Using Confidence Intervals for Graphically Based Data Interpretation

MICHAEL E. J. MASSON, *University of Victoria*
GEOFFREY R. LOFTUS, *University of Washington*

Abstract As a potential alternative to standard null hypothesis significance testing, we describe methods for graphical presentation of data—particularly condition means and their corresponding confidence intervals—for a wide range of factorial designs used in experimental psychology. We describe and illustrate confidence intervals specifically appropriate for between-subject versus within-subject factors. For designs involving more than two levels of a factor, we describe the use of *contrasts* for graphical illustration of theoretically meaningful components of main effects and interactions. These graphical techniques lend themselves to a natural and straightforward assessment of statistical power.

Null hypothesis significance testing (NHST), although hotly debated in the psychological literature on statistical analysis (e.g., Chow, 1998; Cohen, 1990, 1994; Hagen, 1997; Hunter, 1997; Lewandowsky & Maybery, 1998; Loftus, 1991, 1993, 1996, 2002; Schmidt, 1996), is not likely to go away any time soon (Krueger, 2001). Generations of students from multiple disciplines continue to be schooled in the NHST approach to interpreting empirical data, and practicing scientists rely almost reflexively on the logic and methods associated with it. Our goal here is not to extend this debate, but rather to enhance understanding of a particular alternative to NHST for interpreting data. In our view, to the extent that a variety of informative means of constructing inferences from data are made available and clearly understood, researchers will increase their likelihood of forming appropriate conclusions and communicating effectively with their audiences.

A number of years ago, we advocated and described computational approaches to the use of confidence intervals as part of a graphical approach to data interpretation (Loftus & Masson, 1994; see also, Loftus, 2002). The power and effectiveness of graphical data presentation is undeniable (Tufte, 1983) and is common in all forms of scientific communication in experimental psychology and in other fields. In many instances, however, plots of descriptive statistics (typically means) are not accompanied by any indication of variability or stability associated with those descriptive statistics. The diligent reader, then, is forced to refer to a dreary accompanying recital of significance tests to determine how the pattern of means should be interpreted.

It has become clear through interactions with colleagues and from queries we have received about the use of confidence intervals in conjunction with graphical presentation of data, that more information is needed about practical, computational steps involved in generating confidence intervals, particularly with respect to designs involving interactions among variables. In this article, we briefly explain the logic behind confidence intervals for both between-subject and within-subject designs, then move to a consideration of a range of multifactor designs wherein interaction effects are of interest. Methods for computing and displaying confidence intervals for a variety of between-subject, within-subject, and mixed designs commonly used in experimental psychology are illustrated with hypothetical data sets. These descriptions go beyond the range of experimental designs considered in Loftus and Masson (1994). Moreover, we extend the use of contrasts discussed by Loftus (2002) and present a method for using

Preparation of this report was supported by Discovery Grant A7910 from the Natural Sciences and Engineering Research Council of Canada to Michael Masson and by U.S. National Institute of Mental Health grant MH41637 to Geoffrey Loftus. We thank Stephen Lindsay, Raymond Nickerson, and Warren Tryon for helpful comments on an earlier version of this paper.

Correspondence to: Michael Masson, Department of Psychology, University of Victoria, PO Box 3050 STN CSC, Victoria BC V8W 3P5, Canada. e-mail: mmasson@uvic.ca

planned contrasts to examine theoretically motivated effects generated by factorial designs. Finally, we consider an additional, crucial advantage of this graphical approach to data interpretation that is sorely lacking in standard applications of NHST, namely, the ease with which one can assess an experiment's statistical power.

Interpretation of Confidence Intervals

The formal interpretation of a confidence interval associated with a sample mean is based on the hypothetical situation in which many random samples are drawn from a population. For each such sample, the mean, standard deviation, and sample size are used to construct a confidence interval representing a specified degree of confidence, say 95%. Thus, for each sample we have

$$95\%CI = M \pm SE_M\,(t_{95\%}) \qquad (1)$$

Under these conditions, it is expected that 95% of these sample-specific confidence intervals will include the population mean. In practical situations, however, we typically have only one sample from a specified population (e.g., an experimental condition) and therefore the interpretation of the confidence interval constructed around that specific mean would be that there is a 95% probability that the interval is one of the 95% of all possible confidence intervals that includes the population mean. Put more simply, in the absence of any other information, there is a 95% probability that the obtained confidence interval includes the population mean.

The goal of designing a sensitive experiment is to obtain precise and reliable measurements that are contaminated by as little measurement error as possible. To the extent that a researcher accomplishes this goal, the confidence intervals constructed around sample means will be relatively small, allowing the researcher accurately to infer the corresponding pattern of population means. That is, inferences about patterns of population means and the relations among these means can be derived from the differences among sample means, relative to the size of their associated confidence intervals.

Confidence Intervals for Between-Subject Designs

The construction and interpretation of confidence intervals is most directly appropriate to designs in which independent groups of subjects are assigned to conditions—the between-subject design. To illustrate the use of confidence intervals in this context, consider a study in which different groups of subjects are assigned to different conditions in a study of selective attention involving Stroop stimuli. Admittedly, this is the kind of experiment more likely to be conducted using a repeated-measures or within-subject design, but we will carry this example over to that context below. Assume that one group of subjects is shown a series of color words (e.g., *blue*, *green*), each appearing in an incongruent color (e.g., the word *blue* printed in the color green). The task is to name the color as quickly as possible. A second group is shown a series of color words, each printed in a congruent color (e.g., the word *blue* printed in the color blue), and a third group is shown a series of consonant strings (e.g., *kfgh*, *trnds*, etc.), each printed in one of the target colors.

A hypothetical mean response time (RT) for each of 24 subjects (8 per group) is shown in Table 1. The group means are plotted in the top panel of Figure 1; note that the individual subject means are displayed around each group's mean to provide an indication of inter-subject variability around the means. The means could also be plotted in a more typical manner, that is, with a confidence interval shown for each mean. There are two ways to compute such confidence intervals in a between-subject design such as this one, depending on how variability between subjects is estimated. One approach would be to compute SE_M independently for group, based on seven degrees of freedom in each case, then construct the confidence intervals using Equation 1. An alternative and more powerful approach is one that is also the foundation for the computation of analysis of variance (ANOVA) for designs such as this one: A pooled estimate of between-subject variability is computed across all three groups, in exactly the same way as one would compute MS_{Within} for an ANOVA.

Pooling of the different estimates of between-subject variability provides a more stable estimate of variability and delivers a larger number of degrees of freedom. The advantage of having more degrees of freedom is that a smaller *t*-ratio is used in computing the confidence interval; the disadvantage is that it requires the

TABLE 1
Hypothetical Subject Means for a Between-Subject Design

Stimulus Type		
Incongruent	Congruent	Neutral
784	632	651
853	702	689
622	598	606
954	873	855
634	600	595
751	729	740
918	877	893
894	801	822
$M_1 = 801.2$	$M_2 = 726.5$	$M_3 = 731.4$

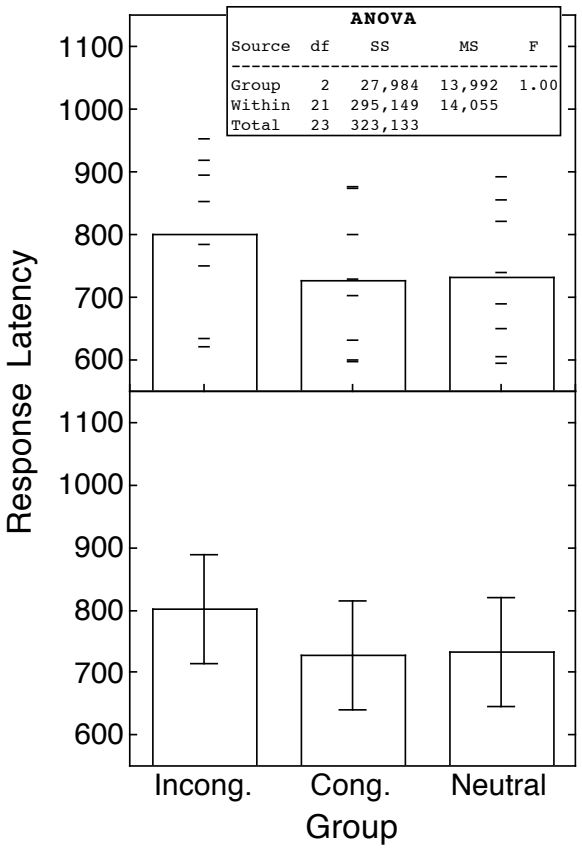

Figure 1. Group means plotted with raw data (top panel) and with confidence intervals (bottom panel) for the hypothetical data shown in Table 1. The ANOVA for these data is shown in the top panel

homogeneity of variance assumption, i.e., the assumption that the population variance is the same in all groups (we return to this issue below). In any event, using the pooled estimate of variability results in the following general equation for confidence intervals in the between-subject design:

$$CI = M_j \pm \sqrt{\frac{MS_{Within}}{n_j}}\ (t_{critical}) \qquad (2)$$

where M_j is the mean for and n_j is the number of subjects in Group *j*. Note that when the *n*'s in the different groups are the same, as is true in our example, a single, common confidence interval can be produced and plotted around each of the group means.

To produce a graphic presentation of the means from this hypothetical study, then, we can compute MS_{Within} using an ANOVA program, then construct the confidence interval to be used with each mean using Equation 2. The MS_{Within} for these data is 14,054 (see Figure 1). With 21 degrees of freedom, the critical *t*-ratio for a 95% confidence interval is 2.080 and $n = 8$, so the confidence interval is computed from Equation 2 to be ±87.18. The resulting plot of the three group means and their associated 95% confidence interval is shown in the lower panel of Figure 1. It is clear from the size of the confidence interval that these data do not imply strong differences between the three groups. Indeed, the ANOVA computed for the purpose of obtaining MS_{Within} generated an *F*-ratio of 1.00, clearly not significant by conventional NHST standards.

Confidence Intervals for Within-Subject Designs

To illustrate the logic behind confidence intervals for means obtained in within-subject designs, we can again consider the data from Table 1, but now treat them as being generated by a within-subject design, as shown in the left side of Table 2. Thus, there are eight subjects, each tested under the three different conditions. The raw data and condition means are plotted in Figure 2, which also includes lines connecting the three scores for each subject to highlight the degree of consistency, from subject to subject, in the pattern of scores across conditions. The ANOVA inset in the top panel of Figure 2 shows that with this design, the data produce a

TABLE 2
Hypothetical Subject Means for a Within-Subject Design

	Raw data				Normalized data			
Subject	Incong.	Cong.	Neutral	*Mi*	Incong.	Cong.	Neutral	*Mi*
1	784	632	651	689.0	848	696	715	753.0
2	853	702	689	748.0	858	707	694	753.0
3	622	598	606	608.7	766.3	742.3	750.3	753.0
4	954	873	855	894.0	813	732	714	753.0
5	634	600	595	609.7	777.3	743.3	738.3	753.0
6	751	729	740	740.0	764	742	753	753.0
7	918	877	893	896.0	775	734	750	753.0
8	894	801	822	839.0	808	715	736	753.0
M_i	801.2	726.5	731.4		801.2	726.5	731.3	

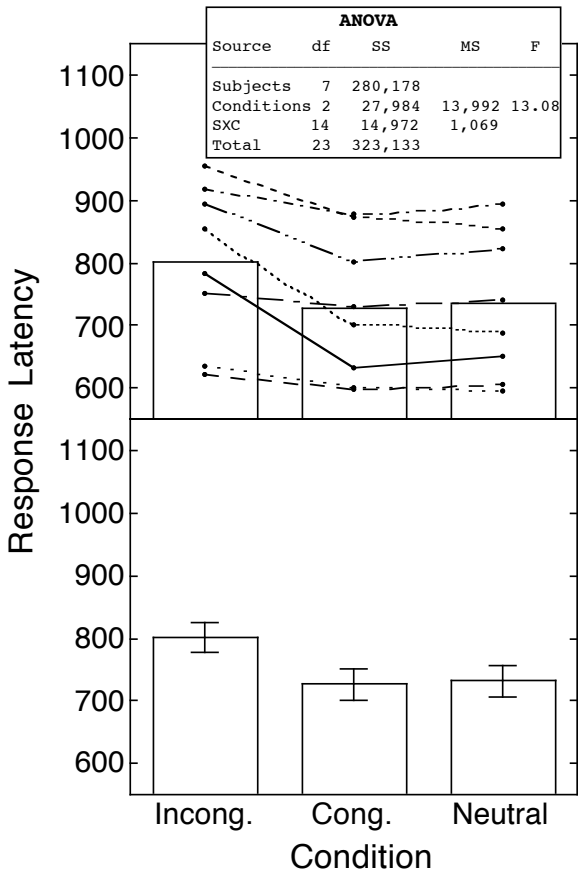

Figure 2. Condition means plotted with raw data (top panel) and with confidence intervals (bottom panel) for the hypothetical data shown in Table 2. The ANOVA for these data is shown in the top panel.

clearly significant pattern of differences between conditions under the usual NHST procedures.

The consistency of the pattern of scores across conditions is captured by the error term for the *F*-ratio, MS_{SXC} (1,069 in our example). To the extent that the pattern of differences is similar or consistent across subjects, that error term, i.e., the Subjects x Conditions interaction, will be small. Moreover, the error term does not include any influence of differences between subjects, as the ANOVA in the top panel of Figure 2 indicates—the between-subject variability is partitioned from the components involved in computation of the *F*-ratio. It is this exclusion of between-subject variability that lends power to the within-subject design (and led, as shown in the upper panel of Figure 2, to a significant effect for a set of data that failed to generate a significant effect when analyzed as a between-subject design).

Now consider how this concept applies to the construction of confidence intervals for a within-subject design. Because the between-subject variability is not relevant to our evaluation of the pattern of means in a within-subject design, we will, for illustrative purposes, remove its influence before establishing confidence intervals. By excluding this component of variability, of course, the resulting confidence interval will not have the usual meaning associated with confidence intervals, namely, an interval that has a designated probability of containing the true population mean. Nevertheless, a confidence interval of this sort will allow an observer to judge the reliability of the *pattern* of sample means as an estimate of the corresponding pattern of population means. The size of the confidence interval provides information about the amount of statistical noise that obscures the conclusions that one can draw about a pattern of means.

The elimination of the influence of between-subject variability, which is automatically carried out in the course of computing a standard, within-subjects ANOVA, can be illustrated by *normalizing* the scores for each subject. Normalization is based on the deviation between a subject's overall mean, computed across that subject's scores in each condition, and the grand mean for the entire sample of subjects (753 in the example in Table 2). That deviation is subtracted from the subject's score in each condition (i.e., $X_{ij} - (M_i - GM)$) to yield a normalized score for that subject in each condition, as shown in the right side of Table 2. Note that, algebraically, the normalized scores produce the same condition means as the raw scores. Also, each subject's pattern of scores remains unchanged (e.g., the Incongruent-Congruent difference for subject 1 is 152 ms in both the raw data and in the normalized data). The only consequence of this normalization is to equate each subject's overall mean to the sample's grand mean, thereby eliminating between-subject variability, while leaving between-condition and interaction variability unchanged.

As Loftus and Masson (1994, p. 489) showed, the variability among the entire set of normalized scores consists of only two components: variability due to Conditions and variability due to the Subjects x Conditions interaction (i.e., error). The variability among normalized scores within each condition can be pooled (assuming homogeneity of variance), just as in the case of the between-subject design, to generate an estimate of the consistency of differences among conditions across subjects—the Subjects x Conditions interaction. Thus, the equation for the construction of a within-subject confidence interval is founded on the mean square error for that interaction (see Loftus & Masson, 1994, Appendix A(3), for the relevant proof):

$$CI = M_j \pm \sqrt{\frac{MS_{SXC}}{n}}\ (t_{critical}) \qquad (3)$$

where n is the number of observations associated with each mean (8 in this example) and the degrees of freedom for the critical t-ratio is df_{SXC}, the degrees of freedom for the interaction effect (14 in this example). Based on the ANOVA shown in the top panel of Figure 2, then, the 95% confidence interval for the pattern of means in this within-subject design is ±24.80, which is considerably smaller than the corresponding interval based on treating the data as a between-subject design. The condition means are plotted with this revised confidence interval for the within-subject design in the bottom panel of Figure 2. The clear intuition one gets from inspecting Figure 2 is that there are differences among conditions (consistent with the outcome of the ANOVA shown in Figure 2), specifically between the incongruent condition and the other two conditions, implying that an incongruent color-word combination slows responding relative to a neutral condition, but a congruent pairing generates little or no benefit.

INFERENCES ABOUT PATTERNS OF MEANS

We emphasize that confidence intervals constructed for within-subject designs can support inferences only about *patterns of means across conditions*, not inferences regarding the value of a particular population mean. That latter type of inference *can*, of course, be made when constructing confidence intervals in between-subject designs. But in most experimental research, interest lies in *patterns*, rather than absolute values of means, so the within-subject confidence interval defined here is well-suited to the purpose and as long as the type of confidence interval plotted is clearly identified, no confusion should arise (cf. Estes, 1997).

Our emphasis on using confidence intervals to interpret patterns of means should be distinguished from standard applications of NHST. We advocate the idea of using graphical display of data with confidence intervals as an alternative to the NHST system, and particularly that system's emphases on binary (reject, do not reject) decisions and on showing what is not true (i.e., the null hypothesis). Rather, the concept of interpreting a pattern of means emphasizes what is true (how the values of means are related to one another), tempered by a consideration of the statistical error present in the data set and as reflected in the size of the confidence interval associated with each mean. In emphasizing the interpretation of a pattern of means, rather than using graphical displays of data as an alternative route to making binary decisions about null hypotheses, our approach is rather different from that taken by, for example, Goldstein and Healy (1995) and by Tryon (2001). These authors advocate a version of confidence intervals that support testing null hypotheses about pairs of conditions. For example, Tryon advocates the use of *inferential confidence intervals*, which are defined so that a statistical difference between two means can be established (i.e., the null hypothesis can be rejected) if the confidence intervals associated with the means do not overlap.

Because our primary aim is not to support the continued interpretation of data within the NHST framework, we have not adopted Tryon's (2001) style of confidence interval construction. Nevertheless, there is a relatively simple correspondence between confidence intervals as we define them here and whether there is a statistically significant difference between, say, a pair of means, according to a NHST-based test. Loftus and Masson (1994, Appendix A(3)) showed that two means will be significantly different by ANOVA or t-test if and only if the absolute difference between means is at least as large as $\sqrt{2} \times CI$, where CI is the 100(1-α)% confidence interval. Thus, as a rule of thumb, plotted means whose confidence intervals overlap by no more than about half the distance of one side of an interval can be deemed to differ under NHST.[1] Again, however, we emphasize that our objective is not to offer a graphical implementation of NHST. Rather, this general heuristic is offered only as an aid to the interested reader in understanding the conceptual relationship between the confidence intervals we describe here and NHST procedures.

ASSUMPTIONS

For both the between- and within-subject cases, computation of confidence intervals based on pooled estimates of variability relies on the assumption that variability is equal across conditions—the homogeneity of variance assumption in between-subject designs and the sphericity assumption for within-subject designs (i.e., homogeneity of variance and covariance). For between-subject designs, if there is concern that the homogeneity assumption has been violated (e.g., if group variances differ from one another by a factor or two or more), a viable solution is to use Equation 1 to compute a confidence interval for each group, based only on the scores with that group. This approach will result in confidence intervals of varying size, but that will not interfere with interpreting the pattern of means, nor is it problematic in any other way.

For within-subject designs, standard ANOVA programs provide tests of the sphericity assumption, often

[1] We note that in their demonstration of this point in their Appendix A3, Loftus and Masson (1994) made a typographical error at the end of this section of their appendix (p. 489), identifying the factor as 2 rather than as $\sqrt{2}$.

by computing a value, ε, as part of the Greenhouse-Geisser and Huynh-Feldt procedures for correcting degrees of freedom under violation of sphericity. The ε value reflects the degree of violation of that assumption (lower values indicate more violation). Loftus and Masson (1994) suggest not basing confidence intervals on the omnibus MS_{SXC} estimate when the value of ε falls below 0.75. Under this circumstance, one approach is to compute separate confidence intervals for each condition, as can be done in a between-subject design. This solution, however, is associated with a potential estimation problem in which variance estimates are computed from the difference between two mean-squares values and therefore a negative variance estimate may result (Loftus & Masson, 1994, p. 490). To avoid this problem, we recommend a different approach, whereby confidence intervals are constructed on the basis of specific, single-degree of freedom contrasts that are of theoretical interest. In the example from Table 2, for instance, the original ANOVA produced $\varepsilon < 0.6$, indicating a violation of sphericity. Here, one might want to test for a standard Stroop effect by comparing the incongruent and congruent conditions and also test for a possible effect of Stroop facilitation by comparing the congruent and neutral conditions. This approach would entail computing an ANOVA for each comparison and using the MS_{SXC} term for the contrast-specific ANOVA as the basis for the corresponding confidence interval. Note that the degrees of freedom associated with the *t*-ratio used to construct these intervals are much less than when the omnibus MS_{SXC} was used. Thus, for the comparison between the incongruent and congruent conditions, MS_{SXC} = 1,451, with seven degrees of freedom, so the 95% confidence interval is

$$95\%CI = \pm\sqrt{\frac{1451}{8}}\,(2.365) = \pm 31.85\ .$$

Similarly, the 95% confidence interval for the congruent-neutral contrast is ±8.85, based on a MS_{SXC} of 112. Thus, there are two different confidence intervals associated with the congruent condition. One possible way of plotting these two intervals is shown in Figure 3, in which the mean for the congruent condition is plotted with two different confidence intervals. Interpreting patterns of means is restricted in this case to pairs of means that share a common confidence interval. Note that the difference in magnitude of these two intervals is informative with respect to the degree of consistency of scores across conditions for each of these contrasts, in keeping with what can be observed from the raw data plotted in the upper panel of Figure 2.

Multifactor Designs

Experimental designs involving factorial combination of multiple independent variables call for some inventiveness when it comes to graphical presentation of data. But with guidance from theoretically motivated questions, very informative plots can be produced. There are two issues of concern when considering multifactor designs: how to illustrate graphically main effects and interactions and how to deal with possible violations of homogeneity assumptions. For factorial designs, particularly those involving within-subject factors, violation of homogeneity of variance and sphericity assumptions create complex problems, especially if one wishes to assess interaction effects. Our general recommendation is that if such violations occur, then it may be best to apply a transformation to the data to reduce the degree of heterogeneity of variance. We turn, then, to a consideration of the interpretation of main effects and interactions in the context of a variety of multifactor designs.

DESIGNS WITH TWO LEVELS OF EACH FACTOR

Between-subject designs. In factorial designs, the primary question is, which *MS* term should be used to generate confidence intervals? For a pure between-subject design, there is only a single *MS* error term (MS_{Within}), representing a pooled estimate of variability. In this case, a single confidence interval can be constructed using a minor variant of Equation 2:

$$\mathrm{CI} = \mathrm{M_{jk}} \pm \sqrt{\frac{MS_{Within}}{n}}\,(\mathrm{t_{critical}}) \qquad (4)$$

where *n* is the number of subjects in each group (i.e.,

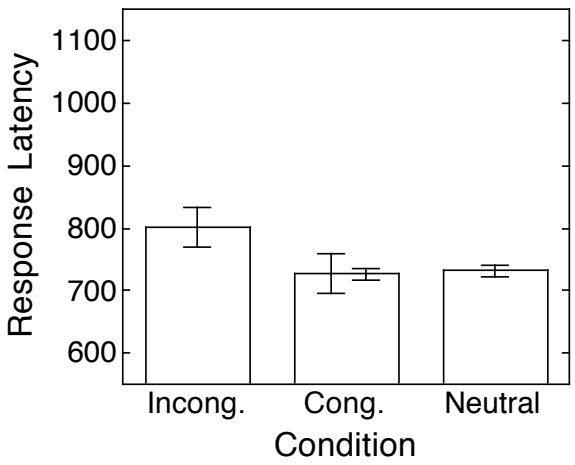

Figure 3. Condition means for the data from Table 2 plotted with separate confidence intervals for each of two contrasts: incongruent vs. congruent and congruent vs. neutral.

TABLE 3
Hypothetical Data for a Two-Factor Between-Subject Design

	Factor B	
Factor A	B1	B2
A1	0.51 (0.08)	0.50 (0.09)
A2	0.58 (0.11)	0.61 (0.08)

ANOVA Summary Table

Source	df	SS	MS	F
A	1	0.098	0.098	11.32
B	1	0.001	0.001	0.06
AxB	1	0.007	0.007	0.79
Within	44	0.381	0.009	
Total	47	0.486		

Note. Standard deviation in parentheses

the number of observations on which each of the j x k means is based). This confidence interval can be plotted with each mean and used to interpret the pattern of means. If there is a serious violation of the homogeneity of variance assumption (e.g., variances differ by more than 2:1 ratio), separate confidence intervals can be constructed for each group in the design using Equation 1.

A set of hypothetical descriptive statistics from a 2 x 2 between-subject factorial design and the corresponding ANOVA summary table for these data are shown in Table 3. Factor A represents two levels of an encoding task, Factor B two levels of type of retrieval cue, and the dependent variable is proportion correct on a cued recall test. There are 12 subjects in each cell. Note that the homogeneity assumption is met, so MS_{Within} can be used as a pooled estimate of variability. Applying Equation 4, yields a 95% confidence interval of

$$CI = \pm\sqrt{\frac{0.009}{12}}\,(2.017) = \pm 0.055\ .$$

The condition means are graphically displayed in the left panel of Figure 4 using this confidence interval. The pattern of means can be deduced from this display. Encoding (Factor A) has a substantial influence on performance, but there is little, if any, influence of Retrieval cue (Factor B). Moreover, the encoding effect is rather consistent across the different types of retrieval cue, implying that there is no interaction. If they are not very large, patterns of differences among means can be

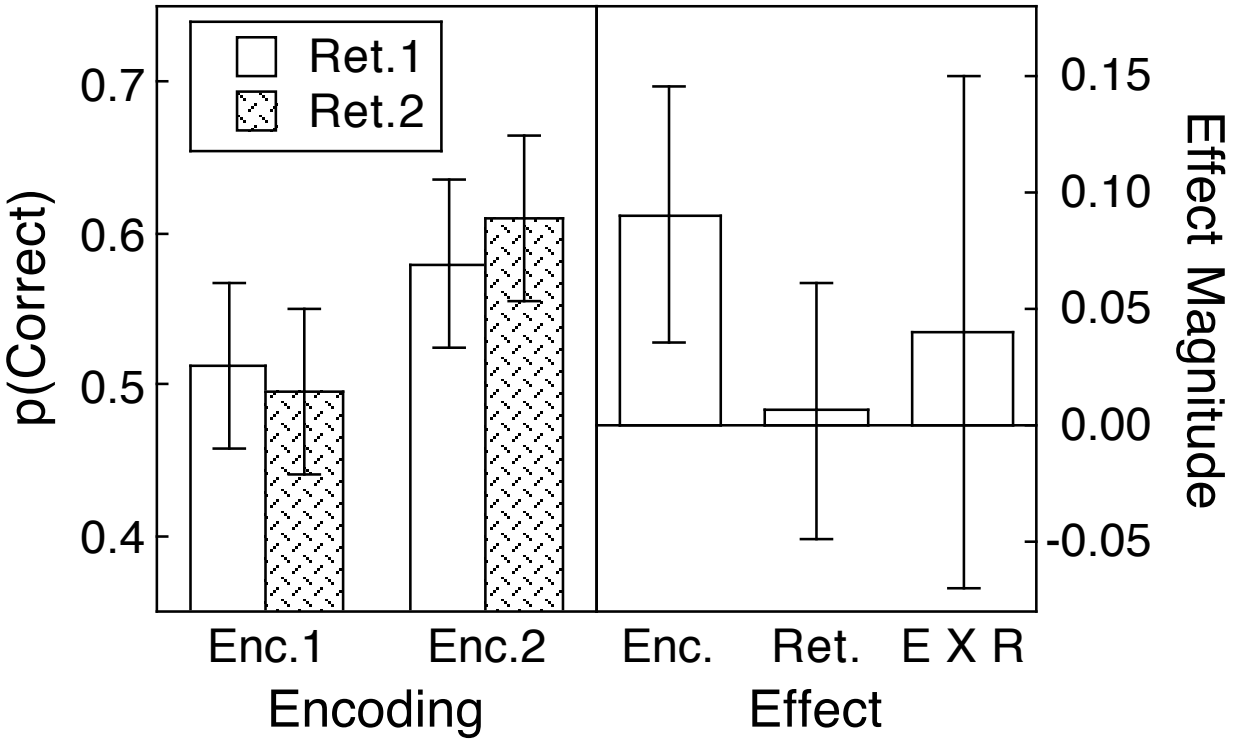

Figure 4. Group means (left panel) and contrasts for each effect (right panel) plotted with 95% confidence interval for data from Table 3. For convenience, each contrast is plotted as a positive value.

difficult to perceive, so it may be useful to highlight selected effects.

We recommend using contrasts as the basis for computing specific effects and their associated confidence intervals. To see how this is done, recall that the confidence intervals we have been describing so far are interpreted as confidence intervals around single means, enabling inferences about the pattern of means. In analyzing an effect defined by a contrast, we are considering an effect produced by a linear combination of means where that combination is defined by a set of weights applied to the means. Contrasts may be applied to a simple case, such as one mean versus another, with other means ignored, or to more complex cases in which combinations of means are compared (as when averaging across one factor of a factorial design to assess a main effect of the other factor). Applying contrast weights to a set of means in a design results in a contrast effect (e.g., a simple difference between two means, a main effect, an interaction effect) and a confidence interval can be defined for any such effect.

In the case of our 2 x 2 example, the main effect of Encoding (Factor A), can be defined by the following contrast of condition means: $(A_1B_1 + A_1B_2) - (A_2B_1 + A_2B_2)$. The weights applied to the four condition means that define this contrast are: 1, 1, –1, –1. More generally, the equation for defining a contrast as a linear combination of means is

$$Contrast\ Effect = \sum w_{jk} M_{jk} \qquad (5)$$

To compute the confidence interval for a linear combination of means, the following equation can be applied

$$CI_{contrast} = CI \sqrt{\sum w_{jk}^2} \qquad (6)$$

where CI is the confidence interval from Equation 2 and the w_{jk} are the contrast weights. Notice that weights can be of arbitrary size, as long as they meet the basic constraint of summing to zero (e.g., 2, 2, -2, -2 would be an acceptable set of weights in place of those shown above). The size of the confidence interval for a contrast effect will vary accordingly, as Equation 6 indicates. We suggest, however, that a particularly informative way to define contrast weights in a manner that reflects the averaging across means that is involved in constructing a contrast. For example, in defining the contrast for the Encoding main effect, we are comparing the average of two means against the average of another two means. Thus, the absolute value of the appropriate contrast weights that reflect this operation would be 0.5 (summing two means and dividing by two is equivalent to multiply each mean by 0.5 and adding the products). Thus, the set of weights that allows the main effect of the Encoding factor to be expressed as a comparison between the averages of two pairs of means would be 0.5, 0.5, –0.5, –0.5. Applying this set of weights to the means from Table 3 produces

$$Contrast\ Effect = 0.5(0.51) + 0.5(0.50) + (-0.5)(0.58) + (-0.5)(0.61) = -0.0905.$$

The mean difference between the two encoding conditions, then, is slightly more than 0.09. The confidence interval for this contrast is equal to the original confidence interval for individual means because the sum of the squared weights equals 1, so the second term in Equation 6 equals 1. This Encoding main effect contrast, representing the main effect of Encoding (Factor A), and its confidence interval are plotted in the right panel of Figure 4. For convenience we have plotted the contrast effect as a positive value. This move is not problematic because the sign of the effect is arbitrary (i.e., it is determined by which level of Factor A we happened to call level 1 and which we called level 2). Note that the confidence interval does not include zero, supporting the conclusion that the Encoding task manipulation favors condition A2, as implied by the pattern of means in the left panel of Figure 4.

For the main effect of Retrieval cue, the weights would be 0.5, –0.5, 0.5, –0.5. The absolute values of these weights are again 0.5 because the same type of comparison is being made as in the case of the Encoding main effect (comparison between averages of pairs of means). The contrast effect for the Retrieval cue effect and its confidence interval (which is the same as for the Encoding task effect because the contrast weights were identical) is shown in the right panel of Figure 4. The confidence interval includes zero, indicating that type of retrieval cue made little or no difference to recall performance.

For the interaction, however, we have a somewhat different situation. The interaction is a difference between differences, rather than a difference between averages. That is, a 2 x 2 interaction is based on computing the difference between the effect of Factor A at one level of Factor B, and the effect of A at the other level of B. In the present example, this contrast would be captured by the weights 1, –1, –1, 1. (This same set of weights can also be interpreted as representing the difference between the effect of Factor B at one level of Factor A and the effect of B at the other level of A.) Applying these weights for the interaction yields the following contrast effect:

$$Contrast\ Effect = 1(0.51) + (-1)(0.50) + (-1)(0.58) + 1(0.61) = 0.04\ .$$

Now consider the confidence interval for this effect. The square root of the sum of these squared weights is 2, so applying Equation 6 to obtain the confidence interval for the effect, we have

$$CI_{contrast} = \pm 0.055(2) = \pm 0.110 .$$

The interaction effect and its confidence interval are plotted in the right panel of Figure 4. Note that this confidence interval is twice the size of the confidence interval for the main effects. This occurs because we are computing a difference between differences, rather than between averages. Even so, this is a somewhat arbitrary stance. We could just as easily have converted the weights for the interaction contrast to ±0.5 and wound up with a confidence interval equal to that for the two main effects. But in treating the interaction contrast as a difference between differences and using ±1 as the weights, the resulting numerical contrast effect more directly reflects the concept underlying the contrast.

The principles described here easily can be scaled up to accommodate more than 2 factors. In such cases it would be particularly useful to plot each of the main effects and interactions as contrasts, as shown in Figure 4, because magnitude of interactions beyond two factors can be very hard to visualize based on a display of individual means. Moreover, main effects in such designs involve collapsing across three or more means, again making it difficult to assess such effects. At the same time, a major advantage of plotting individual means with confidence intervals is that one can examine patterns of means that may be more specific than standard main effects and interactions. And, of course, the plot of individual means, or means collapsed across one of the factors, can reveal the pattern of interaction effects (e.g., a cross-over interaction).

Within-subject designs. In a within-subject design, there are multiple *MS* error terms, one for each main effect and another for each possible interaction between independent variables. For a *J* x *K* two-factor design, for example, there are three *MS* error terms. It is possible that all *MS* error terms in a within-subject design are of similar magnitude (i.e., within a ratio of about 2:1), in which case the most straightforward approach is to combine all such sources of *MS* error to obtain a single, pooled estimate, just as though one had a single-factor design with *JK* conditions. Consider this approach in the case of a two-factor within-subject design for the hypothetical data shown in Table 4. In this case, the experiment involves an implicit measure of memory (latency on a lexical decision task) for words that had or had not been seen earlier. Factor A is Study (nonstudied vs. studied) and Factor B is Word frequency (low vs. high). The data are based on a sample of 12 subjects. To obtain a pooled *MS* error term, one would sum the three sums of squares corresponding to the three error terms in the design and divide by the sum of the degrees of freedom associated with these three terms. For the data in Table 4, the pooled estimate is

$$MS_{SXAB} = \frac{SS_{SXA} + SS_{SXB} + SS_{SXAXB}}{df_{SXA} + df_{SXB} + df_{SXAXB}} =$$

$$\frac{4465 + 3672 + 7045}{11 + 11 + 11} = 460.1 .$$

This estimate would then be used to compute a single confidence interval as follows:

$$CI = M_j \pm \sqrt{\frac{MS_{SXAB}}{n}}\ (t_{critical}) \qquad (7)$$

where *n* is the number of observations associated with each condition mean. Note that the subscript for the *MS* term in this equation reflects a pooled estimate of *MS* error, not the *MS* error term for the interaction alone.

TABLE 4
Hypothetical Data for a Two-Factor Within-Subject Design

	Factor B	
Factor A	B1	B2
A1	588 (69)	504 (78)
A2	525 (90)	478 (67)

ANOVA Summary Table				
Source	df	SS	MS	F
Subjects	11	241,663	21,969	
A	1	23,298	23,298	54.94
SXA	11	4,465	424	
B	1	51,208	51,208	153.40
SXB	11	3,672	334	
AXB	1	4,021	4,021	6.28
SXAXB	11	7,045	640	
Total	47	335,372		

Note. Standard deviation in parentheses

Thus, the degrees of freedom for the critical t-ratio would be the sum of the degrees of freedom for the three *MS* error terms. For the data in Table 4, the 95% confidence interval would be

$$CI = \pm\sqrt{\frac{460.1}{12}}(2.036) = \pm 12.61 \ .$$

This confidence interval can be plotted with each mean in the design and used to interpret patterns among any combination of means. The left panel of Figure 5 presents the data from Table 4 in this manner. One could also display each of the main effects and the interaction as contrasts as was done for the previous between-subjects example in Figure 4. But another alternative would be to highlight the interaction by plotting it as means of difference scores computed for each level of one factor. In the present example, it is of theoretical interest to consider the effect of prior study at each of the two frequency levels. For each subject, then, one could compute the effect of prior study at each level of word frequency, producing two scores per subject. The means of these difference scores are shown in the right panel of Figure 5. The confidence interval for this interaction plot can be computed from a *MS* error term obtained by computing a new ANOVA with only Factor B as a factor and using difference scores on Factor A (i.e., $A_1 - A_2$) for each subject, with one such score computed at each level of B. The resulting *MS* error term is 1,281. The confidence interval for this interaction effect using Equation 3 and a critical *t*-ratio for 11 degrees of freedom is

$$CI = \pm\sqrt{\frac{1281}{12}}(2.201) = \pm 22.74$$

Next, consider how to plot the means and confidence intervals when it is not advisable to pool the three *MS* error terms in a two-factor within-subject design. One could combine a pair of terms if they are sufficiently similar (within a factor of about two) and compute a confidence interval from a pooled *MS* error based on those two sources. A separate confidence interval could be computed for the other effect whose *MS* error is very different from the other two. The latter confidence interval would be appropriate for drawing conclusions only about the specific effect with which it is associated. To illustrate how this might be done, let us assume that in Table 4 the MS_{SXAXB} term was deemed much larger than the other two error terms, so only the MS_{SXA} and MS_{SXB} are pooled and the resulting confidence interval is plotted with the four means of the design. A subsidiary plot, like that shown in the right panel of either Figure 4 (to display all three effects

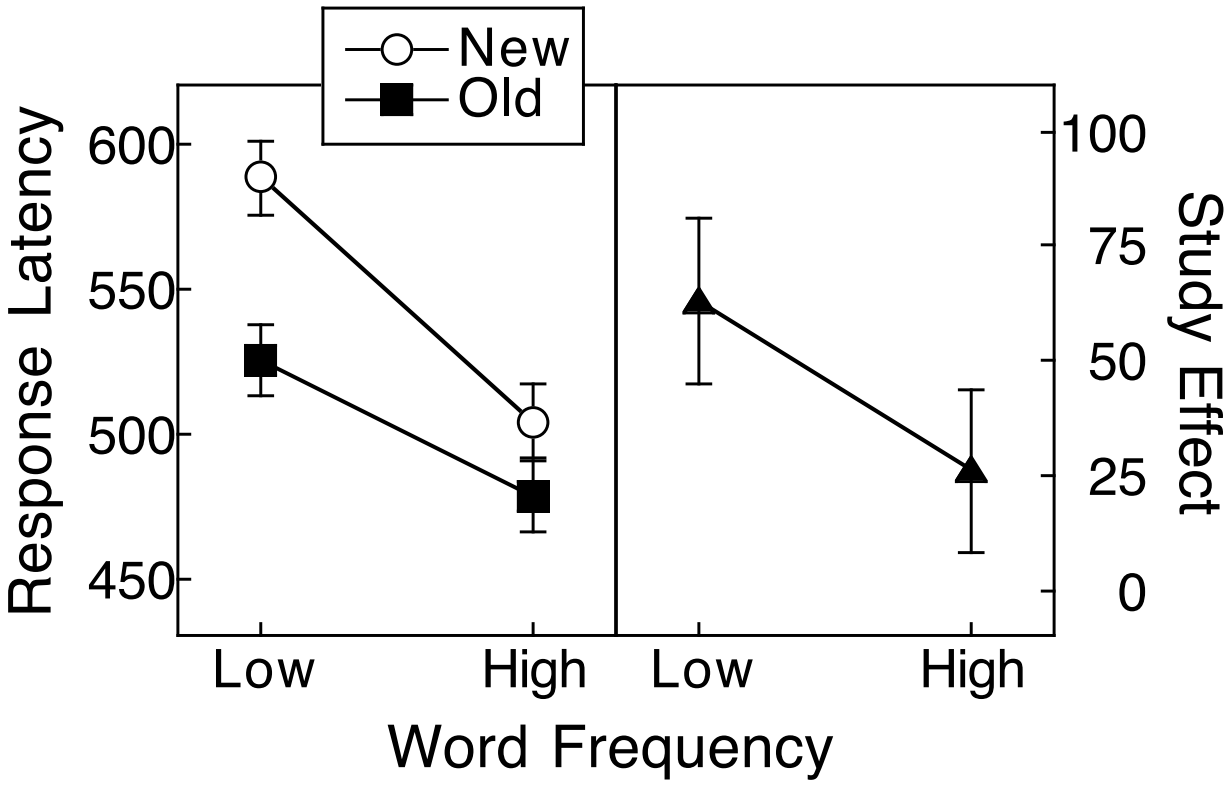

Figure 5. Condition means and interaction plot with 95% within-subject confidence interval for data from Table 4.

TABLE 5
Hypothetical Data for a Two-Factor Within-Subject Design

	Factor B	
Factor A	B1	B2
A1	.33 (.09)	.23 (.11)
A2	.18 (.07)	.20 (.09)

ANOVA Summary Table

Source	df	SS	MS	F
Subjects	15	0.211	0.014	
A	1	0.126	0.126	23.12
SXA	15	0.082	0.005	
B	1	0.023	0.023	2.13
SXB	15	0.16	0.011	
AXB	1	0.052	0.052	13.27
SXAXB	15	0.059	0.004	
Total	63	0.713		

Note. Standard deviation in parentheses

in the design) or Figure 5 (to display just the interaction) could then be constructed specifically for the interaction using a confidence interval computed from MS_{SXAXB}.

As an additional example of a case in which not all *MS* error terms are similar, consider the data set in Table 5. These data are the descriptive statistics and ANOVA for a 2 x 2 within-subject design with 16 subjects. Here, we have a case in which a word stem completion test is used as an implicit measure of memory and the factors are Study (Factor A), referring to whether or not a stem's target completion had been studied previously, and Delay interval (Factor B) between the study and test phase (i.e., the test follows

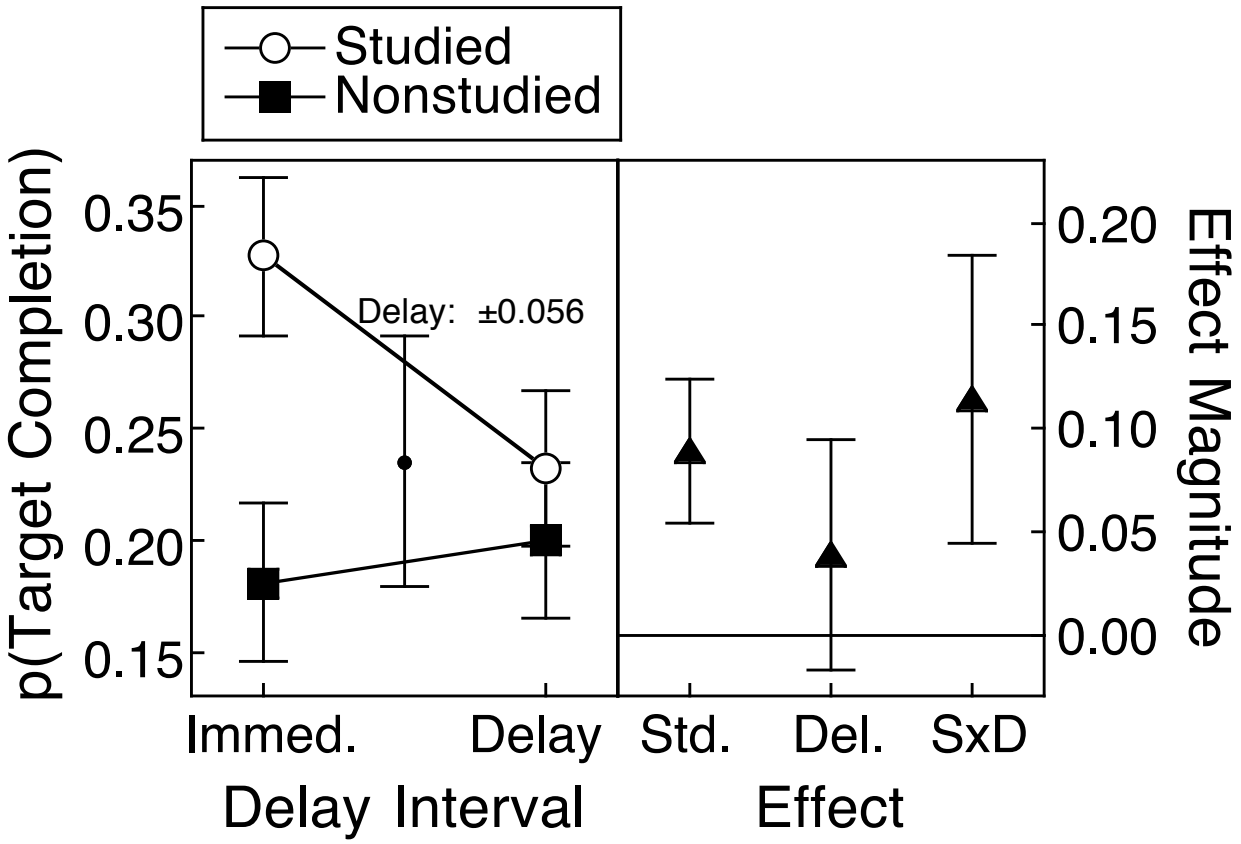

Figure 6. Condition means (left panel) and contrasts for each effect (right panel) plotted with 95% confidence interval for data from Table 5. For convenience, each contrast is plotted as a positive value.

immediately or after a delay). In this case, MS_{SXB} is quite dissimilar from the other two error terms, so we might wish to plot a confidence interval based on pooling MS_{SXA} and MS_{SXAXB}, to be plotted with each mean, then construct a separate confidence interval based on MS_{SXB} for use in the interpretation of Factor B. The confidence interval obtained by pooling across MS_{SXA} and MS_{SXAXB} is found by pooling the *MS* error terms

$$MS_{SxC} = \frac{0.082 + 0.059}{15 + 15} = 0.0047 ,$$

then computing the confidence interval using a *t*-ratio with $df_{SXA} + df_{SXAXB} = 15 + 15 = 30$ degrees of freedom:

$$CI = \pm\sqrt{\frac{0.0047}{16}}\ (2.042) = \pm 0.035 .$$

The confidence interval for the main effect of B is based on a t-ratio with $df_{SXB} = 15$ degrees of freedom:

$$CI = \pm\sqrt{\frac{0.011}{16}}\ (2.132) = \pm 0.056 .$$

This latter confidence interval could be strategically placed so that it is centered at a height equal to the grand mean, as shown in the left panel of Figure 6. This display gives an immediate impression of the magnitude of the B main effect. In addition, the right panel of Figure 6 shows all three effects of this design with the appropriate confidence interval for each. The contrast weights for each effect were the same as in the between-subject design above. Thus, the confidence intervals for the A main effect and for the interaction effect are different, despite being based on the same *MS* error term. The extension of these methods to designs with more than two factors, each having two levels, could proceed as described for the case of between-subject designs.

Mixed designs. In a mixed design, there is at least one between-subject and at least one within-subject factor. We will consider in detail this minimal case, although our analysis can be extended to cases involving a larger number of factors. In the 2-factor case, there are two different *MS* error terms. One is the *MS* within groups and is used to test the between-subject factor's main effect; the other is the *MS* for the interaction between the within-subject factor and subjects (i.e., the consistency of the within-subject factor across subjects) and is used to test the within-subject factor's main effect and the interaction between the between-subject and the within-subject factors. Unlike two-factor designs in which both factors are between- or both are within-subject factors, it would be inappropriate to pool the two MS error terms in a mixed design. Rather, separate confidence intervals must be constructed because the variability reflected in these two MS error terms are of qualitatively different types—one reflects variability between subjects and the other represents variability of the pattern of condition scores across subjects. The confidence interval based on the within-subject *MS* error term (MS_{SxC}) is computed using Equation 3, and the confidence interval for the *MS* error term for the between-subject factor (MS_{Within}) is computed using Equation 2.

Consider a hypothetical study in which subjects perform a lexical decision task and are presented with a semantically related versus unrelated prime for each target (the within-subject factor). The between-subject factor is the proportion of trials on which related primes are used (say, .75 vs. .25), and we will refer to this factor as relatedness proportion (RP). A hypothetical set of data is summarized in Table 6, which presents mean response time for word trials as a function of prime (factor P) and relatedness proportion (factor RP). There are 20 subjects in each RP condition. The table also shows the results of a mixed-factor ANOVA applied to these data. In this case, the prime main effect and the

TABLE 6
Hypothetical Data for a Two-Factor Mixed Design

	Prime	
RP	Related	Unrelated
25	497 (59)	530 (67)
0.75	517 (65)	577 (71)

ANOVA Summary Table

Source	df	SS	MS	F
RP	1	23,018	23,018	2.74
Within	38	318,868	8,391	
Prime	1	43,665	43,665	166.35
RP x P	1	3,605	3,605	13.73
S/RPXP	38	9,974	262	
Total	79	399,130		

Note. Standard deviation in parentheses

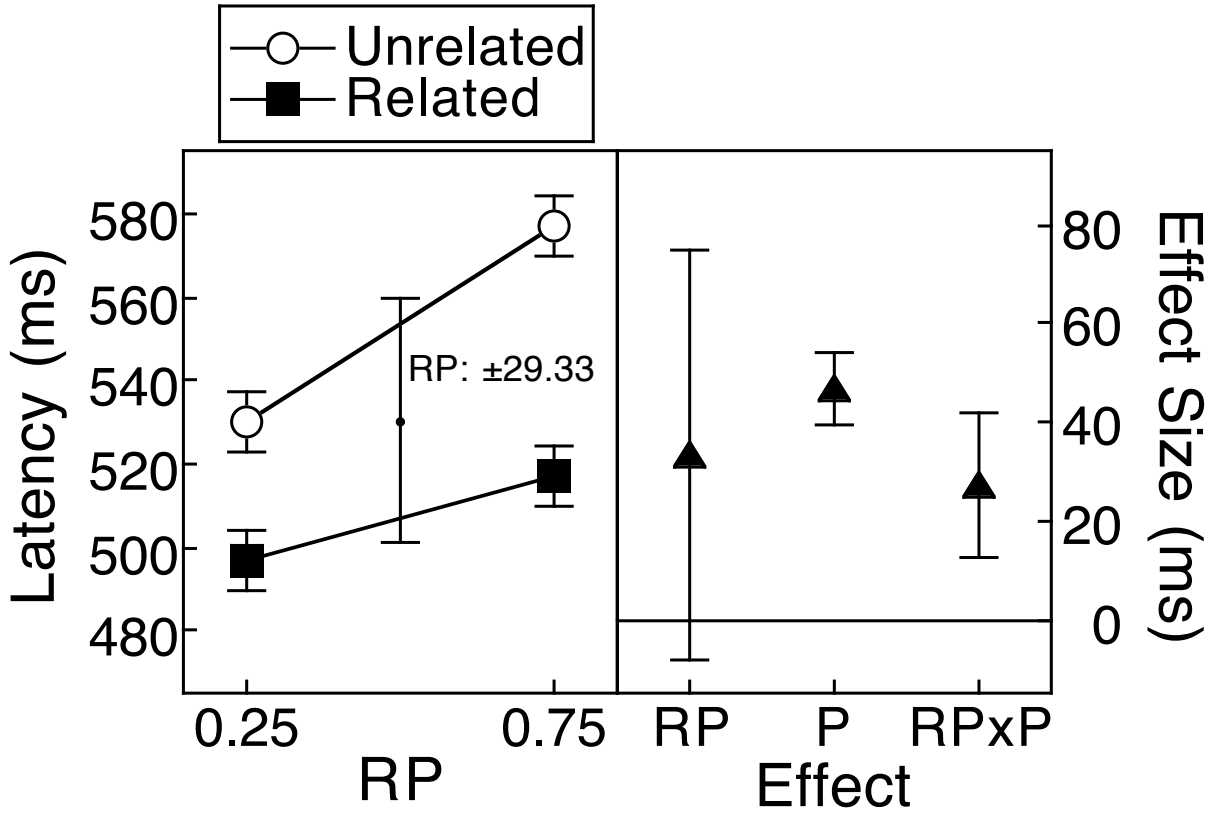

Figure 7. Condition means (left panel) and contrasts for each effect (right panel) plotted with 95% confidence interval for data from Table 6. For convenience, each contrast is plotted as a positive value.

interaction are clearly significant. Notice also that the two *MS* error terms in this design are, as will usually be the case, very different, with the MS_{Within} term much larger than the *MS* error for the within-subject factor and the interaction. The corresponding confidence intervals, then, are also very different. For the main effect of the between-subject factor, RP, the confidence interval (based on the number of observations contributing to each mean; two scores per subject in this case), using Equation 2, is

$$CI = \pm \sqrt{\frac{8,391}{(2)20}}\ (2.025) = 29.33\ .$$

The confidence interval for the within-subject factor, P, and the interaction, using Equation 3, is

$$CI = \pm \sqrt{\frac{262}{20}}\ (2.025) = 7.33\ .$$

The means for the four cells of this design are plotted in the left panel of Figure 7. Here, we have elected to display the confidence interval for the priming factor with each mean and have plotted the confidence interval for the main effect of RP separately. In addition, the right panel of the figure shows each of the main effects and the interaction effect plotted as contrast effects with their appropriate confidence intervals, as computed by Equations 5 and 6. The contrast weights for each effect were the same as in the factorial designs presented above.[2] An alternative approach to plotting these data is to compute a priming effect for each subject as a difference score (unrelated – related).

An ANOVA applied to these difference scores could then be used to compute a confidence interval for comparing the mean priming score at each level of RP and the two means and this confidence interval could be plotted, as was done in Figure 5.

[2]The main effect of RP in this case is based on just two means, one for high and one for low, so the contrast weights are 1 and –1. Therefore, the appropriate confidence interval for this main effect contrast is the original confidence interval multiplied by $\sqrt{2}$, as per Equation 6.

DESIGNS WITH THREE OR MORE LEVELS OF A FACTOR

The techniques described above can be generalized to designs in which at least one of the factors has more than two levels. In general, the first step is to plot means with confidence intervals based on a pooled *MS* error estimate where possible (in a pure within-subject design with *MS* error terms of similar size) or a *MS* error term that emphasizes a theoretically motivated comparison. In addition, one can plot main effects and interactions as illustrated in the previous section. With three or more levels of one factor, however, there is an additional complication: For any main effect or interaction involving a factor with more than two levels, more than one contrast can be computed. This fact does not mean that one necessarily should compute and report as many contrasts as degrees of freedom allow. Indeed, there may be only one theoretically interesting contrast. The important point to note here is simply that one has the choice of defining and reporting those contrasts that are of theoretical interest. Moreover, in designs with one within-subject factor that has two levels, an interaction plot can be generated by computing difference scores based on that factor (e.g., for each subject, subtract the score on level 1 of that factor from the score on level 2), as described above. The mean of these difference scores for each level of the other factor(s) can then be plotted, as in Figure 5.

Let us illustrate, though, an approach in which contrast effects are plotted in addition to individual condition means. For this example, consider a 2 x 3 within-subject design used by a researcher to investigate conscious and unconscious influences of memory in the context of Jacoby's (1991) process-dissociation procedure. In a study phase, subjects encode one set of words in a semantic encoding task and another set in a nonsemantic encoding task. In the test phase, subjects are given a word stem completion task with three sets of stems. One set can be completed with semantically encoded words from the study list, another set can be completed with nonsemantically encoded words, and the final set is used for a group of nonstudied words. For half of the stems of each type, subjects attempt to recall a word from the study list and to use that item as a completion for the stem (inclusion instruction). For the other half of the stems, subjects are to provide a completion that is not from the study list (exclusion instructions). The factors, then, are encoding (semantic, nonsemantic, nonstudied) and test (inclusion, exclusion).

A hypothetical data set from 24 subjects is shown in Table 7, representing mean proportion of stems for which target completions were given. The three *MS* error terms in the accompanying ANOVA summary table are similar to one another, so a combined *MS* error term can be used to compute a single confidence interval for the plot of condition means. The combined *MS* error is

$$MS_{SxC} = \frac{0.615 + 1.084 + 0.675}{23 + 46 + 46} = 0.021 .$$

Based on MS_{SxC}, the confidence interval associated with each mean in this study is

$$CI = \pm \sqrt{\frac{0.021}{24}} \, (1.982) = \pm 0.042 .$$

The means from Table 7 are plotted in the left panel of Figure 8 with this confidence interval.

A number of inferences can be drawn from this pattern of means (e.g., semantic encoding produces more stem completion than nonsemantic encoding under inclusion instructions, but the reverse occurs under

TABLE 7
Hypothetical Data for a Two-Factor Within-Subject Design with Three Levels of One Factor

	Encoding Tasl			ANOVA Summary Table				
Instr	Sem	Nonseman	New	Source	df	SS	MS	F
Incl.	.66 (.11)	.51 (.16)	.30 (.11)	Subjects	23	0.635	0.028	
Excl.	.32 (.16)	.47 (.20)	.28 (.14)	Instr.	1	0.618	0.618	23.13
				S x Instr.	23	0.615	0.027	
				Encoding	2	1.278	0.639	27.13
				S x Enc.	46	1.084	0.024	
				I x E	2	0.802	0.401	27.32
				S x I x E	46	0.675	0.015	
				Total	143	5.707		

Note. Standard deviation in parentheses

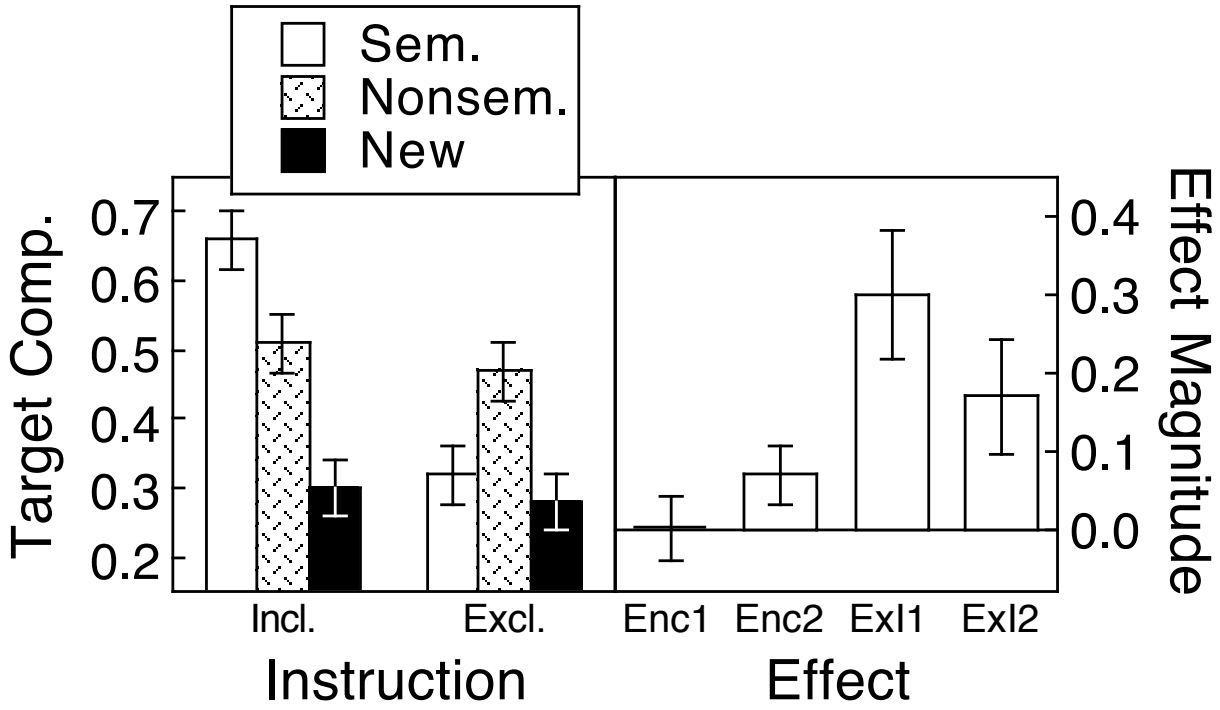

Figure 8. Condition means (left panel) and contrasts for effects involving the Encoding factor and its interaction with the Instruction factor (right panel) plotted with 95% confidence interval for data from Table 7. For convenience, each contrast is plotted as a positive value.

exclusion instructions). In addition, one may wish to emphasize specific aspects of one or more main effects or the interaction. Notice that in the ANOVA shown in Table 7, the encoding main effect and its interaction with test instruction have two degrees of freedom because the encoding factor has three levels. To display either of these effects using the general method illustrated in the earlier discussion of 2 x 2 designs, specific contrasts can be defined and plotted. For example, the main effect of encoding task could be expressed as two orthogonal contrasts combined across instruction condition: (1) semantic vs. nonsemantic and (2) semantic/nonsemantic combined vs. new. The weights defining the first contrast (semantic vs. nonsemantic encoding) for the inclusion instruction and exclusion instruction conditions, respectively, could be 0.5, –0.5, 0, 0.5, –0.5, 0. That is, the semantic conditions within each of the two instructional conditions are averaged, then contrasted with the average of the nonsemantic conditions within each of the instructional conditions. For the other contrast (semantic/nonsemantic combined vs. new), the contrast weights could be 0.25, 0.25, –0.5, 0.25, 0.25, –0.5. These weights reflect the fact that we are averaging across four conditions and comparing that average to the average of the two new conditions. The resulting contrast effects, then, would be

$$\textit{Contrast Effect 1} = 0.5(0.66) + (-0.5)(0.51) + 0.5(0.32) + (-0.5)(0.47) = 0\ ,$$

$$\textit{Contrast Effect 2} = 0.25(0.66) + 0.25(0.51) + (-0.5)(0.30) + 0.25(0.32) + 0.25(0.47) + (-0.5)(0.28) = 0.07\ .$$

The square root of the sum of squared weights for Contrast 1 is equal to 1, so the confidence interval for that contrast is equal to the original confidence interval. For Contrast 2, the square root of the sum of squared weights is 0.87, so the original confidence interval is scaled by this factor, in accordance with Equation 6, to yield a contrast-specific confidence interval of ±0.037. These two contrasts for the main effect of encoding task are plotted as the first two bars in the right panel of Figure 8.

Finally, consider two theoretically relevant contrasts based on the Encoding x Instruction interaction. Let us suppose that, as in the main effect contrasts, the first interaction contrast compares the semantic and nonsemantic conditions, ignoring the new condition. But now the weights are set to produce an interaction, which can be done by using opposite sign contrasts in each of the two instruction conditions, producing the following contrast weights for the semantic vs. nonsemantic interaction contrast: 1, –1, 0, –1, 1, 0. The

weights reflect the intended expression of the interaction effect as a difference between difference scores. This contrast effect is

Contrast Effect 3 = 1(0.66) + (–1)(0.51) + (–1)(0.32) + 1(0.47) = 0.30 .

The square root of the sum of the squared contrast weights in this case is 2, so the corresponding confidence interval is (±0.042)(2) = ±0.084. Let us further assume that the other contrast for the interaction effect will also parallel the second main effect contrast, in which the average of the semantic and nonsemantic conditions is compared to the new condition, but again using opposite signs for the two instruction conditions: 0.5, 0.5, –1, –0.5, –0.5, 1. Applying these weights to the means produces the following contrast:

Contrast Effect 4 = 0.5(0.66) + 0.5(0.51) + (–1)(0.30) + (–0.5)(0.32) +
(–0.5)(0.47) + 1(0.28) = 0.17 .

The confidence interval for this interaction is (±0.042)(1.73) = ±0.073. The two interaction contrasts are plotted as the final two bars in the right panel of Figure 8.

There are, of course, other possible contrasts that could have been defined and plotted. Loftus (2002) presented further examples of contrast effects that can be used to make full use of graphical displays of data from factorial designs.

Power

One particularly important advantage of graphical presentation of data, and especially confidence intervals, is that such presentations provide information concerning statistical power. The power of an experiment to detect some effect becomes very important when the effect fails to materialize and that failure carries substantial theoretical weight. The prescription for computing power estimates under the standard NHST approach requires that one specify a hypothetical effect magnitude. Although it is possible to select such a value on the basis of theoretical expectation or related empirical findings, the estimate of power depends to a large degree on this usually arbitrary value. An alternative approach is to report not just a single power estimate based on a single hypothesized effect magnitude, but a power curve that presents power associated with a wide range of possible effect magnitudes. In practice, this is rarely (if ever) done.

With graphical plots of data, however, we have a ready-made alternative means of displaying the power of an experiment. Specifically, the smaller the confidence intervals, the greater the amount of statistical power and, more important, the greater the confidence we can place in the observed pattern of means. To illustrate this concept, consider two different data sets from a reaction time task, each representing a one-factor within-subject design with two levels of the factor. The means and the MS_{SxC} in the two data sets are the same, but the sample size is much larger in the second case leading to a smaller confidence interval for the means. The means and confidence intervals are shown in Figure 9. The floating confidence interval in each case is the 95% confidence interval for the difference between means, computed using Equation 6 and the contrast weights 1 and –1. It is immediately obvious that we can have greater confidence in the pattern of means in the set of data on the right side of Figure 9.

In normal situations, only one of these two sets of data would be available, so how can one know whether the confidence interval, either for means or for differences between means is small enough to indicate that substantial statistical power is available? There are a number of benchmarks one might rely on. For example, for commonly used tasks there often are well established rules of thumb regarding how small an effect can be detected (e.g., in lexical decision tasks, reliable effects are usually 20 ms or more). Alternatively, as in power estimation under NHST, there may be empirical or theoretical reasons to expect an effect of a particular magnitude. If an expected effect size is larger than the observed effect and larger than the confidence interval

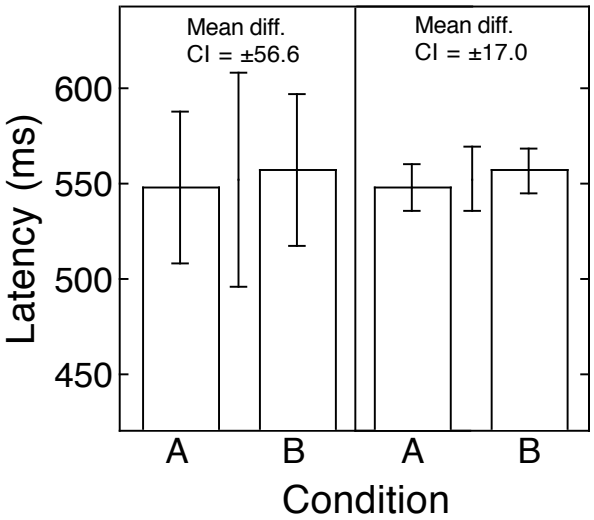

Figure 9. Condition means for hypothetical data representing an experiment with low power (left panel) and an experiment with high power (right panel), plotted with 95% within-subject confidence intervals. The floating confidence interval represents the 95% confidence interval for the difference between means.

for the difference between means, it is reasonable to conclude that the true effect is smaller than what was expected.

Another kind of benchmark for evaluating power comes from inherent limits on scores in a particular experiment. For instance, consider an experiment on long-term priming. If one is assessing differences in priming between two study conditions, it is possible to see immediately whether the study has adequate power to detect a difference in priming, because adequate power depends on whether the confidence interval for the difference between means is smaller or larger than the observed priming effects. If the confidence interval for the difference between means is larger than the priming effects themselves, then clearly there is not adequate power—one condition would have to have a negative priming effect for a difference to be found!

Conclusion

We have described a number of approaches to graphical presentation of data in the context of classical factorial designs that typify published studies in experimental psychology. Our emphasis has been on the use of confidence intervals in conjunction with graphical presentation to allow readers to form inferences about the patterns of means (or whatever statistic the author opts to present). We have also tried to convey the notion that authors have a number of options available with respect to construction of graphical presentations of data and that selection among these options can be guided by specific questions or hypotheses about how manipulated factors are likely to influence behavior. The approach we have described represents a supplement or, for the bold among us, an alternative to standard NHST methods.

References

Chow, S. L. (1998). The null-hypothesis significance-test procedure is still warranted. *Behavioral and Brain Sciences*, *21*, 228-238.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Estes, W. K. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review*, *4*, 330-341.

Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, Series A*, *158*, 175-177.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15-24.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*, 3-7.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30, 513-541.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*, 16-26.

Lewandowsky, S., & Maybery, M. (1998). The critics rebutted: A Pyrrhic victory. *Behavioral and Brain Sciences*, *21*, 210-211.

Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology*, *36*, 102-105.

Loftus, G.R. (1993). A picture is worth a thousand *p*-values: On the irrelevance of hypothesis testing in the computer age. *Behavior Research Methods, Instrumentation & Computers*, *25*, 250-256.

Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161-171.

Loftus, G. R. (2002). Analysis, interpretation, and visual presentation of experimental data. In H. Pashler (Ed.), *Stevens' handbook of experimental psychology* (Vol. 4, pp. 339-390). New York: John Wiley and Sons.

Loftus, G. R. & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476-490.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in Psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115-129.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371-386.

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.