 25 cells of the design matrix in Table 4 (from a low of .044 to a high of .055), this variation does not appear to be systematic. These results strongly suggest that when there truly is no effect of Condition, this mixed models approach, treating both participants and stimuli as random, yields tests of the Condition difference that have appropriate Type 1 error rates.

Table 4
Empirical Type 1 Error Rates for Mixed Models Analysis Treating Participants and Stimuli as Random Effects

	No. of stimuli				
No. of participants	10	30	50	70	90
10	.046	.048	.049	.049	.051
30	.047	.047	.052	.044	.052
50	.051	.050	.051	.048	.052
70	.048	.047	.055	.052	.051
90	.053	.050	.046	.049	.049

We repeated these simulations, this time including a true Condition difference, to begin to examine issues of statistical power as a function of varying the numbers of participants and stimuli in this design where participants are crossed with the two levels of Condition and stimuli are nested under them. The true value of the Condition difference was set at 2 units with all other variance components left as they were in the previous simulations. In power analyses one typically expects power to be reported in terms of some standardized true effect size, comparing the magnitude of the treatment difference to an estimate of error. Because of multiple variance components which might all be considered error in this mixed design, there is no way to easily specify a single standardized effect estimate for these power analyses. Accordingly, we provide the value of true Condition difference in absolute units and the various variance components as specified in the power analysis.[12]

Figure 1 plots the empirical statistical power estimates that resulted from these simulations as a function of the number of participants and the number of stimuli. Unsurprisingly, as both the number of participants and the number of stimuli increase, statistical power increases. What is perhaps a bit surprising is that the power benefit of increasing the number of participants seems to be relatively small after 30 while increasing the number of stimuli pays benefits beyond 30 (i.e., beyond 15 per condition). This conclusion is undoubtedly specific to this design, where participants are crossed with condition while stimuli are nested under the two conditions. These power results also depend on the values of the variance components we have used, which admittedly were rather arbitrary. In general, as the random variance components are larger, designs will tend to have less power.

[12] Although there is no commonly agreed-upon effect size estimate in linear mixed models, one can begin to specify the general conditions for estimating an effect size. Basically, one would like to compare the magnitude of a fixed effect estimate with the expected variation in responses in a single condition. This expectation can be estimated, depending on the details of the design, from the estimated random variance components, although the details of doing this are beyond the scope of this article.

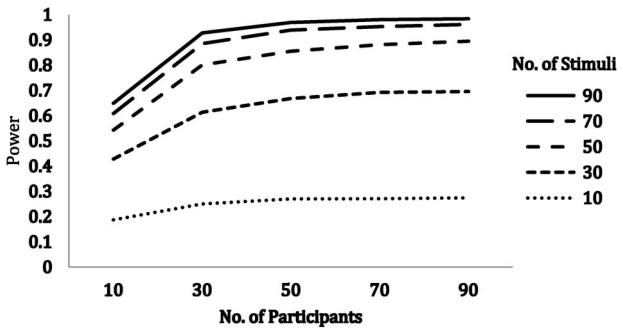

Figure 1. Statistical power levels as a function of the number of participants and the number of stimuli in a design where participants are crossed with Condition and Stimuli are nested under its two levels. (Note: $\sigma_e^2 = 16$, $\sigma_P^2 = \sigma_{P\times C}^2 = \sigma_{PS(C)}^2 = 4$, and the true condition difference is 2 units.)

Brief Consideration of Other Designs

The model specification and power estimates that we have just given are specific to the illustrative research design that we have so far considered, with participants crossed with condition and stimuli nested under its levels. There are many alternative designs that we might have considered. In the paragraphs that follow, we briefly discuss some alternative designs and model specification in these, before turning to the analysis of some actual social psychological data sets.

First, consider a design in which the roles of participants and stimuli are reversed: each stimulus is found once in each condition, and participants are nested under condition. For instance, in Hamilton, Katz, and Leirer (1980), participants encounter a set of stimulus persons and are either given memory or impression formation instructions. Thus, participants are either in one condition or the other (i.e., instructions), and each stimulus person is encountered in both conditions. For this design, the model specification would include random intercepts for participants and both random intercepts and random condition slopes for stimuli, exactly the reverse of the design that we have considered. And the power conclusions we reached about the relative impact of increasing the numbers of participants and stimuli would be exactly reversed.

Second, it might be that both participants and stimuli are crossed with condition. For instance, there is one set of stimuli, and each participant does two different tasks with all such stimuli (e.g., two different judgment tasks where task is the independent variable of interest, i.e., condition). Here the model specification would include random intercepts and random slopes for both participants and stimuli.

A variation on this second design might be that each participant does both tasks, but one set of stimuli is used for one task and a second set is used for the second task, and which stimulus set is used in which task is counterbalanced across participants. In this case, both task and the counterbalancing variable (i.e., which stimulus set is in which task) might be treated as fixed effects. Task varies within participants while the counterbalancing factor varies between them. Both task and the counterbalancing factor vary within stimuli. The model specification would thus include random intercepts and task slopes for participants and random intercepts, task slopes, and counterbalancing slopes for stimuli.

There are a myriad of further designs, increasingly complex, as the number of fixed effects proliferates and as there are multiple stimulus sets. Rather than continue this abstractly, we chose to illustrate further design complexities and model specification with some actual data sets in social psychology.

Mixed Model Illustrations With Actual Data Sets

Our goal in this section of the article is to examine some well-known social cognitive data sets that have been reported in the literature to illustrate the use and advantages of the mixed models approach treating both participants and stimuli as random effects.[13]

"Shooter" Data

The first data set is taken from the "shooter" paradigm that examined whether the race of a target influences the speed with which participants are able to correctly discriminate between targets who are holding a weapon (where the correct response is to "shoot") and targets who are unarmed (where the correct response is "not shoot"; Correll et al., 2002; 2007). In the specific data set examined, 36 participants responded to 100 randomly ordered shooter trials in which there were 25 armed White targets, 25 armed African American targets, 25 unarmed White targets, and 25 unarmed African American targets. Each specific target (of which there were 25 White males and 25 African American males) appeared twice in this sequence, once armed and once unarmed. Thus, participants are crossed with both target race and gun (armed vs. unarmed), whereas targets are crossed with gun but nested under target race.

Analyses were conducted on log-transformed response latencies involving only trials where correct responses were given. In total across all participants, there were $36 \times 100 = 3{,}600$ trials. Of these, correct responses were given on 3,402 trials. Thus, 5.5% of the data were missing from this analysis. The basic shooter hypothesis is that correct responses to stereotype-congruent targets (unarmed White and armed African American) would be faster than correct responses to stereotype-incongruent targets (unarmed African American and armed White). Thus, the prediction is a race of target by gun interaction.

Three different analyses were conducted. First, we treated participant as random but ignored target, analyzing four mean response latencies for each participant (mean latencies to armed and unarmed White and African American targets). This by-participant analysis yielded a significant race of target by gun interaction, $F_{1,\ 35} = 57.89$, $p < .001$. This is the analysis reported in the published "shooter" articles.

The second analysis was the by-stimulus analysis, treating target as random and collapsing across participants. Specifically, for each target (25 Whites and 25 African Americans), two mean response latencies were computed (one for when the target was armed and one for when he was unarmed), averaging across all participants who provided correct responses to each target in each condition. In this analysis, race of target is between targets, and gun is within

[13] The full syntax and output for these analyses are available from the authors upon request.

them. This analysis yielded a marginally significant race of target by gun interaction, $F_{1, 48} = 3.74$, $p = .059$.

The final analysis was based on the mixed models approach, specifying both participants and targets as random, with random error components for the intercept, gun effect, target race effect, and gun by target race interaction for participants and random error components for the intercept and gun effect for targets, as well as the covariances between these random effects. Effect or contrast coding was used to code the fixed factors, gun (.5 = gun; −.5 = no-gun) and race (.5 = black target; −.5 = white target) and the interaction was computed as the product of these codes. This analysis revealed a significant main effect of gun such that responses were faster to armed targets than to unarmed ones (as did the by-participant and the by-stimulus analyses). Additionally there was a marginally significant race of target by gun interaction, in the expected direction with faster responses to stereotype-congruent targets than to stereotype-incongruent targets, $b = -.06$, $F_{1, 48.1} = 3.39$, $p = .072$.

Testing the variance components in this mixed model revealed that the variance of the gun by target race effect across participants was not significantly different from 0, meaning that the shooter effect is basically the same for all participants. This suggests that one might look in vain for individual differences that might moderate the effect. There was, however, considerable random variation between targets in the magnitude of the gun effect, suggesting considerable variation from target to target in latency differences as a function of whether the target is armed. It is largely because of this variation that the by-participant analysis (which ignores this variation) overestimates the significance of the shooter effect.

Afrocentric Features Data

The second data set is taken from a line of work conducted by Blair and colleagues, examining how stereotypic judgments ensue from Afrocentric facial features that vary from target to target even within racial categories (Blair et al., 2005; Blair, Judd, & Chapleau, 2004; Blair, Judd, Sadler, & Jenkins, 2002). The specific data set examined was reported in Blair et al. (2005). Participants in this study made judgments about 64 African American male target individuals. Participants were shown a photograph of a target individual as well as a record of whether that target had acted aggressively in four previous situations. The photographs presented people who varied in the degree to which they possessed Afrocentric facial features (as determined by pretest participants—each photograph was scaled on this variable as the mean judgment provided by pretest participants, $\alpha = .87$). All target individuals had consensually been identified as African American. In addition, the attractiveness of each individual face was also scaled, again as the mean from pretest participants ($\alpha = .93$). The other information provided for each target was a score that varied from 0 to 4, indicating the number of previous situations in which the target had acted aggressively. These previous situations had been described for participants as well as the behaviors that were either aggressive or not. Each participant's task was to judge the probability that each target individual would act aggressively in a fifth situation that had also been described. Individual target photographs were randomly paired with scores on the previous situations individually for each participant. The researchers clearly expected high probability of aggression judgments in the fifth situation for target individuals who had acted more aggressively in the four previous situations. Over and above that, the prediction was that target individuals with more Afrocentric facial features would be seen as having a higher probability of aggressive behavior in the fifth situation. Thus, in these data, the critical independent variable (Afrocentric facial features) varied continuously across targets, as did facial attractiveness and levels of previous aggression that needed to be controlled in the analysis.

The original analysis of these data as reported in Blair et al. (2005) treated participants as random but not targets. Because the independent variable and covariates were continuous, they could not compute means to conduct the by-participant analysis. Instead, they proceeded as follows: First, for every participant, a within-participant model was estimated, regressing probability judgments on three predictor variables: Afrocentric facial features of the targets, facial attractiveness, and the levels of previous aggression attributed to the target. These individual-participant multiple regression models were estimated with the 64 targets as the unit of analysis. From these regression models, one for each participant, the individual slopes for each predictor became the data for further analyses to test whether their means differed significantly from zero on average across participants. One certainly expected a significant average effect of prior levels of aggression on the probability estimates. The more interesting prediction was that over and above this effect and the effect for attractiveness, the average slope for the within-participant effect of Afrocentric features would also differ significantly from zero. The resulting t, with 46 degrees of freedom, equaled 2.53, indicating that targets having more Afrocentric features were given higher probability of aggression judgments, controlling for prior aggression levels and attractiveness.

The mixed models analysis treats both participants and targets as random effects. Each participant by target observation is the unit of analysis, with each row of data indicating the probability of aggression judgment given by that participant to that target, the target's level of prior aggression (from 0 to 4), the target's attractiveness, and the target's Afrocentric facial features. The fixed effects are the intercept, the effect of level of prior aggression, the effect of attractiveness, and the effect of Afrocentric features. Only the intercept in this model varies randomly across targets, whereas the intercept and the slopes of prior aggression, attractiveness, and Afrocentric facial features vary randomly across participants (and these may covary as well). To make the intercept meaningful, all three predictors were centered around their grand means. Accordingly, the intercept equals the mean probability of aggression attributed to each target by each participant, and the variance components associated with the intercept estimate the variability in these means either across targets or participants.

The mixed model estimated five random variance components (intercept across targets, intercept across participants, and three slopes across participants), the covariances between these random effects, and fixed effects for the intercept and each of the three predictors. Examining the tests of the fixed effects, we find unsurprisingly that there is a highly significant effect of prior aggression, $b = 20.49$, $t_{44.9} = 24.92$, $p < .0001$. Controlling for this (and also for facial attractiveness), faces with more Afrocentric features were given higher aggression probability estimates, $b = 0.75$, $t_{51.1} = 2.08$, $p = .0430$. Facial attractiveness did not have a

statistically significant effect on these probability estimates, $b = -0.64$, $t_{50.5} = -0.99$, *ns*.

Turning to the random components of variance, there is significant variability across target faces in the intercepts, meaning that the mean probability estimates varied across target faces, even with their prior aggression, Afrocentric features, and attractiveness controlled. Across participants, the intercepts also showed highly significant variance across participants. Additionally, there was significant variation across participants in the degree to which probability judgments were influenced by prior levels of aggression. Interestingly, variation in the slopes for Afrocentric features was not significant, suggesting no individual differences in the degree to which higher aggression probability estimates were given for targets with more Afrocentric facial features.

"Retroactive Priming" Data

Our third data set is from Bem (2011), a controversial article recently published in this journal that claimed to find evidence for "psi" or premonition of future events. This study has been the subject of much criticism suggesting that the findings are surely nonreplicable, with some critics going so far as to proclaim that the Bem article demonstrates that experimental psychologists in general must abandon their traditional "frequentist" statistics and begin using Bayesian data analysis techniques (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). It is our opinion that there is nothing inherently wrong with standard hypothesis testing if it is based upon correct models (i.e., treating factors as random when they should be). One problem, we suspect, in Bem's (2011) research and elsewhere is that random variation due to stimuli has not been considered and handled in the analyses. We therefore obtained the data from one of the experiments reported by Bem (Experiment 4) to begin to assess whether the findings reported in the original article would continue to be significant when stimuli, as well as participants, were treated as random.

In the experiment, Bem used a variation of the standard evaluative priming procedure, in which participants view a series of target stimuli, each one very briefly preceded by either a positive or negative prime word, and are asked to classify the target stimulus as either "good" or "bad" as quickly as possible. The classic evaluative priming effect is that participants are faster to correctly classify target stimuli when they are preceded by an evaluatively congruent prime compared with an incongruent prime (Fazio, Sanbonmatsu, Powell & Kardes, 1986). In the present experiment by Bem (2011), participants underwent a series of such priming trials, half of which involved the classic or "forward priming" procedure, and half of which reversed the order of primes and targets such that participants responded to each target stimulus *before* encountering the prime, in a "retroactive priming" procedure. The assumption is that a statistically reliable effect of evaluative congruency on response times in the retroactive priming procedure would indicate evidence for psi.

The experiment included 99 participants, each of whom responded to 64 trials, half using the forward priming procedure and half using the retroactive priming procedure. A fixed set of 16 positive and 16 negative photographs were used as the target stimuli in the forward priming condition, and a different fixed set of 16 positive and 16 negative photographs were used in the retroactive condition. Each photograph was preassigned two prime words, one positive and one negative, and each participant viewed either the positive or the negative prime for the associated target at random.

Bem (2011) reported analyses of these data using different transformations of the response latency dependent variable to correct for positive skew (either a log transformation or an inverse transformation) and using different criteria for excluding outlying trials (excluding trials where the response time exceeded either 1,500 ms or 2,500 ms), all with similar results. We focus on only one of these analyses here, that using the inverse transformation and excluding responses that exceeded 2,500 ms. This is because we found that the inverse transformation was generally more successful in correcting positive skew in the model residuals than was the log transformation for these data. Our initial analyses of these data, to replicate Bem's results, treated participants as random but both primes and stimuli as fixed, yielding statistically significant priming effects for both the forward priming trials, $b = 31.4$ ms, $t_{98} = 4.71$, $p < .001$, and the retroactive priming trials, $b = 23.9$ ms, $t_{98} = 2.57$, $p = .012$.

We reanalyzed these data using mixed models with random effects for participants, primes, and targets. Tests of the fixed effects in the models showed that while the priming effect remained statistically intact for the classic or forward priming trials, $b = 34.7$ ms, $t_{46.91} = 3.82$, $p = .0292$, the priming effect was no longer significant for the retroactive priming trials, $b = 11.3$ ms, $t_{27.58} = 1.53$, $p = .136$.

As in earlier data sets, tests of the random variance components in this analysis are of substantive theoretical interest. Our mixed models allowed both priming effects to vary randomly with respect to participants, allowing for individual differences in the magnitude of these effects. We might reasonably expect such variation in the forward priming trials: individuals may differ in the extent to which they are influenced by evaluative primes. However, if the phenomenon of retroactive priming does not in fact exist, then it cannot be that some participants are "better" at it than others. A likelihood ratio test on the random participant priming slope for the forward priming trials verifies that there are systematic individual differences in the tendency to show the classic evaluative priming effect, $\chi^2_1 = 31.33, p < .001$. However, there was not significant variation in the priming effect for the retroactive priming trials, $\chi^2_1 = 1.64, p = .200$.

Conclusion

We began this article by saying that the issue of stimulus sampling in social psychology is an old issue that has reared its head from time to time, only to be generally ignored in the analysis of social psychological data. Our goals in this article have been (a) to highlight, once again, the dangers of implicitly treating stimuli as fixed when they are in fact random, and (b) to show the way toward a new and comprehensive approach for analyzing data with multiple crossed random effects.

In spite of strong warnings in the past that stimuli in social psychological experiments ought to be treated as random, it is rare indeed in the social psychological literature to find analyses that take into account sampling variability of stimuli. As our simulations make clear, in many commonly used designs in social cognitive research, a likely consequence of only treating participants as a random effect is a large inflation of Type I statistical errors,

well above the nominal .05 rate. The reanalyses we have reported for some of Bem's (2011) experimental data, seemingly demonstrating extrasensory perception, are particularly compelling in this regard. As we have said, Bem's work led to critiques of standard statistical practices and calls for increased reliance on Bayesian approaches. We have demonstrated that one problem in Bem's research is that random variation due to stimuli was not considered in the analyses. Once stimuli were treated as random, there remained little evidence for retroactive priming in the experimental data we examined.

Because the literature is filled with designs where stimuli should be treated as random but are not, we as a field are probably faced with many Type 1 errors, leading to persistent failures to replicate effects when different experimenters use different experimental stimuli (Lehrer, 2010). And when experimenters attempt to replicate effects using the same experimental stimuli as in previous work but analyze these data using traditional procedures that ignore random stimulus variation, it can never be clear whether a successful replication indicates a truly reliable treatment effect or merely a consistent bias in the set of experimental stimuli used. Mixed models can give us greater confidence to rule out this second possibility by allowing researchers to quantify and account for random stimulus variation in experimental data.

We have to this point considered only designs involving continuous response variables (e.g., reaction times), but there are many common social psychological paradigms that involve categorical dependent variables, such as the analysis of error rates in the go/no-go association task (Nosek & Banaji, 2001). Data analysis in these paradigms often involves computing statistics from signal detection theory (most commonly the *d*' and *c* statistics) separately for each participant and then submitting these statistics to an analysis of variance, in very much the same way that the by-participant analysis that we discussed previously involves computing and analyzing participant-level mean scores. This widely used procedure for analyzing categorical response variables consequently suffers from many of the same statistical problems as the by-participant analyses that we have discussed at length, as do analyses based more simply on analyzing within-participant proportions (Jaeger, 2008; Rouder et al., 2007). Categorical data of this kind can be handled under the mixed models approach by adopting a logit link function, extending the familiar logistic regression model to include crossed random effects for participants and stimuli.

Although we have emphasized the costs of not treating both participants and stimuli as random, we hope to have also conveyed some of the benefits of using a mixed models approach to data with both factors random. A pronounced benefit is that one can obtain estimates of the various variance components, and these may lead in turn to new insights about factors that might be responsible for unexplained variance in data, either associated with stimuli or participants. Additionally, if one has estimates of the relative magnitude of these variance components, one can begin to figure out the relative power benefits of adding participants versus adding stimuli to a design. In general, it should be the case that as variance components become larger, power benefits will accrue with the inclusion of additional stimuli or participants across which those particular effects vary. So, for instance, in our classic design, if one believed that there was more intercept variance due to stimuli than variance in condition slopes due to participants, one would be better off increasing the number of stimuli than the number of participants. Full details of this, however, await further work that more adequately evaluates power for a range of research designs, varying the magnitude of the various relevant variance components. And this further work needs to develop more precisely appropriate measures of effect size to permit meta-analytic integration of results across studies that utilize mixed models analyses.

Thinking about these models clarifies often-perplexing issues about the sampling of stimuli in studies such as those we have discussed. Should one ensure considerable stimulus variability or should one attempt to have stimuli that resemble each other as closely as possible? In the shooter data set, the targets were chosen so they varied considerably, and this was considered a strength of the design. On the other hand, in the Afrocentric features data, the targets were all of similar age, dress, and so on. This difference explains why the by-participant analysis was much more biased in the case of the shooter data set than in the Afrocentric features data set. At the same time, however, the wider stimulus sampling in the shooter data set permits us to further explore stimulus characteristics that may explain the variability of the armed versus unarmed difference from target to target in that data set. As in sampling participants, sampling less variable stimuli may lead to power benefits, but more narrowly defined samples of stimuli also mean that one is unable to identify significant further moderators of effects of theoretical interest and that the conclusions make reference only to a narrower range of stimuli.

It is exciting to point to new methods that permit one to include multiple random factors in the analysis of social psychological data. Unlike older methods that were suggested but rarely implemented, mixed models permit estimation of appropriate models in the presence of missing data, nonorthogonal factors, and independent variables that are not categorical. The mixed models approach that we have outlined, based in large part on recent work by Baayen et al. (2008) and others, seems very promising indeed. It is our hope that researchers in social psychology will now realize the importance and benefits of treating stimuli as random and will begin to implement the sort of analyses we have outlined.

References

Alnosaier, W. (2007). *Kenward–Roger approximate F test for fixed effects in mixed linear models* (Unpublished doctoral dissertation). Oregon State University, Corvallis, OR.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge, England: Cambridge University Press.

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

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes* (R package version 0.999375–39). Retrieved from http://CRAN.R-project.org/package=lme4

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100,* 407–425. doi:10.1037/a0021524

Blair, I. V., Chapleau, K. M., & Judd, C. M. (2005). The use of Afrocentric features as cues for judgment in the presence of diagnostic information. *European Journal of Social Psychology, 35,* 59–68. doi:10.1002/ejsp.232

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of

Afrocentric facial features in criminal sentencing. *Psychological Science, 15,* 674–679. doi:10.1111/j.0956-7976.2004.00739.x

Blair, I. V., Judd, C. M., Sadler, M. S., & Jenkins, C. (2002). The role of Afrocentric features in person perception: Judging by features and categories. *Journal of Personality and Social Psychology, 83,* 5–25. doi:10.1037/0022-3514.83.1.5

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

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

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

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83,* 1314–1329. doi:10.1037/0022-3514.83.6.1314

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology, 92,* 1006–1023. doi:10.1037/0022-3514.92.6.1006

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50,* 229–238. doi:10.1037/0022-3514.50.2.229

Fears, T. R., Benichou, J., & Gail, M. H. (1996). A reminder of the fallibility of the Wald Statistic. *The American Statistician, 50,* 226–227. doi:10.2307/2684659

Forster, K. I., & Dickenson, R. G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F_1, F_2, F', and *min F'*. *Journal of Verbal Learning and Verbal Behavior, 15,* 135–142. doi:10.1016/0022-5371(76)90014-1

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, England: Cambridge University Press.

Green, B. F., & Tukey, J. W. (1960). Complex analyses of variance: General problems. *Psychometrika, 25,* 127–152. doi:10.1007/BF02288577

Halekoh, U., & Højsgaard, S. (2011). *pbkrtest: Parametric bootstrap Kenward–Roger based methods for mixed model comparison* (R package version 0.1.3). Retrieved from http://CRAN.R-project.org/package=pbkrtest

Hamilton, D. L., Katz, L. B., & Leirer, V. O. (1980). Cognitive representation of personality impressions: Organizational processes in first impression formation. *Journal of Personality and Social Psychology, 39,* 1050–1063. doi:10.1037/h0077711

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Erlbaum.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434–446. doi:10.1016/j.jml.2007.11.007

Kenny, D. A. (1985). Quantitative methods for social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 1, 3rd ed., pp. 487–508). New York, NY: Random House.

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 1, 4th ed., pp. 233–268). New York, NY: McGraw–Hill.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53,* 983–997. doi:10.2307/2533558

Lehrer, J. (2010, December 13). The truth wears off: Is there something wrong with the scientific method? *The New Yorker*. Retrieved from http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition, 19,* 625–666. doi:10.1521/soco.19.6.625.20886

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effect models in S and S-PLUS.* New York, NY: Springer.

Raaijmakers, J. G. W. (2003). A further look at the "language-as-fixed-effect fallacy" [Special issue]. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 57,* 141–151. doi:10.1037/h0087421

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language, 41,* 416–426. doi:10.1006/jmla.1999.2650

Raudenbusch, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Rietveld, T., & van Hout, R. (2007). Analysis of variance for repeated measures designs with words as a nested random or fixed factor. *Behavior Research Methods, 39,* 735–747. doi:10.3758/BF03192964

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika, 72,* 621–642. doi:10.1007/s11336-005-1350-6

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models using SAS Proc MIXED. *Journal of Agricultural, Biological, and Environmental Statistics, 7,* 512–524. doi:10.1198/108571102726

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford University Press.

Snijders, T., & Bosker, R. (1999). *An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage.

Spilke, J., Piepho, H.-P., & Hu, X. (2005). A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. *Journal of Agricultural, Biological, and Environmental Statistics, 10,* 374–389. doi:10.1198/108571105X58199

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100,* 426–432. doi:10.1037/a0022790

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

Wickens, T. D., & Keppel, G. (1983). On the choice of design and of test statistics in the analysis of experiments with sampled materials. *Journal of Verbal Learning and Verbal Behavior, 22,* 296–309. doi:10.1016/S0022-5371(83)90208-6

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

(*Appendix follows*)

Appendix

Mixed Model Analyses of the Illustrative Dataset in R, SAS, and SPSS

R

Linear mixed effects models can be analyzed in R using the lmer() function from the *lme4* package (Bates, Maechler, & Bolker, 2011), which uses optimization methods specifically suited for fitting models that include crossed random effects. We first assume that the data set is loaded in R in the "long" format (i.e., 900 rows of data, one for each of 30 participant ratings of each of 30 stimuli). Each row of data indicates the response or outcome variable (*y*), the stimulus (*j*), the participant (*i*), and the condition. We assume the data set is stored in a data frame object named `dat`. We can then fit the initial model by using the following commands:

```
> library(lme4)
> model_1 <- lmer(y ~ c + (1 | j) + (c | i),
data=dat)
```

Note that the > symbol at the start of each line simply indicates the input prompt and is not an actual command entered by the user. The first input line says to load the lme4 package. This is necessary to access the functions of lme4 and assumes that lme4 has already been installed. If lme4 has not been installed, this can be done very simply from the input prompt by calling the `install.packages()` function. The second input line says to create a new object, called `model_1` and to assign to that object the output of our call to the `lmer()` function. Within the call to `lmer()`, we specify the response variable on the left side of the ~ operator, followed by the fixed effects, followed by the random effects. For each random effect, we must specify the grouping factor to which the effect is random. We do this for each random effect by encapsulating the effect in parentheses and then indicating the effect on the left side of the | symbol and the grouping factor on the right side. Thus, our first random effect, (`1 | j`), indicates that the intercept (denoted with a "1") is random with respect to stimuli (*j*). When a *slope* is given as random, `lmer()` assumes, unless explicitly told otherwise, that the user also wishes to add random intercepts for the grouping factor in question, as well as to estimate the covariance between the random effects. This is a reasonable assumption because it rarely makes sense to allow random slopes but not random intercepts for a given grouping factor. So our second random effect, (c | i), indicates that both the intercept and the Condition slope are random with respect to participant (*i*) and that a covariance between the two effects should be estimated. Finally, we indicate that our data set is contained in a data frame object named dat.

To view the results, we can simply enter the name of the fitted model object as a command. The resulting output is contained in Figure A1.

It is noteworthy that the fixed effects summaries obtained from models fit by lmer() include parameter estimates, standard errors, and *t* statistics, but no degrees of freedom or *p*-values. As Baayen, Davidson, and Bates (2008) discussed at some length, this is a deliberate choice reflecting the fact that it is not obvious how these quantities should be computed in the context of mixed effects models. One method for obtaining *p* values in the face of these conceptual difficulties, which seems to work well in many situations, is to rely on the Kenward–Roger approximation (Kenward & Roger, 1997; see also Alnosaier, 2007; Schaalje, McBride, & Fellingham, 2002, Spilke, Piepho, & Hu, 2005;). This procedure is a modification of the Satterthwaite approximation; it differs in that in some cases it will rescale the *F* ratio in addition to computing the quantity for degrees of freedom that results in a better approximation to an appropriate *F* distribution. The *F* tests with the Kenward–Roger approximation can be conducted in R using the `KRmodcomp()` function from the *pbkrtest* package (Halekoh & Højsgaard, 2011), which compares two nested models. In the following syntax, we first specify a model, `model_2`, which has the same random effects structure as `model_1` but which excludes Condition from the fixed effects specification (again, a "1" indicates the intercept, which is implicit in the fixed effects specification for `model_1`). We then compare the two models using `KRmodcomp()` to obtain the estimated effect of Condition under the Kenward–Roger approximation (see Figure A2).

```
Linear mixed model fit by REML
Formula: y ~ c + (1 | j) + (c | i)
   Data: dat
  AIC  BIC logLik deviance REMLdev
 5193 5227  -2590     5181    5179
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 j        (Intercept)  3.6703  1.9158
 i        (Intercept)  4.2940  2.0722
          c            4.1822  2.0450   0.271
 Residual             15.5573  3.9443
Number of obs: 900, groups: j, 30; i, 30

Fixed effects:
            Estimate Std. Error t value
(Intercept)  -0.1804     0.5318  -0.339
c             2.5211     0.8354   3.018

Correlation of Fixed Effects:
  (Intr)
c 0.086
```

Figure A1. Output for mixed models estimation in R.

SAS

In SAS, one uses *PROC MIXED* to estimate mixed models. The following syntax estimates the full model for these data, allowing random variance components for the intercept with respect to *j*, the intercept and *c* slope with respect to *i,* and the covariance between these latter two:

(*Appendix continues*)

```
> library(pbkrtest)
> model2 <- lmer(y ~ 1 + (1 | j) + (c | i), data=dat)
> KRmodcomp(model1,model2)
F-test with Kenward-Roger approximation; computing time: 11.35 sec.
Large : y ~ c + (1 | j) + (c | i)
small : y ~ 1 + (1 | j) + (c | i)
    Fstat df1      df2   p.value F.scaling
 9.106821   1 38.51768 0.0044978         1
```

Figure A2. Input and output for Kenward–Roger test in R.

```
proc mixed covtest;
class i j;
model y=c/solution ddfm=kr;
random intercept c/sub=i type=un;
random intercept/sub=j;
run;
```

The *covtest* option asks SAS to provide tests of the various random effects (albeit not the likelihood ratio chi-square tests that we have recommended). In the *model* statement, the *solution* option specifies that the fixed effects parameter estimates be printed (otherwise only their associated *F*s are output) and the *ddfm=kr* option specifies that the Kenward–Roger approximation be used for the degrees of freedom in testing the fixed effects. The two *random* statements specify the random effects in the model, first the intercept and the *c* slope with respect to participant (*i*) and then the intercept only with respect to stimuli (*j*). The *type=un* option directs SAS to estimate the covariance between the two random components with respect to *i.*

The resulting output (which we have edited to keep things short) is given in Figure A3.

The variances of the random components are given as *UN(1,1) i* for the intercept with respect to *i*, *UN(2,2) i* for the slope with respect to *i,* and *Intercept j* for the intercept with respect to *j. UN(2,1) i* is the covariance between the *i* intercepts and slopes. Each of these variance components is tested by computing an estimated standard error and an approximate *Z* statistic. When sample sizes are sufficiently large, these tests will give results that are quite close to the likelihood-ratio chi-square difference tests that we and others prefer (Fears, Benichou, & Gail, 1996). At the bottom of the output, the fixed effects parameter estimates and associated *t* statistics are given, along with the Kenward–Roger degrees of freedom approximation.

SPSS

Mixed effects models are estimated in SPSS using the MIXED procedure. We can use the following set of commands to estimate the full model that allows covariance between the random effects for *i*.

```
MIXED y WITH c
/FIXED=c
/PRINT=SOLUTION TESTCOV
/RANDOM=INTERCEPT c | SUBJECT(i) COVTYPE(UN)
/RANDOM=INTERCEPT | SUBJECT(j).
```

The first line indicates the dependent and independent variables. Note that we use `WITH` rather than `BY` because `c` is precoded at −0.5 versus +0.5. We therefore tell SPSS to treat this variable as continuous so that it does not attempt to recode c. In the next lines, we specify c as a fixed effect and then request that the parameter estimates for the fixed effects and tests of the random covariance parameters be printed with the output. In the final two lines, we specify the intercept and the Condition effect both as being random with respect to *i* and the intercept additionally as being random with respect to *j*. For both of these grouping factors, we specify an unstructured covariance matrix so that all possible random effect covariances are estimated.

```
The Mixed Procedure

      Iteration History
      Iteration    Evaluations    -2 Res Log Like       Criterion
       0              1      5420.61838245
       1              1      5179.04463914      0.00000000
           Convergence criteria met.

               Covariance Parameter Estimates
                                   Standard         Z
Cov Parm      Subject    Estimate       Error     Value        Pr Z

UN(1,1)       i          4.2938      1.2641      3.40      0.0003
UN(2,1)       i          1.1464      1.0409      1.10      0.2707
UN(2,2)       i          4.1821      1.6462      2.54      0.0055
Intercept     j          3.6703      1.1198      3.28      0.0005
Residual                15.5573      0.7721     20.15      <.0001

          Fit Statistics
-2 Res Log Likelihood          5179.0
AIC (smaller is better)        5189.0
AICC (smaller is better)       5189.1
BIC (smaller is better)        5179.0

                  Solution for Fixed Effects
                        Standard
Effect       Estimate       Error      DF    t Value    Pr > |t|

Intercept     -0.1804      0.5317    50.5      -0.34      0.7358
c              2.5211      0.8354    38.5       3.02      0.0045
```

Figure A3. Output for mixed models estimation in SAS.

(*Appendix continues*)

Information Criteria

-2 Restricted Log Likelihood	5179.045
Akaike's Information Criterion (AIC)	5189.045
Hurvich and Tsai's Criterion (AICC)	5189.112
Bozdogan's Criterion (CAIC)	5218.045
Schwarz's Bayesian Criterion (BIC)	5213.045

Estimates of Fixed Effects

						95% Confidence Interval	
Parameter	Estimate	Std. Error	df	t	Sig.	Lower Bound	Upper Bound
Intercept	-.180422	.531749	50.475	-.339	.736	-1.248222	.887378
c	2.521067	.835418	38.516	3.018	.004	.830596	4.211538

Estimates of Covariance Parameters

			Std.			95% Confidence Interval	
Parameter		Estimate	Error	Wald Z	Sig.	Lower Bound	Upper Bound
Residual		15.557	.772	20.149	.000	14.115304	17.146643
Intercept + c [subject = i]	UN (1,1)	4.294	1.264	3.397	.001	2.411326	7.645878
	UN (2,1)	1.146	1.041	1.101	.271	-.893741	3.186625
	UN (2,2)	4.182	1.646	2.540	.011	1.933423	9.046049
Intercept [subject = j]	Variance	3.670	1.120	3.278	.001	2.018369	6.674347

Figure A4. Output for mixed models estimation in SPSS.

These commands give rise to the SPSS output given in Figure A4.

These results closely match those obtained with R and SAS. Note that the degrees of freedom reported in the fixed effects summary by SPSS are under the Satterthwaite approximation. The Kenward–Roger is generally slightly favored over the Satterthwaite; however, in many circumstances, as in the present example, the two methods yield identical or nearly identical results.

Received October 12, 2011
Revision received March 6, 2012
Accepted March 7, 2012 ■