Psychonomic Bulletin & Review
1994, 1, 476-490

Using Confidence Intervals in Within-Subject Designs

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

We argue that to best comprehend many data sets, plotting judiciously selected sample statistics with associated confidence intervals can usefully supplement, or even replace, standard hypothesis-testing procedures. We note that most social sciences statistics textbooks limit discussion of confidence intervals to their use in between-subjects designs. Our central purpose in this article is to describe how to compute an analogous confidence interval that can be used in within-subjects designs. This confidence interval rests on the reasoning that because between-subjects variance typically plays no role in statistical analyses of within-subjects designs, it can legitimately be ignored; hence, an appropriate confidence interval can be based on the standard within-subject error term—i.e., on the variability due to the subject-by-condition interaction. Computation of such a confidence interval is simple and is embodied in Equation 2 on p. xx of this article. This confidence interval has two useful properties. First, it is based on the same error term as is the corresponding ANOVA, and hence leads to comparable conclusions. Second, it is related by a known factor ($\sqrt{2}$) to a confidence interval of the difference between sample means; accordingly it can be used to infer the faith one can put in some pattern of sample means as a reflection of the underlying pattern of population means. These two properties correspond to analogous properties of the more widely used between-subjects confidence interval.

Most data analysis within experimental psychology consists of *statistical* analysis, most of which revolves in one way or another around the question, "What is the correspondence between a set of observed sample means and the associated set of population means that the sample means are estimating[1]?" If this correspondence were known, then most standard statistical analyses would be unnecessary. Imagine, for example, an ideal experiment which incorporated such a vast amount of statistical power that all population means could be assumed essentially equal to the corresponding observed sample means. With such an experiment, it would make little sense to carry out a standard test of some null hypothesis because the test's outcome would be apparent from inspection of the sample means. Data analysis could accordingly be confined to the scientifically useful processes of parsimoniously characterizing the observed pattern of sample means and/or determining the implications of the observed pattern for whatever question the experiment was addressing to begin with.

With a real, as opposed to an ideal experiment, population means are typically not known but only estimated, which is why we *do* do statistical analyses. Thus, some determination of how much faith can be put in the observed pattern of sample means must form a

The writing of this manuscript was supported by an NIMH grant to Loftus and a Canadian NSERC grant to Masson. We thank Jim Colton, Steve Edgell, Rich Gonzalez, David Lane, Jeff Miller, Rich Schweickert, Saul Sternberg, George Wolford, and two anonymous reviewers for very useful comments on earlier drafts of the manuscript. Requests for reprints may be sent to Geoffrey R. Loftus, Department of Psychology, University of Washington, Seattle, WA 98195; email: gloftus@u.washington.edu.

[1]For expositional simplicity, we will couch our arguments using sample means, realizing that analogous arguments could be made about any sample statistic.

preliminary step to be carried out prior to evaluating what the observed pattern might imply for the question at hand.

This preliminary step can take one of several forms. In the social sciences, the overwhelmingly dominant form is that of hypothesis testing: one formulates a null hypothesis, typically that some set of population means are all equal to one another and, based on the pattern of sample means along with some appropriate error variance, decides either to reject or to not reject the null hypothesis. In this article we echo suggestions (e.g., Tukey, 1974; 1977; Wainer & Thissen, 1993) that graphical procedures—particularly construction of confidence intervals—can be carried out as a supplement, or even as a replacement for standard hypothesis-testing procedures. Before doing so, however, we briefly consider the origins of procedures now in common use.

Historical Roots

The hypothesis-testing procedures that now dominate data analysis techniques in the behavioral sciences have evolved as something of an expedient compromise between a number of ideologically conflicting approaches to drawing conclusions from statistical data (see Gigerenzer, Swijtink, Porter, Daston, Beatty, & Krüger, 1989 for a thorough discussion of this assertion).

Bayesian Techniques

One of these approaches, which turned out not to have a strong influence on the techniques that are widely used in behavioral sciences today, is based on ideas developed by Bayes (1763; see Berger & Berry, 1988, and Winkler, 1993 for clear introductions to Bayesian statistical analysis; see Box & Tiao, 1973, and Lewis, 1993, for extensive treatments of the Bayesian approach to analysis of variance). In the Bayesian approach, the goal is to estimate the probability that a hypothesis is true and/or to determine some population parameter's distribution given the obtained data.

Computing this probability or probability distribution requires specification of an analogous probability or probability distribution prior to data collection (the *prior probability*) and, in experimental designs, specification of the maximal effect that the independent variable can have on changing these prior probabilities. An important feature of the Bayesian approach is that interpretation of data depends crucially on the specification of such prior probabilities. When there is no clear basis for such specification, data interpretation will vary across researchers who hold different views about what ought to be the prior probabilities.

Null Hypotheses and Significance Testing

An alternative to the Bayesian approach was developed by Fisher (1925, 1935, 1955), who proposed that data evaluation is a process of *inductive inference* in which a scientist attempts to reason from particular data to draw a general inference regarding a specified null hypothesis. In this view, statistical evaluation of data is used to determine how likely is an observed result under the assumption that the null hypothesis is true. Note that this view of data evaluation is opposite to that of the Bayesian approach, in which an observed result influences the probability that a hypothesis is true. In Fisher's approach, results with low probability of occurrence are deemed statistically significant and taken as evidence against the hypothesis in question. This concept is known to any modern student of statistical applications in the behavioral sciences.

Less familiar, however, is Fisher's emphasis on significance testing as a formulation of belief regarding a single hypothesis, and, in keeping with the grounding of this approach in inductive reasoning, the importance of both replications and replication failures in determining the true frequency with which a particular kind of experiment has produced significant results. Fisher was critical of the Bayesian approach, however, particularly because of problems associated with establishing prior probabilities for hypotheses. When no information about prior probabilities is available, there is no single accepted method for assigning probabilities. Therefore, different researchers would be free to use different approaches to establishing prior probabilities, a form of subjectivity that Fisher found particularly irksome. Moreover, Fisher emphasized the point that a significance test does not allow one to assign a probability to a hypothesis, but only to determine the likelihood of obtaining a result under the assumption that the hypothesis is valid. One's

degree of belief in the hypothesis might then be modified by the probability of the result, but the probability value itself was not to be taken as a measure of the degree of belief in the hypothesis.

Competing Hypotheses

In contrast to Fisher's emphasis on inductive *reasoning*, a third approach to statistical inference was developed by Neyman and Pearson (1928, 1933; Neyman, 1957). They were primarily concerned with inductive *behavior*. For Neyman and Pearson, the purpose of statistical theory was to provide a *rule* specifying the circumstances under which one should reject or provisionally accept some hypothesis. They shared Fisher's criticisms of the Bayesian approach, but went one step further. In their view, even degree of belief in a hypothesis did not enter into the picture. In a major departure from Fisher's approach, Neyman and Pearson introduced the concept of two competing hypotheses, one of which is assumed to be true. In addition to establishing a procedure based on two hypotheses, they also developed the concept of two kinds of decision error: rejection of a true hypothesis (Type I error) and acceptance of a false hypothesis (Type II error). In the Neyman-Pearson approach, both hypotheses are stated with equal precision so that both types of error can be computed. The relative importance of the hypotheses, along with the relative costs of the two types of error are used to set the respective error probabilities. The desired Type I error probability is achieved by an appropriate choice of the rejection criterion, while the desired Type II error probability is controlled by varying sample size. In this view, Type I and Type II error rates will vary across situations according to the seriousness of each error type within the particular situation.

Neyman and Pearson's hypothesis-testing approach differs from Fisher's approach in several ways. First, it requires consideration of two, rather than just one, precise hypotheses. This modification enabled computation of *power estimates*, something that was eschewed by Fisher who argued that there was no scientific basis for precise knowledge of the alternative hypothesis. In Fisher's view, power could generally not be computed, although he recognized the importance of sensitivity of statistical tests (Fisher, 1947). Second, Neyman and Pearson provided a prescription for *behavior*—i.e., for a *decision* about whether to reject a hypothesis. Fisher's, on the other hand, emphasized the use of significance testing to measure the degree of discordance between observed data and the null hypothesis. The significance test was intended to influence the scientist's belief in the hypothesis, not simply to provide the basis for a binary decision (the latter, is a stance that Neyman and Pearson critically viewed as "quasi-Bayesian"; see Gigerenzer *et al.*, 1989, p. 103).

The issues raised in the Fisher vs Neyman/Pearson debate have not been settled, and are still discussed in the statistical literature (e.g., Camilli, 1990; Lehmann, 1993). Nevertheless, there has been what Gigerenzer *et al.* (1989) referred to as a "silent solution" within the behavioral sciences. This solution has evolved from statistical textbooks written for behavioral scientists, and consists of a combination of ideas drawn from Fisher and from Neyman and Pearson. For example, drawing on Neyman and Pearson, researchers are admonished to specify the significance level of their test prior to collecting data. But little if anything is said about why a particular significance level is chosen and few texts discuss consideration of the costs of Type I and Type II error in establishing the significance level. Following the practice established by Fisher, however, researchers are taught to draw no conclusions from a statistical test that is not significant. Moreover, concepts from the two viewpoints have been mixed together in ways that contradict the intentions of the originators. For instance, in current applications, probabilities associated with Type I and Type II errors are not used only for reaching a binary decision about a hypothesis, as advocated by Neyman and Pearson but often are also treated as measures of degree of belief, as per Fisher's approach. This tendency has on many occasions led researchers to state the most stringent possible level of significance (e.g., $p < .01$) when reporting significant results, apparently with the intent of convincing the skeptical reader.

Perhaps the most disconcerting consequence of the hypothesis testing approach as it is now practiced in behavioral science is that it often is a mechanistic enterprise that is ill-

suited for the complex and multidimensional nature of most social-science data sets. Both Fisher and Neyman and Pearson (as well as the Bayesians) clearly realized this, in that they consider *judgment* to be a crucial component drawing inferences from statistical procedures. Similarly, judgment is called for in other solutions to the debate between the Fisher and the Neyman-Pearson schools of thought, as in the suggestion to apply different approaches to the same set of data (e.g., Box, 1986).

Graphical Procedures

Traditionally, as we have suggested, the primary data-analysis emphasis in the social sciences has been on *confirmation:* the investigator considers a small number of hypotheses and attempts to confirm or disconfirm them. Over the past twenty years, however, a consensus has been (slowly) growing that exploratory, primarily *graphical* techniques are at least as useful as confirmatory techniques in the endeavor to maximally understand and use the information inherent in a data set (see, Tufte, 1983; 1990 for superb examples of graphical techniques, and Wainer & Thissen, 1993, for an up-to-date review of them).

A landmark event in this shifting emphasis was publication (and dissemination of prepublication drafts) of John Tukey's (1977) book, *Exploratory Data Analysis* which heralded at least an "official" toleration (if not actually a widespread use) of exploratory and graphical techniques. Tukey's principal message is perhaps best summarized by a remark that previewed the tone of his book: "The picturing of data allows us to be sensitive not only to the multiple hypotheses that we hold, but to the many more we have not yet thought of, regard as unlikely or think impossible" (Tukey, 1974, p. 526). It is in this spirit that we focus on a particular facet of graphical techniques, that of confidence intervals.

Confidence Intervals

We have noted that, whether framed in a hypothesis-testing context or in some other context, a fundamental statistical question is, "How well does the observed pattern of sample means represent the underlying pattern of population means?" Elsewhere, one of us has argued that construction of confidence intervals, which directly addresses this question, can profitably supplement (or even replace) the more common hypothesis-testing procedures (Loftus, 1991; 1993a, 1993b, 1993c; see also Bakan, 1966; Cohen, 1990). These authors offer many reasons in support of this assertion. Two of the main ones are as follows. First, hypothesis testing is primarily designed to obliquely address a restricted, convoluted, and usually uninteresting question, "Is it not true that some set of population means are all equal to one another?" whereas confidence intervals are designed to directly address a simpler and more general question, "What *are* the population means?" Estimation of population means, in turn, facilitates evaluation of whatever theory-driven alternative hypothesis is under consideration.

A second argument in favor of using confidence intervals (and against sole reliance on hypothesis testing) is that it is a rare experiment in which any null hypothesis could plausibly be true. That is, it is rare that a set of population means corresponding to different treatments could all be identically equal to one another. Therefore it usually makes little sense to test the validity of such a null hypothesis; a finding of statistical significance typically implies only that the experiment has enough statistical power to detect the population mean differences that one can assume apriori must exist[2].

We assert that at the very least, plotting a set of sample means along with their confidence intervals can provide an initial, rough-and-ready, intuitive assessment of (1) the best estimate of the underlying pattern of population means and (2) the degree to which the observed pattern of sample means should be taken seriously as a reflection of the underlying pattern of population means, i.e., the degree of statistical *power* (an aspect of statistical analysis that is usually ignored in social-science research).

---

[2]Some caveats should be noted in conjunction with these assertions. First, on occasion, a plausible null hypothesis does exist (e.g., that performance is at chance in a parapsychological experiment). Second, in a two-tailed z- or t-test situation, rejection of some null hypothesis can establish the *directionality* of some effect. (Note, however, that even this latter situation rests on a logic by which which one tests the validity of some usually implausible null hypothesis).

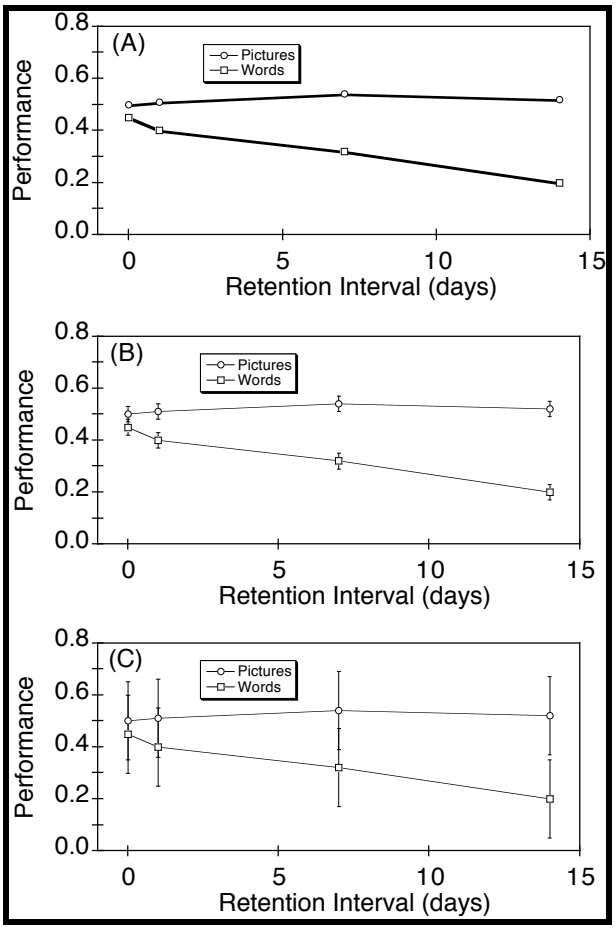

Figure 1. Hypothetical data without confidence intervals (Panel A) and with confidence intervals (Panels B and C).

Consider for example the hypothetical data shown in Figure 1A, which depicts memory performance following varying retention intervals for picture and word stimuli. Although Figure 1A provides the best estimate of the pattern of underlying population means, there is no indication as to how seriously this best estimate should be taken—that is there is no indication of error variance. In Figures 1B and 1C, 95% confidence intervals provide this missing information (indicating that the observed pattern should be taken very seriously in the case of Figure 1B—which depicts something close to the ideal experiment described above—but not seriously at all in the case of Figure 1C, which would clearly signal the need for additional statistical power in order to make any conclusions at all from the data). Furthermore, a glance at either Figure 1B or 1C would allow a quick assessment of how the ensuing hypothesis-testing procedures would likely turn out. Given the Figure-1B data pattern, for instance, there would be little need for further statistical analyses.

Among the reactions to the advocacy of routinely publishing confidence intervals along with sample means has been the observation (typically in the form of personal communication to the authors) that most textbook descriptions of confidence intervals are restricted to between-subject designs; hence many investigators are left in the dark about how to compute analogous confidence intervals in within-subjects designs. Our purpose here is to fill this gap, i.e., to describe a rationale and a procedure for computing confidence intervals in within-subject designs. Our reasoning is an extension of that provided by a small number of introductory statistics textbooks, generally around page 400 (e.g., Loftus and Loftus 1988, pp. 411-429; Anderson & McLean, 1974, pp. 407-412). It goes as follows.

A standard confidence interval in a between-subjects design has two useful properties. First, the confidence interval's size is determined by the same quantity that serves as the error term in the ANOVA; thus the confidence interval and the ANOVA, based as they are on the same information, lead to comparable conclusions. Second an X% confidence interval around a sample mean and an X% confidence interval around the difference between two sample means are related by a factor of $\sqrt{2}$.[3] This forms the basis of our assertion that confidence in patterns of means (of which the difference between two means is a basic unit) can be judged based on the confidence intervals plotted around the individual sample means. The within-subjects confidence interval that we will describe has these same two key properties.

[3]Because we are interested in comparing within- and between-subjects designs, we restrict ourselves to between-subjects situations in which equal numbers of subjects are assigned to all J conditions. We also assume homogeneity of variance, which implies that confidence intervals around all sample means are determined by a common, pooled error term. In a later section, we consider the case in which this assumption is dropped.

Table 1

A Between-Subjects Design: Number of Words Recalled (out of 20) for each of 10 Subjects in each of Three Conditions. (NOTE: M_j: Mean of Condition j).

Exposure Duration Per Word (sec)		
1 Sec	2 Sec	5 Sec
10	13	13
6	8	8
11	14	14
22	23	25
16	18	20
15	17	17
1	1	4
12	15	17
9	12	12
8	9	12
$M_1 = 11.0$	$M_2 = 13.0$	$M_3 = 14.2$

In the text that follows, we present the various arguments at an informal, intuitive level. The Appendixes to this article provide the associated mathematical underpinnings.

A Hypothetical Experiment

Consider a hypothetical experiment designed to measure effects of study time in a free-recall paradigm. In this hypothetical experiment, to-be-recalled 20-word lists are presented at a rate of either 1, 2, or 5 sec per word. Of interest is the relation between study time and number of recalled list words.

Between-Subject Data

Suppose first that the experiment is run as a between-subject design in which N = 30 subjects are randomly assigned to three groups of n = 10 subjects per group. Each group then participates in one of the three study-time conditions, and each subject's number of recalled words is recorded. The data are presented in Table 1 and Figure 2A. Both figure and table show the mean number of words recalled by each subject (shown as small dashes in Figure 2A) as well as the means over subjects (shown as closed circles connected by the solid line).

Table 1 and Figure 2A elicit the intuition that the study-time effect would not be significant in a standard ANOVA: there is too much variability over the subjects within each condition (reflected by the spread of individual-subject points around each condition mean and quantified as MS_W) compared to the rather meager variability across conditions (reflected by the differences among the three means and quantified as MS_C). Sure enough, as shown in the ANOVA table at the lower right of Figure 2A, the study-time effect is not statistically significant, $F(2,27) < 1$.

Between-Subject Confidence Intervals

Figure 2B shows the 95% confidence interval around the three condition means. This confidence interval is based on the pooled estimate of the within-condition variance, i.e., on MS_W. It is therefore based on $df_W = 27$, and is computed by the usual formula,

$$CI = M_j \pm \sqrt{\frac{MS_W}{n}} \text{ [criterion t(27)]} \quad (1)$$

which, as indicated on the figure, is ±3.85 in this example.

Figure 2B provides much the same information as does the ANOVA shown in Figure 2A. In particular, a glance at Figure 2B indicates the same conclusion reached via the ANOVA: given our knowledge about the values of the three condition population means, we can't exclude the possibility that they are all equal. More generally, the confidence intervals indicate that any possible ordering of the three population means is well within the realm of possibility. Note that the intimate correspondence between the ANOVA and the confidence interval comes about because their computations involve the common error term, MS_W.

Individual Population Means vs Patterns of Population Means

A confidence interval by definition provides information about the value of some specific population mean; e.g., the confidence interval around the left-hand mean of Figure 2B provides information about the population mean corresponding to the 1-sec condition. However, in psychological experiments, it is rare (although, as we discuss in a later section, not unknown) that one is genuinely interested in inferring the specific value of a population mean. More typically, one is interested in inferring the *pattern* formed by a *set* of population means. In the present example, the primary interest is not so much in the absolute values of the three population means, but rather in how they are related to one another. A hypothesis that might be of interest, for example, is that the population means increase with longer study times. In short, isolating the values of the individual population means is generally interesting only insofar as it reveals something about the pattern that they form.

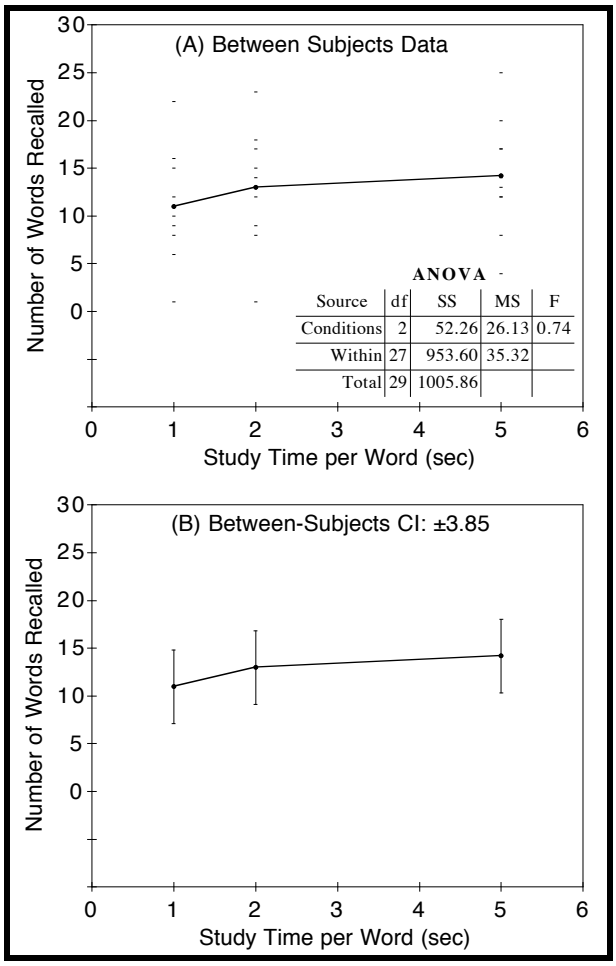

Figure 2. An example of a between-subject design. Top panel: Means surrounded by individual data points. Bottom panel: Confidence intervals around the three data points.

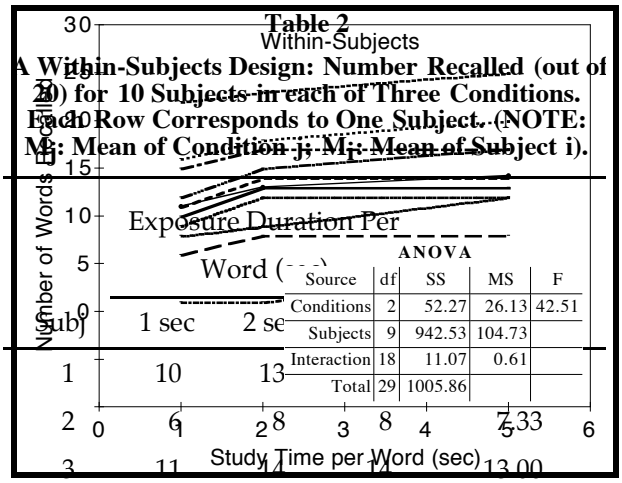

Figure 3. An example of a within-subject design: Means (connected by the heavy solid line) are shown with individual subject curves (other lines).

Table 2
Within-Subjects

A Within-Subjects Design: Number Recalled (out of 30) for 10 Subjects in each of Three Conditions. Each Row Corresponds to One Subject. (NOTE: M_j: Mean of Condition j; M_i: Mean of Subject i).

Subj	1 sec	2 sec	4 sec	M_i
1	10	13	14	12.33
2	6	8	8	7.33
3	11	14	14	13.00
4	22	23	25	23.33
5	16	18	20	18.00
6	15	17	17	16.33
7	1	1	4	2.00
8	12	15	17	14.67
9	9	12	12	11.00
10	8	9	12	9.67
M_j	M_1 = 11.0	M_2 = 13.0	M_3 = 14.2	M = 12.73

ANOVA

Source	df	SS	MS	F
Conditions	2	52.27	26.13	42.51
Subjects	9	942.53	104.73	
Interaction	18	11.07	0.61	
Total	29	1005.86		

Within-Subject Data

Let us now suppose that the numbers from Table 1 came not from a between-subject design, but from a within-subject design. Suppose, that is, that the experiment included a *total* of n = 10 subjects, each of whom participated in all three study-time conditions. Table 2 reproduces the Table-1 data from each of the three conditions, showing in addition the mean for each subject (row mean) along with the grand mean, M = 12.73 (Table 2, bottom right). Figure 3 shows these data in graphical form: the individual subject curves (thin lines) are shown along with the curve for the condition means (heavy line). Note that the condition means, based as they are on the same numbers as they were in Table 1 and Figure 2, do not change.

Error Variance and the Notion of "Consistency"

The Figure-3 data pattern should suffice to convince the reader that an effect of study time can be reasonably inferred (specifically, a monotonically increasing relation between study time and performance). This is because each of the 10 subjects shows a small *but consistent* study-time effect. Statistically, this consistency is reflected in the small mean square due to interaction (MS_{SxC} = 0.61) in the ANOVA table at the bottom right of Figure 3. And, indeed, the F for the study-time conditions, now computed as MS_C/MS_{SxC} is highly significant, $F(2,18) = 42.51$.

Constructing a Confidence Interval

Suppose that we wished to construct a confidence interval based on these within-subject data. As shown in Appendix A(2), a *bona fide* confidence interval—one designed to provide information about values of individual population means—would be exactly that shown in Figure 2B, i.e., ±3.85. That is, if we wish to provide information about, say, the value of the 1-sec condition population mean, we must construct the confidence interval that includes the same intersubject variability that constituted the error variance in the between-subject design.

Intuitively, this seems wrong. An immediately obvious difficulty is that such a confidence interval would yield a different conclusion than that yielded by the within-subject ANOVA. We argued earlier that the Figure-2B confidence interval shows graphically that we could not make any strong inferences about the ordering of the three condition means (e.g., we could not reject the null hypothesis of no differences). In the between-subject example, this conclusion was entirely in accord with nonsignificant F yielded by the between-subject ANOVA. In the within-subject counterpart, however, such a conclusion would be entirely at odds with the highly significant F yielded by the within-subject ANOVA. This conflict is no quirk; it occurs because the intersubject variance, which is irrelevant in the within-subject ANOVA, partially determines (and in this example would almost completely determine) the size of the confidence interval. More generally, because the ANOVA and the confidence interval are based on different error terms, they provide different (and seemingly conflicting) information.

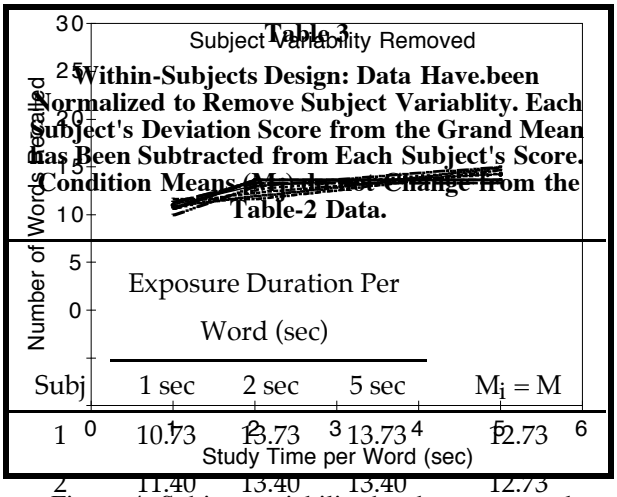

Figure 4. Subject variability has been removed from the Figure-2 data using the procedure described in the text.

Table 3

Within-Subjects Design: Data Have.been Normalized to Remove Subject Variablity. Each Subject's Deviation Score from the Grand Mean has Been Subtracted from Each Subject's Score. Condition Means Are Unchanged From the Table-2 Data.

1	10.73	13.73	13.73	12.73
2	11.40	13.40	13.40	12.73
3	10.73	13.73	13.73	12.73
4	11.40	12.40	14.40	12.73
5	10.73	12.73	14.73	12.73
6	11.40	13.40	13.40	12.73
7	11.73	11.73	14.73	12.73
8	10.07	13.07	15.07	12.73
9	10.73	13.73	13.73	12.73
10	11.07	12.07	15.07	12.73
M_j	M_1 = 11.0	M_2 = 13.0	M_3 = 14.2	M = 12.73

A Within-Subject Confidence Interval

To escape this conundrum, one can reason as follows. Given the irrelevance of intersubject variance in a within-subject design, it can legitimately be ignored for purposes of statistical analysis. In Table 3 we have eliminated intersubject variance without changing anything else. In Table 3, each of the three scores for a given subject has been *normalized* by subtracting from the original (Table-2) score a subject-deviation score consisting of that subject's mean, M_i (rightmost column of Table 2) minus the grand mean, M = 12.73 (Table 2, bottom right). Thus each subject's pattern of scores over the three conditions remains unchanged, and in addition the three condition means remain unchanged. But, as is evident in Table 3, rightmost column, each subject has the same normalized mean, equal to 12.73, the grand mean.

Figure 4 shows the data from Table 3; it is the Figure 3 data minus the subject variability. As shown in Appendix A(3), there are now only two sources of variability in the data: the condition variance is, as usual, reflected by the differences among the three condition means, while the remaining variance—the interaction variance—is reflected by the variability of points around each of the three means.

Figure 5A shows the Figure-4 data redrawn with the individual-subject curves removed, leaving only the mean curve and the individual data points. It is evident that there is an intimate correspondence between Figure 5A and Figure 2A. In both cases, the condition means are shown surrounded by the individual data points, and in both cases, the variability of the individual points around the condition means represents the error variance used to test for the Condition effect in the ANOVA. Intuitively therefore, it is sensible to compute from the Figure-5A data something very much like the between-subject confidence interval that was computed from the Figure 2A data (cf. Figure 2B). Because the variability in Figure 5A is entirely interaction variance, the appropriate formula is, as shown in Appendix A(3),

$$CI = M_j \pm \sqrt{\frac{MS_{SxC}}{n}} \text{ [criterion t(18)]}$$

which, in this example, is ±0.52. More generally,

$$CI = M_j \pm \sqrt{\frac{MS_{SxC}}{n}} \text{ [criterion t}(df_{SxC})\text{]} \quad (2)$$

Thus, Equation 2 embodies a within-subjects confidence interval. Note that there are two differences between Equations 1 and 2. First, the "error variance" in Equation 2 is the interaction mean squares rather than the within mean squares. Second, the criterion t in Equation 2 is based on df_{SxC} rather than df_W.

Figure 5B shows the resulting confidence intervals around the three condition means. It is abundantly clear that the information conveyed by Figure 5B mirrors the result of the ANOVA, clearly implying differences among the population means. (To illustrate this clarity *a fortiori*, the small plot embedded in Figure 5B shows the same data with the ordinate appropriately rescaled.) We emphasize that this confidence interval, and the associated ANOVA, now provide concordant information because they are based on the same error term (MS_{SxC})—just as in a between-subjects design, the ANOVA and a confidence interval provide concordant information because they are both based on the same error term, MS_W.

Inferences About Patterns of Population Means

As we have noted, the "confidence interval" generated by Equation 2 is not a *bona-fide* confidence interval, in the sense that it does not provide information about the value of some relevant population mean. We have also noted that in either a between- or a within-subject design, a *bona-fide* confidence interval—one truly designed to provide information about a population mean's value—must be based on intersubject variance as well as interaction variance. However, this Figure-5B confidence interval has an important property that justifies its use in a typical within-subject design. This property has to do with inferring *patterns* of population means across conditions.

Earlier we argued that a psychologist is typically interested not in the specific values of relevant population means, but instead is interested in the pattern of population means across conditions. In the present hypothetical study for example it might, as noted, be of interest to confirm a hypothesis that the three condition population means form a monotonically increasing sequence.

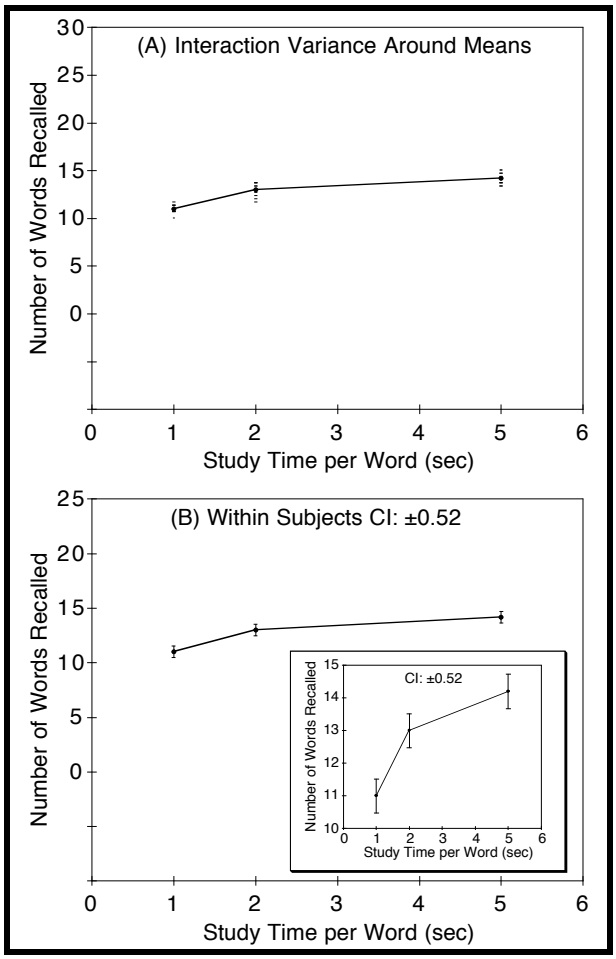

Figure 5. Construction of a within-subject confidence interval. Top panel: the only remaining variance is interaction variance. Bottom panel: a confidence interval constructed on the basis of the top-panel data. Note the analogy between this Figure and Figure 1.

In a within-subject design, as in a between-subject design, an ANOVA is designed to address the question: are there any differences among the population means? The within-subject confidence interval addresses the same question. In its simplest form, the question boils down to: are two sample means significantly different? In a between-subject design, there is a precise correspondence between the results of an ANOVA and the results of using confidence intervals: as shown in Appendix A(1) two sample means, M_j and M_k are significantly different given a particular α if and only if

$$|M_j - M_k| > \sqrt{2} \times CI$$

where "CI" is the (100 x (1.0 - □))% confidence interval. As demonstrated in Appendix A(3), the within-subjects confidence interval also has the property that it is related by a factor of $\sqrt{2}$ to the confidence interval around the difference between two means.

In summary, a between-subjects and a within-subjects confidence interval function similarly in two ways. First they both provide information that is consistent with that provided by the ANOVA, and second they both provide a clear, direct picture of the (presumably important) underlying pattern of population means. In addition, they both provide a clear, direct picture of relevant statistical power in that, smaller the confidence interval, the greater the power.

Additional Issues

The foregoing constitutes the major thrust of our remarks. In this section, we address a number of other issues involving the use of confidence intervals in general and within-subjects confidence intervals in particular.

Assumptions

In our discussions thus far, we have made the usual assumptions (see Appendix A for a description of them). In this section, we discuss several issues regarding effects of and suggested procedures to be used in the event of assumption violations.

In repeated-measures ANOVAs applied to cases in which there are more than two conditions, the computed F-ratio is, strictly speaking, correct only under the assumption of *sphericity*. A strict form of sphericity (called *compound symmetry*) requires that population variances for all conditions be equal (homogeneity of variance) and that the correlations between each pair of conditions be equal (homogeneity of covariance). If the sphericity assumption is violated (and it is arguable that this typically is the case, e.g., O'Brien & Kaiser, 1985), two problems arise. First, the F-ratio for the test of the conditions effect tends to be inflated (Box, 1954). Corrections for this problem have been developed in which the degrees of freedom used to test the obtained F-ratio are adjusted according to the seriousness of the departure from sphericity (Greenhouse & Geisser, 1959; Huynh & Feldt, 1976).

Second, violation of the sphericity assumption compromises the use of the omnibus error term (and its associated degrees of freedom) when testing planned or other types of contrasts. The omnibus error term is the average of the error terms associated with all possible one-degree-of-freedom contrasts that could be performed with the set of conditions that were tested. When sphericity is violated, these specific error terms may vary widely, so the omnibus error term is not necessarily a valid estimate of the error term for a particular contrast (O'Brien & Kaiser, 1985).

One solution to the problem of violation of the sphericity assumption is to conduct a multivariate analysis of variance (MANOVA) in place of a univariate analysis of variance, an approach that some advocate as a general solution (e.g., O'Brien & Kaiser, 1985). The MANOVA test avoids the problem of sphericity because it does not use pooled error terms. Instead, MANOVA is a multivariate test of a set of orthogonal, one-degree-of-freedom contrasts, with each contrast treated as a separate variable (not pooled as in ANOVA).

The use of MANOVA in place of ANOVA for repeated measures designs is not, however, universally recommended. For example, Hertzog and Rovine (1985) recommend estimating violations of sphericity using the measure □ as an aid in deciding whether to use MANOVA in place of ANOVA (e.g., Huynh & Feldt, 1970). Huynh and Feldt point out that such violations do not substantially influence the type I error rate associated with univariate F-tests unless □ is less than about 0.75. For values of □ between 0.90 and 0.75, Hertzog and Rovine recommend using the F-tests with adjusted degrees of freedom, and only for values of □ below 0.75 do they suggest using MANOVA.

More important for our purposes is that the sphericity assumption problem arises only when considering omnibus tests. As soon as one considers specific, one-degree-of-freedom contrasts, as is often done after MANOVA is applied, the sphericity assumption is no longer in effect. Thus, a viable solution is to use the appropriate specific error term for each contrast (e.g., Boik, 1981) and avoid the sphericity assumption altogether.

The problems that result from violation of sphericity have implications for the implementation of confidence intervals as graphic aids and as alternatives to hypothesis testing. The computation of the confidence interval as shown in Equation 2 uses the omnibus error term, and is the interval that would be plotted with each mean, as in Figure 5. Given that a crucial function served by the plotted confidence interval is to provide an impression of the pattern of differences among means, we must be sensitive to the possibility that violation of sphericity causes an underestimate of the interval's size.

To counteract the underestimation stemming from inappropriately high degrees of freedom, one could use the Greenhouse-Geisser or Huynh-Feldt procedure (as computed by ANOVA packages such as BMDP) to adjust the degrees of freedom used in establishing the criterion t-value.

It is important to note that although the confidence interval computed by applying the adjustment to degrees of freedom may be used to provide a general sense of the pattern of means, more specific questions about pairs of means should be handled differently. If the omnibus error term is not appropriate for use in contrasts when sphericity is violated, then the confidence interval plotted with each mean should be based on a specific error term. The choice of the error term to use will depend on the contrast that is of interest. For example, in Figure 5 it might be important to contrast the first and second duration conditions and the second and third conditions. The confidence interval plotted with the means of the first and second conditions would be based on the error term for contrasting those two conditions. The confidence interval plotted with the mean for the third condition would be based on the error term for the contrast between the second and third conditions. For ease of comparison, one might plot both intervals, side by side, around the mean for the second condition. The choice of which interval(s) to plot will depend on the primary message that the graph is intended to convey. Below we provide a specific example of plotting multiply-derived confidence intervals to illustrate different characteristics of the data.

Another means of treating violation of the homogeneity of variance assumption is to compute separate confidence intervals for the separate condition means. In a between-subjects design, this is a simple procedure: one estimates the population variance for each group, j (MS_{Wj} based on n_j - 1 degrees of freedom, where n_j is the number of observations in Group j) and then computes the confidence interval for that group as

$$CI_j = \sqrt{\frac{MS_{Wj}}{n_j}} \times [\text{criterion } t(n_j - 1)]$$

An analogous procedure for a within-subjects design is described in Appendix B. The general idea underlying this procedure is that one allows the subject-by-condition interaction variance to differ from condition to condition; the confidence interval for condition j is then based primarily on the interaction variance from Condition j. The equation for computing the best estimate of this Condition-j interaction variance ("$estimator_j$") is,

$$estimator_j = \left[\frac{J}{J-1} MS'_{Wj}\right] - \left[\frac{MS_{SxC}}{J}\right]$$

Here, MS_{SxC} is the overall mean square due to interaction, and

$$MS'_{Wj} = \frac{\sum (y'_{ij} - M_j)^2}{n(n-1)} = \frac{\sum_i y'^2_{ij} - T_j^2/n}{n(n-1)}$$

(where T_j is the Group-j total and again n is the number of subjects). Thus, MS'_{Wj} is the "mean-square within" obtained from Condition j of the *normalized* (y'_{ij}) scores (e.g., in this article's example a mean square within a given column of Table 3). Having computed the estimator, the Group-j confidence interval is computed as,

$$CI_j = \sqrt{\frac{estimator_j}{n}} \times \text{criterion } t\,(n - 1)(3)$$

Mean Differences

Above we discussed the relation between confidence intervals around sample means and around the difference between two sample means. Because this relation is the same (involving a factor of $\sqrt{2}$) for both between-subjects and the within-subjects confidence intervals, one could convey the same informa-

tion in a plot by simply including a single confidence interval appropriate for the difference between two sample means.

Which type of confidence interval is preferable is partly a matter of taste, but is also a matter of the questions being addressed in the experiment. Our examples in this article involved parametric experiments in which an entire pattern of means was at issue. In our hypothetical experiments, one might ask, for instance, whether the relation between study time and performance is monotonic, or perhaps whether it conforms to some more specific underlying mathematical function, such as an exponential approach to an asymptote. In other experiments, more qualitative questions are addressed (e.g., "what are the relations among conditions involving a positive, neutral, or negative prime?") Here, the focus would be on specific comparisons between sample means, and a confidence interval of mean differences might be more useful.

Multifactor Designs

The logic that we have presented here is based on a simple design in which there is only a single factor that is manipulated within subjects. In many experiments, however, there two or more factors. In such cases, all factors may be manipulated within subjects, or some factors may be within subjects, while others are between subjects.

Multifactor Within-Subjects Designs

Consider a design in which there are two fixed factors, A and B, with J and K levels per factor, combined with n subjects. In such a design, there are three error terms, corresponding to the interactions of subjects with factors A, B, and the AxB interaction. Roughly speaking, one of two basic results can occur in this design: either the three error terms are all approximately equal, or they differ substantially from one another.

As discussed in any standard design text (e.g., Winer, 1971) when the error terms are all roughly equal, they can be pooled by dividing the sum of the three sums of squares by the sum of the three degrees of freedom (which amounts to treating the design as if it were a single-factor design with JK conditions). A single confidence interval can then be computed using Equation 2,

$$CI = M_j \pm \sqrt{\frac{MS_{SxAB}}{n}} \; [\text{criterion } t(df_{SxAB})] \qquad (4)$$

where SxAB refers to the interaction of subjects with the combined JK conditions formed by combining factors A and B (based on (n - 1)(JK - 1) degrees of freedom). This confidence interval is appropriate for comparing any two means (or any pattern of means) with one another.

As discussed above, the use of the omnibus error term depends on meeting the sphericity assumption. When this assumption is untenable (as indicated, for example, by a low value of ε computed in conjunction with the Greenhouse-Geisser or Huynh-Feldt procedure for corrected degrees of freedom, or by substantially different error terms for main effects and interactions involving the repeated-measures factors), different mean differences are distributed with different variances, as shown in the Appendix A(4). For instance, the standard error appropriate for assessing $(M_{jk} - M_{jr})$ may be different from that appropriate for assessing $(M_{jk} - M_{qk})$ or $(M_{jk} - M_{qr})$. In such cases, one should adopt the strategy of plotting confidence intervals that can be used to assess patterns of means or contrasts that are of greatest interest. One might even plot more than one confidence interval for some means, or construct more than one plot for the data. Finally, one could treat the design as a one-way design with "conditions" actually encompassing all JxK cells; one could then drop the homogeneity-of variance assumption and compute an individual confidence interval for each condition as discussed in the "Assumptions" section above (see Equation 3). Here the interaction term would be MS_{SxAB} described as part of Equation 4.

Mixed Designs

Other designs involve one or more factors manipulated within subjects in conjunction with one or more factors manipulated between subjects. Here, matters are further complicated, as evaluation of the between-subjects effect is almost always based on a different error term than is evaluation of the within-sub-

jects or the interaction effects. Here again, one could, at best, construct different confidence intervals depending on which mean differences are to be emphasized.

Data Reduction in Multifactor Designs

An alternative to treating multifactor data as simply a collection of (say) JxK means is to assume a model that implies some form of preliminary data reduction. Such data reduction can functionally reduce the number of factors in the design (e.g., could reduce a two-fixed-factor design to a one-fixed-factor design).

An Example. To illustrate, suppose that one were interested in slope differences between various types of stimulus materials (e.g., digits, letters, words) in a Sternberg (1966) memory-scanning task. One might design a completely within-subjects experiment in which J levels of set size were factorially combined with K levels of stimulus type and n subjects. If it were assumed that the function relating reaction time to set size were fundamentally linear, then one could compute a slope for each subject, thereby functionally reducing the design to a one-factor (stimulus type), within-subject design in which "slope" was the dependent measure. Confidence intervals around mean slopes for each stimulus-type level could be constructed in the manner that we have described. Alternatively, if stimulus type were varied *between* subjects, then computing a slope for each subject would allow one to treat the design as one-way, between-subjects design (again with "slope" as the dependent measure), and standard between-subjects confidence intervals could be computed.

Contrasts. The slope of an assumed linear function is, of course, a special case of a one-degree-of-freedom contrast by which a single dependent variable can be computed from a J-level factor as,

$$y = \sum_j w_j M_j$$

where the M_j are the means of the j levels and the w_j (constrained such that $\sum w_j = 0$) are the weights corresponding to the contrast. Thus the above examples can be generalized to any case in which the effect of some factor can be reasonably well specified.

The Case of a J x 2 Design. One particular fairly common situation bears special mention. When the crucial aspect of a multifactor design is the interaction between two factors, and one of the factors has only two levels, the data can be reduced to a set of J difference scores. These difference scores can be plotted along with the confidence interval computed from the error term for an ANOVA of the difference scores. A plot of this kind addresses whether the differences between means are different and provides an immediate sense of (1) whether an interaction is present and (2) the pattern of the interaction. Such a plot can accompany the usual plot showing all condition means.

To illustrate the flexibility of this approach, consider a semantic priming experiment in which subjects name target words that are preceded by either a semantically related or unrelated prime word. Prime relatedness is factorially combined within subjects with the prime-target stimulus onset asynchrony (SOA) which, suppose, is 50, 100, 200, or 400 ms. Hypothetical response latency data from six subjects are shown in Table 4. The mean latency for each of the eight conditions is plotted in Figure 6A. Confidence intervals in Figure 6A are based on the comparison between related and unrelated prime conditions within a particular SOA (i.e., on the 5-degree-of-freedom error terms stemming from individual two-level one-way ANOVAs performed at each SOA level). This plot thus illuminates the degree to which priming effects are reliable at the different SOAs.

Table 4

Data (Reaction Times) from Six Subjects in a Hypothetical Priming Experiment. Four Values of SOA are Combined with Two Priming Conditions (Colums Labeled R: Primed; Columns Labeled U: Unprimed). Columns Labeled D Represent Unprimed Minus Primed Difference Scores at each SOA Level.

	SOA = 50 ms			SOA = 100 ms			SOA = 200 ms			SOA = 400 ms		
Subject	R	U	D	R	U	D	R	U	D	R	U	D
1	450	462	12	460	482	22	460	497	37	480	507	27
2	510	492	-18	515	530	15	520	534	14	504	550	46
3	492	508	16	512	522	10	503	553	50	520	539	19
4	524	532	8	530	543	13	517	546	29	503	553	50
5	420	409	-11	424	452	28	431	468	37	446	472	26
6	540	550	10	538	528	-10	552	575	23	562	598	36
M_j	489	492	3	497	510	13	497	529	32	503	537	34

Suppose that a further avenue of investigation revolves around the degree to which priming effects differ in magnitude across the SOAs. In ANOVA terms the question would be: is there a reliable interaction between prime type and SOA? A standard 4x2 within-subjects ANOVA applied to these data shows that the interaction is significant , $F(3, 15) = 7.08$, $MS_{SxC} = 95.39$. The nature of the interaction can be displayed by plotting mean difference scores (which, for individual subjects, are obtained by subtracting the latency in the related prime condition from the latency in the unrelated prime condition) as a function of SOA. These difference scores (representing priming effects) are included in Table 4, and the mean difference scores are plotted in Figure 6B. The confidence intervals in Figure 6B are based on the MS_{SxC} term for a one-factor repeated measures ANOVA of the difference scores ($MS_{SxC} = 190.78$). (Note that, as with any difference score, the error term in this ANOVA is twice the magnitude of the corresponding error terms in the full, two-factor ANOVA that generated the F-ratio for the interaction.) Examination of the bottom panel of Figure 6 indicates immediately that indeed, reliably different priming effects occurred at different SOAs (consistent with the significant interaction obtained in the two-factor ANOVA) and also reflects the range of pattern that this interaction could assume.

Knowledge of Absolute Population Means

Central to our reasoning up to now is that knowledge of absolute population means is not critical to the question being addressed. Although this is *usually* true, it is not, of course, *always* true. For instance, one might be carrying out a memory experiment in which one were interested in whether performance in some condition differed from a 50% chance level. In this case, the within-subjects confidence interval that we have described would be inappropriate. If one were to use a confidence interval in this situation, it would be necessary to use the confidence interval that included the between-subject variation that we removed in our examples.

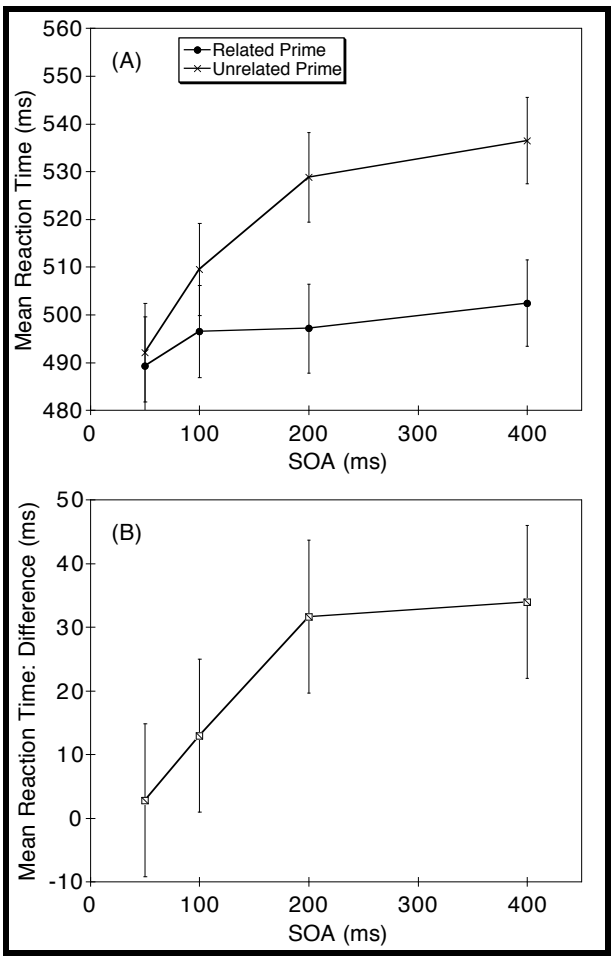

Figure 6. Illustration of multiple ways of plotting to demonstrate different aspects of the data. Panel A: Mean RT as a function of the 8 conditions. Confidence intervals at each SOA level is based on the 5-df error term from the one-way, two-level (primed vs unprimed) ANOVA done at that SOA level. Panel B: Mean (primed - unprimed) RT difference. Confidence intervals are based on the SxC error term from the ANOVA of RT differences.

Meta-Analysis

One advantage of reporting the results of ANOVA and tables of means and standard deviations is that it makes tasks associated with meta-analysis easier and more precise. In cases in which an author relies on graphical depictions of data using confidence intervals, as described here, it would be helpful to include in the Results section a statement of the effect size associated with each main effect, interaction, or other contrast of interest in the design. This information is not typically included even in articles that apply standard hypothesis testing techniques with ANOVA. All researchers would benefit if both the hypothesis testing method and the graphical approach advocated here were supplemented by estimates of effect size.

Conclusions: Data Analysis as Art, not Algorithm

In this article, we have tried to accomplish a specific goal: to describe an appropriate and useful confidence interval to be used in within-subjects designs that serves the same functions as does a confidence interval in a between-subjects design. Although we have attempted to cover a variety of "ifs, ands, and buts" in our suggestions, we obviously cannot cover all of them. We would like to conclude by underscoring our belief that each experiment constitutes its own data-analysis challenge in which (1) specific (often multiple) hypotheses are to be evaluated, (2) standard assumptions may (or may not) be violated to varying degrees, and (3) certain sources of variance or covariance are more important than others. Given this uniqueness, it is almost self-evident that no one set of algorithmic rules can appropriately cover all possible situations.

REFERENCES

Anderson, V. & McLean, R.A. (1974). *Design of Experiments: A Realistic Approach*. New York: Marcel Dekkar Inc.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423-437.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society, 53*, 370-418.

Berger, J.O. & Berry, D.A. (1988). Statistical analysis and the illustion of objectivity. *American Scientist, 76*, 159-165.

Boik, R. J. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. *Psychometrika, 46*, 241-255.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effect of inequality of variance and of correlation between errors in the two-way classifica-

tion. *Annals of Mathematical Statistics, 25*, 484-498.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

Box, G. E. P. (1986). An apology for ecumenism in statistics. In G. E. P. Box, T. Leonard, & C.-F. Wu (Eds.), *Scientific Inference, Data Analysis, and Robustness* (pp. 51-84). new york: academic press.

Camilli, G. (1990). The test of homogeneity for 2 X 2 contingency tables: A review of and some personal opinions on the controversy. *Psychological Bulletin, 108*, 135-145.

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

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society, 98*, 39-54.

Fisher, R. A. (1947). *The Design of Experiments*. New York: Hafner Press.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Ser. B, 17*, 69-78.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The Empire of Chance*. Cambridge: Cambridge University Press.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95-112.

Hays, W. (1973). *Statistics for the Social Sciences* (second edition). New York: Holt.

Hertzog, C., & Rovine, M. (1985). Repeated-measures analysis of variance in developmental research: Selected issues. *Child Development, 56*, 787-809.

Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measures designs have exact F distributions. *Journal of the American Statistical Association, 65*, 1582-1589.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in the randomized block and split plot designs. *Journal of Educational Statistics, 1*, 69-82.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association, 88*, 1242-1249.

Lewis, C. (1993). Bayesian methods for the analysis of variance. In Kerens, G. and Lewis, C. (Eds.) *A Handbook for Data Analysis in the Behavioral Sciences: Statistical Issues*. Hillsdale NJ: Erlbaum.

Loftus, G.R. & Loftus, E.F. (1988). *Essence of Statistics*, 2nd. Edition (New York: Random House).

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

Loftus, G.R. (1993a). Editorial Comment. *Memory & Cognition, 21*, 1-3.

Loftus, G.R. (1993b). Visual data representation and hypothesis testing in the microcomputer age. *Behavior Research Methods, Instrumentation, and Computers, 25*, 250-256.

Loftus, G.R. (1993c). On the overreliance on hypothesis testing in the social sciences (talk presented at the Psychonomic Society meetings, Washington DC,).

Neyman, J. (1957). "Inductive behavior" as a basic concept of philosophy of science. *Review of the International Statistical Institute, 25*, 7-22.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 20A*, 175-240, 263-294.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical transactions of the royal society of london, Ser. A, 231*, 289-337.

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin, 97*, 316-333.

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153*, 652-654.

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

Tufte, E.R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tukey, J.W. (1974). The future of data analysis. *Annals of Mathematical Statistics, 33*, 1-67.

Tukey, J.W. (1977) *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Wainer, H. & Thissen, D. (1993) Graphical data analysis. In Kerens, G. and Lewis, C. (Eds.) *A Handbook for Data Analysis in the Behavioral Sciences: Statistical Issues*. Hillsdale NJ: Erlbaum.

Winer, B.J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Winkler, R.L. (1993) Bayesian statistics: An overview. In Kerens, G. and Lewis, C. (Eds.) *A Handbook for Data Analysis in the Behavioral Sciences: Statistical Issues*. Hillsdale NJ: Erlbaum.

Appendix A

We begin by considering a between-subjects design, and providing the logic underlying the computation of the usual standard error of the mean. We then articulate the assumptions of the within-subjects standard error, and demonstrate its relation to its between-subjects counterpart. Our primary goal is to show that the within-subjects standard error plays the same role as does the standard, between-subjects standard error in two senses: its size is determined by the error term used in the ANOVA; and it is related by a factor of $\sqrt{2}$ to the standard error of the difference between two means. Appendix sections are numbered for ease of reference in the text.

(1) Between-Subjects Designs

Consider a standard one-factor, between-subjects design in which subjects from some population are assumed to be randomly sampled and randomly assigned to one of J conditions. Each subject thus contributes one observation to one condition. All observations are independent of one another. Because we are primarily concerned with comparing between- to within-subjects designs, we lose little generality by assuming that there are equal numbers, n, of subjects in each of the J conditions. The fixed-effects linear model underlying the ANOVA is,

$$y_{ij} = \mu + \alpha_j + \beta_j + g_{ij} \qquad (1)$$

Here, y_{ij} is the score obtained by Subject i in Condition j, μ is the grand population mean, α_j is an effect of the Condition-j treatment ($\sum \alpha_j = 0$), β_j is an effect due to Subject ij, and g_{ij} is an interaction effect of Subject ij's being in Condition j. We include both β_j and g_{ij} for completeness although they cannot, of course, be separated in a between-subjects design. We assume that the β_j are normally distributed over subjects in the population with means of zero and variances of σ^2_β. We assume that the g_{ij} are likewise normally distributed over subjects in the population with means of zero and variances of σ^2_g for all J conditions. We assume that for each condition, j, the β_j and the g_{ij} are independent and that the g_{ij} are independent of one another over conditions. Notationally, we let $\beta_j + g_{ij} = e_{ij}$, which, given our assumptions so far, means that we can de-

fine "error variance," σ^2_e, to be $\sigma^2_\alpha + \sigma^2_g$ and that Equation 1 can be rewritten as

$$y_{ij} = \mu + \alpha_j + e_{ij} \qquad (2)$$

As is demonstrated in any standard statistics textbook (e.g., Hays, 1973, Chapters 12-13), the following is true given the model embodied in Equation 2 and the assumptions that we have articulated.

1. The expectation of M_j, the mean of Condition j, is μ_j, where μ_j, the population mean given condition j, is equal to $\mu + \alpha_j$.

2. An unbiased estimate of the error variance, σ^2_e, is provided by (MS_W) based on J(n - 1) degrees of freedom. The condition means, M_j are distributed over samples of size n with variance σ^2_e/n. Thus the standard error of M_j, SE_j, computed as $\sqrt{MS_W/n}$, is, determined by the same quantity (MS_W) that constitutes the error term in the ANOVA.

3. The standard error of the *difference* between any two means, M_j and M_k, is computed by,

$$SE_{j-k} = \sqrt{2 \times \frac{MS_W}{n}} = \sqrt{2} \times SE_j$$

Therefore, the standard error of the mean and the standard error of the difference between two means are related by $\sqrt{2}$. A corollary of this conclusion is that two means, M_j and M_k, are significantly different by a two-tailed t-test at some significance level, x, if and only if,

$$\frac{|M_j - M_k|}{\sqrt{2 \times \frac{MS_W}{n}}} > \text{Criterion } t(df_W)$$

where Criterion $t(df_W)$ is two-tailed at the (100 x (1.0 - x))% level, or,

$$\frac{|M_j - M_k|}{\sqrt{2}} > \sqrt{\frac{MS_W}{n}} \text{ Criterion } t(df_W) = CI$$

where "CI" at the right of the equation refers to the (100 x (1.0 - x)) confidence interval. Thus, as asserted in the text, M_j and M_k differ significantly at the x% level when,

$$|M_j - M_k| > \sqrt{2} \times CI$$

(2) Within-Subjects Designs

Now consider a standard one-factor, within-subjects design in which n subjects from the population are assumed to be randomly sampled; however, each subject participates in all J conditions. Each subject thus contributes one observation to each condition. The linear model underlying the ANOVA is,

$$y_{ij} = \mu + \alpha_j + \pi_i + g_{ij} \qquad (3)$$

As above, y_{ij} is the score obtained by Subject i in Condition j, μ is the population mean, and α_j is an effect of the Condition-j treatment. Again π_i is an effect due to Subject i (note that π_i now has only a single subscript, i, since each subject participates in all conditions). Again, g_{ij} is the interaction effect of Subject i's being in Condition j. We make the same assumptions about π_i and the g_{ij} as we did in the preceding section.

The mean of Condition j, M_j, has an expectation of,

$$E(M_j) = E[\frac{1}{n_i}(\mu + \alpha_j + \pi_i + g_{ij})]$$

$$= E[\mu + \alpha_j + \frac{1}{n_i}(\pi_i + g_{ij})]$$

$$= \mu + \alpha_j + \frac{1}{n_i}E(\pi_i) + \frac{1}{n_i}E(g_{ij})$$

or, because $E(\pi_i) = E(g_{ij}) = 0$ for all j,

$$E(M_j) = \mu + \alpha_j = \mu_j \qquad (4)$$

where μ_j is the population mean given condition j.

The expectation of $(y_{ij} - M_j)^2$ is,

$$E(y_{ij} - M_j)^2 =$$

$$E[(\alpha_j + \pi_i + g_{ij} - \frac{1}{n_i}(\alpha_j + \pi_i + g_{ij})]^2$$

which reduces to,

$$E(y_{ij} - M_j)^2 = (\sigma_\varepsilon^2 + \sigma_g^2)[(n - 1)/n] \quad (5)$$

Thus, the variablity of the y_{ij} scores around M_j includes variance both from subjects (σ_ε^2) and from interaction (g_{ij}).

The variance of the M_j's around μ_j is,

$$E(M_j - \mu_j)^2 = E[\frac{1}{n}\sum_i(\mu_j + \varepsilon_i + g_{ij} - \mu_j)]^2$$

$$= E[\frac{1}{n}\sum_i(\varepsilon_i + g_{ij})]^2$$

$$= E[\sum_i \varepsilon_i^2/n + \sum_i g_{ij}^2/n]$$

$$= (\sigma_\varepsilon^2 + \sigma_g^2)/n \quad (6)$$

That is, over random samples of subjects, the variability of the M_j's includes both subject and interaction variability. An unbiased estimate of $(\sigma_\varepsilon^2 + \sigma_g^2)/n$ is obtained by $(MS_S + MS_{SxC})/n$. Therefore, the bona fide standard error of M_j is $\sqrt{(MS_S + MS_{SxC})/n}$.

(3) Removal of Intersubject Variance

We now consider our proposed correction to each score designed to remove subject variance (that resulted in the transformation from Table 2 to Table 3 in the text). This correction consisted of subtracting from each of Subject i's scores an amount equal to Subject i's over-condition mean, M_i, minus the grand mean, M. Thus, the equation for the transformed dependent variable, y'_{ij}, is

$$y'_{ij} = \mu + \varepsilon_j + \varepsilon_i + g_{ij} - M_i + M \quad (7)$$

It can easily be demonstrated that the transformed mean of Condition j, M'_j, equals the untransformed mean, M_j. A comparison of Equations (3) and (7) indicates that M'_j and M_j differ by mean over the n subjects of $(M - M_i)$, or,

$$M'_j - M_j = \frac{1}{n}\sum_i(M - M_i) = M - \frac{1}{nJ}\sum_j\sum_i y_{ij} =$$

$$= M - \frac{1}{Jn}(JnM) = 0$$

which means that $M_j = M'_j$. Therefore, by Equation (4), we conclude that the expectation of $M'_j = M_j$, the mean of condition j, is μ_j, the population mean given condition j.

Next, we consider the within-condition variance of the y'_{ij} scores. The variance of the $(y'_{ij} - M_j)$ scores is,

$$E(y'_{ij} - M_j)^2 = E(y_{ij} - M_i + M - M_j)^2 =$$

$$= E[\mu + \varepsilon_j + \varepsilon_i + g_{ij} - \frac{1}{J}\sum_j(\mu + \varepsilon_j + \varepsilon_i + g_{ij}) +$$

$$+ \frac{1}{Jn}\sum_j\sum_i(\mu + \varepsilon_j + \varepsilon_i + g_{ij}) - \frac{1}{n}\sum_i(\mu + \varepsilon_j + \varepsilon_i + g_{ij})]^2$$

which can be reduced to,

$$E(y'_{ij} - M_j)^2 = \sigma_g^2[(n - 1)/n]$$

Thus the within-cell variance of the y'_{ij} scores includes only the interaction component. Moreover, the only additional variance of the y'_{ij} scores is variance due to conditions.

We have asserted in the text that the variance of the y'_{ij} scores within each condition plays a role analogous to the variance of the individual subject scores within each condition of a between-subjects design. More precisely, we consider a "sum of squares within" over the y'_{ij} scores, which can be computed as:

$$SS'_W = \sum_j\sum_i (y'_{ij} - M_j)^2$$

$$= \sum_j\sum_i (y_{ij} - M_i + M - M_j)^2$$

which reduces to,

$$SS'_W = \sum_j\sum_i y_{ij}^2 - J\sum_i M_i^2 - n\sum_j M_j^2 + JnM^2$$

which is equal to sum of squares due to the subject by condition interaction. This means that the variance of the y'_{ij} scores within each condition are distributed with a variance that can be estimated by MS_{SxC}, whose expectation is σ^2_g. Therefore the standard error is computed by,

$$SE_j = \sqrt{\frac{MS_{SxC}}{n}}$$

As in the between-subject case, accordingly, the size of the standard error is determined by the same quantity (MS_{SxC}) that constitutes the error term in the ANOVA.

Now, consider the difference between two sample means,

$$\begin{aligned}(M_j - M_k) &= \\ &\frac{1}{n}\sum_i(\mu + \alpha_j + \gamma_i + g_{ij}) - \frac{1}{n}\sum_i(\mu + \alpha_k + \gamma_i + g_{ik}) \\ &= \frac{1}{n}\sum_i(\alpha_j - \alpha_k + g_{ij} - g_{ik}) \\ &= (\alpha_j - \alpha_k) + \frac{1}{n}\sum_i(g_{ij} - g_{ik}) \qquad (6)\end{aligned}$$

The g_{ij} and g_{ik} are assumed to be distributed independently, with means of zero; thus the expectation of $(M_j - M_k)$ is $(\alpha_j - \alpha_k)$. By the homogeneity of variance assumption, the g_{ij} and g_{ik} are distributed identically in Conditions j and k, with variance σ^2_g. The $(M_j - M_k)$ are therefore distributed with a variance equal to $2\sigma^2_g/n$ which, in turn, is estimated by $2MS_{SxC}/n$. This implies that the standard deviation of the difference between any two means, M_j and M_k, is

$$SE_{j-k} = \sqrt{\frac{2MS_{SxC}}{n}} = \sqrt{2}\ \text{x}\ SE_j$$

Therefore the standard error is related to the standard error of the difference between two means by $\sqrt{2}$, just as it is in a between-subjects case.

(4) Multifactor Designs

Consider a JxK design in which each of n subjects participates in each JK level. The model describing this situation is

$$y_{ij} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \gamma_i + g_{ijk} + ag_{ij} + bg_{ik} + abg_{ijk} \qquad (7)$$

Here, y_{ij}, μ, and α_j are as in Equation (3), while $\alpha\beta_{jk}$ is an effect due to the AxB interaction. The subject-by-cell interaction term, g_{ijk} acts like g_{ij} in Equation 3, but has an extra subscript to refer to all JK cells. The ag_{ij}, the interaction of subjects with factor A, sums to zero over the J levels of A, and is identical at each level k for each subject, i. Likewise, the bg_{ik}, the interaction of subjects with factor B, sums to zero over all K levels of B, and is identical at each level j for each subject, i. Finally, the abg_{ijk}, the interaction of subjects with the AxB interaction sum to zero both over all J levels of Factor A for each level k and across all K levels of Factor B for each level j.

There are three mean squares involving subject interactions: the AxS, BxS, and AxBxS interactions which are the ANOVA error terms for the effects of A, B, and AxB. As is shown in any standard design text (e.g., Winer, 1971, Chapter 7), the expected mean squares of each of these interactions (MS_{SxA}, MS_{SxB}, and MS_{SxAxB}) contains both a σ^2_g and a $\sigma^2_{\ }$ component along with another variance component corresponding to the specific effect ($\sigma^2_{\ }$, $\sigma^2_{\ }$, etc.). If, using standard procedures, one can infer that $\sigma^2_{ag} = \sigma^2_{bg} = \sigma^2_{abg} = 0$, then for making inferences about effects of A, B, or AxB, there remains only a single source of error, σ^2_g that is estimated by the pooled variances due to AxS, BxS, and AxBxS.

We have been emphasizing that confidence intervals—both standard confidence intervals and the within-subjects confidence intervals—are appropriate for assessing patterns of means or, most basically, differences between means. Given the model in Equation 7, the standard error of the difference between two means depends on which two means are being compared. In particular, for comparisons...

Across different columns within a row:

$$E(SE^2_{jk-qk}) = \frac{\sigma^2_g + \sigma^2_{ag} + \sigma^2_{abg}}{n}$$

Across different rows within a column:

$$E(SE^2_{jk\text{-}jr}) = \frac{\sigma^2_g + \sigma^2_{bg} + \sigma^2_{abg}}{n}$$

Across different rows and columns:

$$E(SE^2_{jk\text{-}qr}) = \frac{\sigma^2_g + \sigma^2_{ag} + \sigma^2_{bg} + \sigma^2_{abg}}{n}$$

The standard errors appropriate for for any arbitrary mean difference will, accordingly, be equal only when $\sigma^2_{ag} = \sigma^2_{bg} = 0$.

Appendix B

In our previous considerations, we have made a homogeneity-of-variance assumption for the g_{ij} (interaction components). That is, we have assumed that σ_{g_j} is identical for each of the j conditions. We now drop this assumption, and assume that the variance of the g_{ij} is σ_{g_j}. Denote the mean variance of the J σ_{g_j}'s (over the J conditions) as $\bar{\sigma}_g$.

Consider Group j. The variance of the g_{ij} for that group, σ_{g_j}, is estimated as follows. First, the expectation of the variance of the normalized scores, y'_{ij} is,

$$E\,(y'_{ij} - M_j)^2 = \\ = E(\mu + \alpha_j + \beta_i + g_{ij} - M_i + M - M_j)^2$$

which (after not inconsiderable algebra) reduces to,

$$E\,(y'_{ij} - M_j)^2 = \\ = f([\sigma^2_{g_j}(n-1)(J-2)] + [\bar{\sigma}^2_g(n-1)], nJ)$$

Thus the expectation of the sum of squares within the y'_{ij} scores of group j is,

$$E\left[\sum_i (y'_{ij} - M_j)^2\right] = \\ \frac{[\sigma^2_{g_j}(n-1)(J-2)] + [\bar{\sigma}^2_g(n-1)]}{J}$$

which means that the expectation of the mean square within Group j, MS'_{Wj}, is,

$$E\left[\frac{\sum_i (y'_{ij} - M_j)^2}{n-1}\right] = \sigma^2_{g_j}\left[\frac{J-2}{J}\right] + \frac{\bar{\sigma}^2_g}{J}$$

or,

$$E\left[\frac{\sum_i (y'_{ij} - M_j)}{n-1}\left[\frac{J}{J-2}\right]\right] = \sigma^2_{g_j} + \frac{\bar{\sigma}^2_g}{J-2}$$

Because the first factor in this equation is MS'_{Wj},

$$E\left[MS_{Wj}\left[\frac{J}{J-2}\right]\right] = \sigma^2_{g_j} + \frac{\bar{\sigma}^2_g}{J-2} \qquad (1)$$

It can be shown that the expectation of the overall mean square due to interaction, MS_{SxC} is $\bar{\sigma}^2_g$. Therefore,

$$E\left[\frac{MS_{SxC}}{J-2}\right] = \frac{\bar{\sigma}^2_g}{J-2} \qquad (2)$$

Substituting the left side of Equation 2 for the rightmost term of Equation 1 and rearranging

terms, the estimator of $\omega^2_{g_j}$ is,

$$E\left[\frac{J}{J-1}\left(MS'_{W_j} - \frac{MS_{SxC}}{J}\right)\right] = \omega^2_{g_j} \quad (3)$$

as indicated in the text.

There is one potential problem with the Equation-3 estimator: because it is the difference of two estimated variances it can turn out to be negative. In such an instance, two possible solutions are (1) to use the overall estimator or (2) to average estimates from multiple groups which, for some a priori reason can be considered to have equal variances.