Inference by Eye

Confidence Intervals and How to Read Pictures of Data

Geoff Cumming and Sue Finch
La Trobe University

Wider use in psychology of confidence intervals (CIs), especially as error bars in figures, is a desirable development. However, psychologists seldom use CIs and may not understand them well. The authors discuss the interpretation of figures with error bars and analyze the relationship between CIs and statistical significance testing. They propose 7 rules of eye to guide the inferential use of figures with error bars. These include general principles: Seek bars that relate directly to effects of interest, be sensitive to experimental design, and interpret the intervals. They also include guidelines for inferential interpretation of the overlap of CIs on independent group means. Wider use of interval estimation in psychology has the potential to improve research communication substantially.

Inference by eye is the interpretation of graphically presented data. On first seeing Figure 1, what questions should spring to mind and what inferences are justified? We discuss figures with means and confidence intervals (CIs), and propose *rules of eye* to guide the interpretation of such figures. We believe it is timely to consider inference by eye because psychologists are now being encouraged to make greater use of CIs.

Many who seek reform of psychologists' statistical practices advocate a change in emphasis from null hypothesis significance testing (NHST) to CIs, among other techniques (Cohen, 1994; Finch, Thomason, & Cumming, 2002; Nickerson, 2000). The American Psychological Association's (APA) Task Force on Statistical Inference (TFSI) supported use of CIs (Wilkinson & TFSI, p. 599), and the APA *Publication Manual* states that CIs "are, in general, the best reporting strategy" (APA, 2001, p. 22).

Statistical reformers also encourage use of visual representations that make clear what data have to say. Figures can "convey at a quick glance an overall pattern of results" (APA, 2001, p. 176). The TFSI brought together advocacy of CIs and visual representations by stating the following: "In all figures, include graphical representations of interval estimates whenever possible" (Wilkinson & TFSI, 1999, p. 601). In other words, CIs should be displayed in figures. We applaud this recommendation and believe it has the potential to enhance research communication in psychology. However, two difficulties are likely to hinder its adoption. First, according to evidence presented by Belia, Fidler, Williams, and Cumming (2004) and Cumming, Williams, and Fidler (2004), many researchers in psychology and some other disciplines have important misconceptions about CIs. Second, there are few accepted guidelines as to how CIs should be represented or discussed. For example, the *Publication Manual* (APA, 2001) gives no examples of CI use and no advice on style for reporting CIs (Fidler, 2002).

Four main sections follow. In the first, we discuss basic issues about CIs and their advantages. The second presents our rules of eye for the interpretation of simple figures showing means and CIs. Our main focus is CIs, but in the third section we discuss standard error (*SE*) bars. We close with comments about some outstanding issues.

CIs and Error Bars: Basic Issues

What Is a CI?

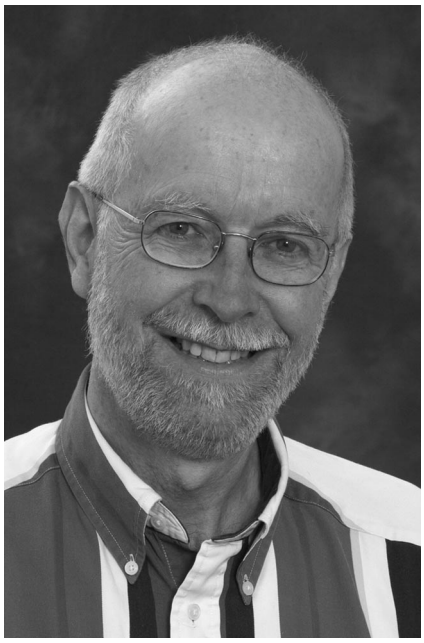
Suppose we wish to estimate the verbal ability of children in Melbourne, Australia. We choose a recognized test of verbal ability and are willing to assume its scores are normally distributed in a reference population of children. We test a random sample of Melbourne children ($n = 36$) and find the sample mean (M) is 62 and the sample standard deviation (SD) is 30. Then M is our *point estimate* of the population mean verbal ability of Melbourne children. We seek also a 95% CI, which is an *interval estimate* that indicates the precision, or likely accuracy, of our point estimate. The 95% is the *confidence level*, or C, of our CI, and we are following convention by choosing $C = 95$. The CI will be a range centered on M, and extending a distance w either side of M, where w (for *width*) is called the *margin of error*. The margin of error is based on the SE, which is a function of SD and n. In fact, $SE = SD/\sqrt{n} = 30/\sqrt{36} = 5$, and w is the SE multiplied by $t_{(n-1),C}$, which is a critical

Geoff Cumming and Sue Finch, School of Psychological Science, La Trobe University, Melbourne, Victoria, Australia.

This research was supported by the Australian Research Council. We thank Kevin Bird, Mark Burgman, Ross Day, Fiona Fidler, Ken Greenwood, Richard Huggins, Chris Pratt, Michael Smithson, Neil Thomason, Bruce Thompson, Eleanor Wertheim, Sabine Wingenfeld, and Rory Wolfe for valuable comments and Rodney Carr for showing what Excel can do.

Correspondence concerning this article should be addressed to Geoff Cumming, School of Psychological Science, La Trobe University, Melbourne, Victoria 3086, Australia, or to Sue Finch, who is now at the Statistical Consulting Centre, University of Melbourne, Melbourne, Victoria 3010, Australia. E-mail: g.cumming@latrobe.edu.au or s.finch@ms.unimelb.edu.au

Vol. 60, No. 2, 170–180 DOI: 10.1037/0003-066X.60.2.170



Geoff Cumming
Photo by David Walsh

value of the t statistic that depends on our chosen value of C.

For $C = 95$, we need the value of t, with $df = n - 1 = 35$, that cuts off the lower 2.5% and upper 2.5% of the t distribution; this critical value is 2.03. Our margin of error is $w = 2.03 \times 5 = 10.15$. The lower limit of our CI is $M - w = 51.85$, and the upper limit is $M + w = 72.15$, and so the 95% CI we seek is (51.85 to 72.15), also written as (51.85, 72.15). This is our interval estimate of the mean verbal ability of Melbourne children.

More generally, in the simple cases we consider the CI estimates μ, the population mean, and the margin of error is given by $w = t_{(n-1),C} \times SE$. The CI is $(M - w, M + w)$ and so the full extent of the CI is twice the margin of error, or $2 \times w$. Different levels of confidence give different sizes of CI, because $t_{(n-1),C}$ depends on C. To be more confident that our interval includes μ, we need a wider interval: A 99% CI is wider than a 95% CI based on the same data, and a 90% CI is narrower.

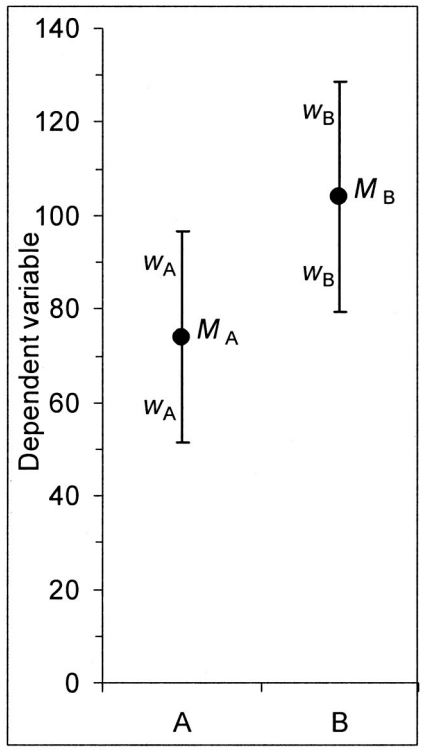
Hays (1973) described a CI as "an estimated range of values with a given high probability of covering the true population value" (p. 375). It is essential, however, to be extremely careful whenever probability is mentioned in connection with a CI. It is correct to state that the probability that $(M - w \leq \mu \leq M + w) = .95$, but this is a probability statement about the lower and upper limits, which vary from sample to sample. It would be incorrect to state that our interval (51.85, 72.15) has probability .95 of including μ, because that suggests that μ varies, whereas μ is fixed, although unknown.

Figure 2 illustrates how M and w, and thus the 95% CI, vary if the experiment is repeated many times. In the long run, we expect 95% (more generally, C%) of the CIs to include μ. As Figure 2 also illustrates, μ is more often captured by the central part of a CI than by either extreme. The occasional CI (two cases in Figure 2; 5% of cases in the long run) will not include μ. Running an experiment is equivalent to choosing just one CI like those shown in Figure 2, and of course we do not know whether our interval does or does not capture μ. Our CI comes from an infinite sequence of potential CIs, 95% of which include μ, and in that sense there is a chance of .95 that our interval includes μ. However, probability statements about individual CIs can so easily be misinterpreted that they are best avoided. Bear in mind Figure 2 and that our calculated CI is just one like those illustrated.

Why Use CIs?

Four major advantages of CIs are that (a) they give point and interval estimates in measurement units that should be readily comprehensible in the research situation; (b) there is a link between CIs and p values and hence familiar NHST; (c) CIs help combine evidence over experiments: They support meta-analysis and meta-analytic thinking focused on estimation; and (d) CIs give information about

Figure 1
Two Fictitious Sample Means With Error Bars



Note. Sample size is $n = 25$ for each, the means are M_A and M_B, and the length of a single bar is w_A or w_B, which is half the extent of the whole interval. If the error bars depict confidence intervals, w_A and w_B are the margins of error. The first ambiguity is that the bars may instead depict standard error bars or may even show standard deviation. The second ambiguity concerns experimental design, as Figure 3 illustrates.

Sue Finch

precision, and this may be more useful than a calculation of statistical power. In our primer on CIs (Cumming & Finch, 2001), we discussed these four aspects further.

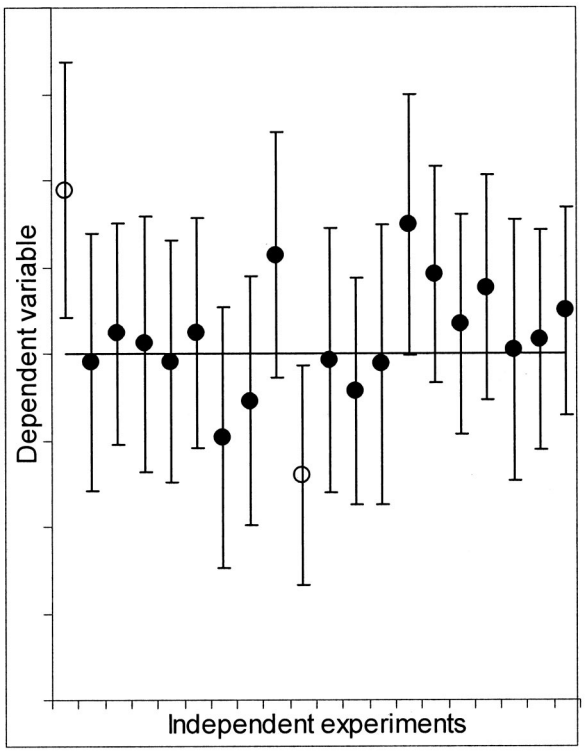
There are complex situations, including multivariate analyses and assessment of the fit of models, where it may be difficult or impossible to find appropriate CIs. However the home territory of CIs is being expanded. Smithson (2000) wrote a statistics textbook for the behavioral sciences that places CIs center stage. Guidance for the calculation of CIs is given for a broad range of situations by Altman, Machin, Bryant, and Gardner (2000), Smithson (2002), and Kline (2004). More specific guidance is given by Steiger and Fouladi (1997) for CIs requiring noncentral distributions; Cumming and Finch (2001) for the standardized effect size measure known as Cohen's *d*; Fidler and Thompson (2001), Bird (2002), and Steiger (2004) for some analysis of variance effect sizes; and Smithson (2001) for some regression effect sizes and parameters.

In this article, we refer often to the link between CIs and *p* values, which we hope may assist researchers develop their CI thinking and practices, but this link is only one of the four points above. We agree with the numerous writers, including Krantz (1999), Oakes (1986), and Rossi (1997), who argue that CIs offer important advantages beyond the link with statistical significance testing. When interpreting CIs, any or all of the four aspects may give insight, and it is important to avoid thinking only of *p* values or whether a null hypothesis should be rejected.

The very common choice of $C = 95$, which corresponds to $\alpha = .05$, may be a legacy of NHST, but researchers can use $C = 90$, or 99, or some other value if there is good reason. In this article, however, we focus on 95% CIs not to reinforce the link with NHST but for consistency with most common CI practice. We aim first to encourage good intuitions about 95% CIs rather than discussing intervals that differ because of various choices of *C*.

In traditional NHST practice, a dichotomous decision is made: A null hypothesis is rejected if $p < .05$, and otherwise is not rejected. The corresponding, but not necessarily appropriate, conclusion is often that an effect is real or not. The *Publication Manual* (APA, 2001) describes dichotomous decision making and also describes the practice of reporting exact *p* values. It concludes by saying "in general it is the exact probability (*p* value) that should be reported" (p. 25). Reporting exact *p* values encourages a move from NHST as dichotomous decision making to the consideration of *p* values as useful input to interpretation. We believe this is a positive move and, further, that CIs can also assist psychologists to move beyond dichotomous NHST. Development of better ways to present and discuss CIs, and the wider use of CIs, could lead psychology researchers to formulate their ideas more in estimation terms and to design their experiments accordingly. This should lead to theories that are more quantitative (Wilkin-

Figure 2
The 95% Confidence Interval (CI) for the Population Mean μ, for 20 Independent Replications of a Study



Note. Each sample has size $n = 36$. The horizontal line is μ. The CIs are based on sample estimates of the population variance and so vary in width from sample to sample. Filled circles are the means whose bars include μ, and open circles are those whose bars do not include μ. In the long run, 95% of the CIs are expected to include μ (18 do here). Note that the CI varies from sample to sample, but μ is fixed and usually unknown.

son & TFSI, 1999), experiments that are more informative, and generally to a stronger empirical discipline.

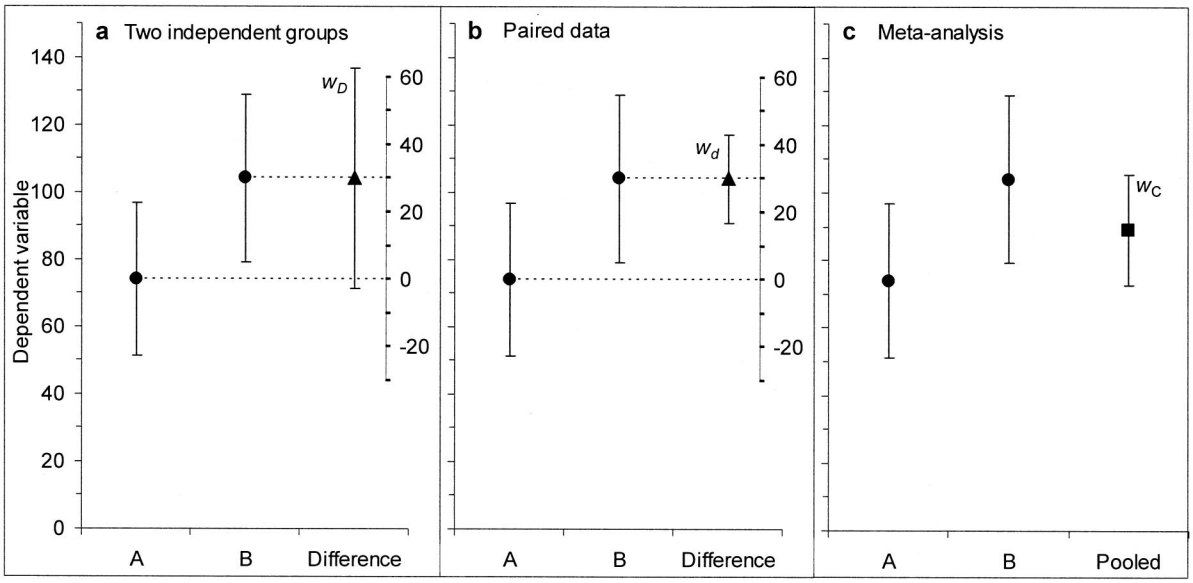
Graphical Ambiguity and Experimental Design

Consider the mean M_A in Figure 1. Error bars extend above and below the mean, and each has length 23, marked as w_A. Error bars like these represent some measure of variability associated with the mean. If they depict a CI, each bar has a length equal to the margin of error. Unfortunately, the same graphic has also traditionally been used to depict the *SE* or the *SD*. *SE* bars extend 1 *SE* above and 1 *SE* below the mean, similarly for *SD* bars. Error bars of the three types obviously give quite different information. Furthermore, even if readers know that the bars show a CI, there is potential ambiguity, because *C* may be 95 or some other value. It is essential that figure captions explain error bars, as the *Publication Manual* requires (APA, 2001, pp. 180, 182). Our focus is on CIs, but in some journals *SE* bars are often shown in figures, so we discuss *SE* bars in a separate section toward the end of the article. From here on, we assume that all bars in Figures 1 and 3 depict 95% CIs.

The second ambiguity concerns experimental design, and Figure 3 illustrates how additions to Figure 1 can distinguish three cases. In Figure 3a, the original means, M_A and M_B, are of two independent groups, each shown with its CI. The difference between the two means ($M_B - M_A$) is shown as a triangle against a difference axis. The 95% CI on this difference has margin of error w_D and is the interval $[(M_B - M_A) - w_D, (M_B - M_A) + w_D]$, which is our interval estimate of the difference between the population means ($\mu_B - \mu_A$). Now w_D is about $\sqrt{2}$ (or 1.4) times either of the original margins of error, w_A and w_B, assuming these are similar. It makes sense that this CI is larger than either of the original intervals because sampling error in the difference is a compounding of sampling error from each of the two independent means. If, as usual for two independent groups, we are interested in the difference between means, then Figure 3a provides on the right the CI directly relevant for our desired inference.

Another possibility is that Figure 1 summarizes paired data. The A and B means may, for example, be of pre- and posttest scores for a single group of participants. There is a single repeated-measure independent variable (IV) and the two scores are almost certainly correlated. Our interest is probably in the differences, and Figure 3b shows as a triangle the mean M_d of the posttest minus pretest differences on a difference axis. The CI on this mean has margin of error w_d and is the interval $(M_d - w_d, M_d + w_d)$, which is our interval estimate of the population mean of differences. Note that M_d equals $M_B - M_A$ and that w_d, like the *t* value for the paired *t* test, is based on the *SE* of the differences. We discuss the case of paired data in a separate section later.

Figure 3
Additions to Figure 1 That Illustrate Three Possibilities



Note. (We assume that Figures 1 and 3 show 95% confidence intervals [CIs].) (a) The A and B means are of independent groups, and their difference ($M_B - M_A$) is plotted as a triangle on a difference axis. The CI on this difference has margin of error w_D that is virtually always greater than w_A and w_B, the margins of error on the original means. (b) A repeated-measure design, with correlated A and B scores, for example, pre- and posttest scores for a single group of participants. The triangle is M_d, the mean of the posttest minus pretest differences, which also equals $M_B - M_A$. The small margin of error w_d of its CI reflects a high correlation, .85, between pre- and posttest scores. (c) The A and B means are for separate experiments examining the same question. The square is the mean resulting from pooling the two experiments—a simple form of meta-analysis—and its CI has a smaller margin of error w_C.

A third possibility is that A and B are separate studies that investigate the same issue. Figure 3c illustrates pooling of the two studies, which is the bare-bones meta-analysis of Hunter and Schmidt (2004). The pooled mean, marked by the square, is our best estimate of μ based on combining the original studies, and its CI is typically smaller than that of either original study. Such a summary plot of more than one experiment is known in medicine as a forest plot (Altman et al., 2000, pp. 134–135). The value of a CI display of a meta-analysis is discussed by Cumming and Finch (2001), Light, Singer, and Willett (1994), Schmidt (1996), and Thompson (2002).

Figures 1 and 3 emphasize that experimental design is crucial for the interpretation of figures and that displays of means and CIs can easily leave experimental design ambiguous. Clear figure captions are necessary to resolve such ambiguity. Further, Gardner and Altman (1986) advised that "the major contrasts of a study should be shown directly, rather than only vaguely in terms of the separate means" (p. 748); they would therefore approve of Figure 3 but not Figure 1. In the three cases discussed above, our interest is in a difference or combination of means, rather than the two separate means shown in Figure 1. In each case, Figure 3 displays a single mean and CI that are directly relevant to our interest and sufficient to support inferential interpretation; this is perhaps the most valuable representation we could be given, regardless of whether the original two means are also shown.

The argument extends beyond simple differences. In more complex designs, Figure 1 would show more than two cell means. The mean that is added for Figure 3 could be of a main effect, simple main effect, interaction, or indeed any contrast. The CI on that effect or contrast would be shown, although it may be challenging to select an appropriate error term to use to calculate the CI (Loftus & Masson, 1994; Masson & Loftus, 2003).

The present article is about interpreting, rather than designing, figures, but the discussion above highlights the value of presenting the mean and CI that are directly relevant for the effect or comparison of primary research interest. Good figure design is vital for appropriate data interpretation, and we expect to see increasing use of figures that present means and CIs for contrasts of inferential interest. It remains necessary, however, to consider the interpretation of figures showing separate means with error bars, as in Figure 1, because they appear so frequently in journals. We now turn to our main topic, discussion of how simple pictures of means with CIs might be read.

Rules of Eye for Reading Data Pictures With CIs

How should a figure with CIs—whether printed in a journal or projected on a screen during a conference presentation—be interpreted? We propose five rules of eye that apply to figures that show means and CIs; the Appendix summarizes these rules and also two rules for *SE* bars we describe later. By analogy with rules of thumb, our rules are intended to be useful heuristics or pragmatic guidelines. They are not intended to be numerically exact or to replace statistical calculations: If exact p values are desired, they should also be presented. Our focus is on the broad inferential understanding that rules of eye may prompt. Some of the rules have general application to inference beyond the interpretation of figures, but they are essential for graphical interpretation—the focus of this article—so we include them as rules of eye.

Rule of Eye 1: Identify what the means and error bars represent.

At the beginning of this article, we asked "On first seeing Figure 1, what questions should spring to mind?" We can now specify four questions, whose answers should together satisfy Rule 1: (a) What is the dependent variable? Is it expressed in original units, or is it standardized in some way, for example as Cohen's *d*? (b) Does the figure show 95% CIs, *SE* or *SD* bars, or possibly CIs with a different confidence level *C*? (c) What is the experimental design? (d) What effect or comparison is our major interest, and how do the displayed means and CIs relate to this? Where should we focus our inferential attention?

Rule of Eye 2: Make a substantive interpretation of the means.

The first interpretive focus should be the means, or combinations or pattern of means, and these should be assessed against any theoretical predictions. Use knowledgeable judgment in the research situation and consider the extent to which an effect is (a) important or interesting and (b) large. Distinguish practical or clinical significance from statistical significance (Kendall, 1999; Kirk, 1996).

Interpretation of a CI

Rule of Eye 3: Make a substantive interpretation of the CI.

We suggest four approaches to the interpretation of any CI. We refer to a 95% CI, but there are generalizations for a *C*% CI.

Our CI is just one from an infinite sequence. As we discussed earlier, and referring to Figure 2, if the experiment were repeated many times and a CI calculated for each, in the long run 95% of the intervals would include μ. This is the fundamental and correct way to think about a CI. Equivalently, a researcher who routinely reports 95% CIs can expect over a lifetime that about 95% of those intervals will capture the parameters estimated (Cohen, 1995).

Focus on the interval and values in the interval. The general idea is that values within the CI are a good bet for μ and those outside it are not, but scholars differ on what words give an acceptable way to express this, without implying inappropriate statements about probability. We give several alternatives, each of which may be queried by some authorities. There are not yet widely agreed ways to say all that needs to be said when interpreting CIs.

- This is our favorite: Our CI is a range of plausible values for μ. Values outside the CI are relatively implausible.

- We can be 95% confident that our CI includes μ.
- Our data are compatible with any value of μ within the CI but relatively incompatible with any value outside it.
- The lower limit is a likely lower bound estimate of the parameter; the upper limit a likely upper bound.
- Consider substantive interpretation of values anywhere in the interval. For example, consider interpretations of the lower and upper limits and compare these with interpretation of the mean (Rule 2).

The CI, NHST, and p values. Any value outside the CI, when considered as a null hypothesis, gives two-tailed $p < .05$. Any value inside the CI, when considered as a null hypothesis, gives $p > .05$. Considering our example CI (51.85, 72.15), the null hypothesis $\mu = 50$ would give $p < .05$ because 50 is outside the interval, but the null hypothesis $\mu = 60$ would give $p > .05$.

If the lower or upper CI limit is considered as a null hypothesis, the p value is exactly .05 or, more generally, 1 – C/100. Recall that changing C changes the size of the CI. Suppose we increase C to widen our example CI until the lower limit is exactly 50; this requires $C = 97.8$. The corresponding $p = (1 – 97.8/100) = .022$, and this is the p value calculated from our data for the null hypothesis $\mu = 50$. If we can adjust C, CIs are not limited to indicating whether $p < .05$ but can give exact p values, which we recommended earlier.[1]

An index of precision: w. We can be 95% confident that our point estimate is no more than w from the true value of μ, so w is the largest error of estimation we are likely to make—although larger errors are possible—and w is an index of the precision of the study. Make a substantive interpretation of w.

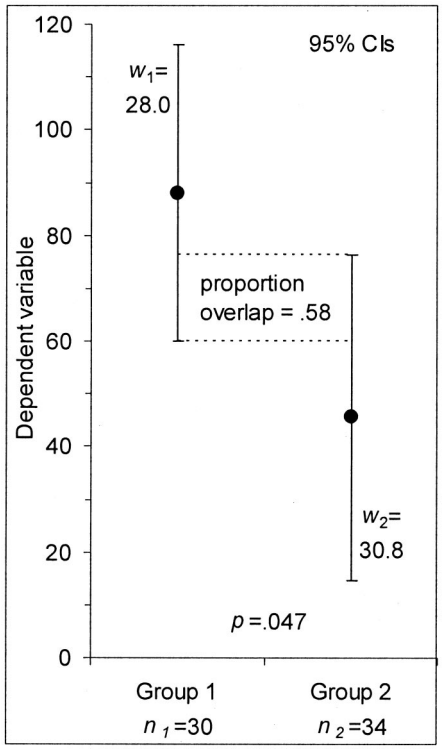
The margin of error w may come to be recognized as more useful than a statistical power estimate for planning an experiment (What n is needed to achieve a desired w?) and also for reporting results (Cumming & Finch, 2001). Smithson (2002, chap. 7) also discussed the relation between CIs and power.

Great caution is needed whenever NHST tempts the acceptance of a null hypothesis. CIs do not reduce the need for caution, but w is likely to give strong guidance. A large w may suggest an experiment is of little value. On the other hand, if the CI is narrow and close to a null hypothesized value μ (which may be contained in the interval), and if we judge every value in the CI to be for practical purposes equal to μ, then the CI justifies our regarding μ for practical purposes as the true value (Tryon, 2001).

Two Independent Groups

Our first three rules are the most general. The fourth rule applies to the two-independent-groups case. First, we describe an example, which we suspect may be surprising to some. Referring to Figure 4, suppose Groups 1 and 2 are independent and that 95% CIs are displayed. Knowing the means, sample sizes, and margins of error (and of course that the error bars depict 95% CIs) is sufficient to determine the p value for the t test comparison between the means.

Figure 4
Means With 95% Confidence Intervals (CIs) for a Two-Independent-Groups Example



Note. The group sizes are $n_1 = 30$ and $n_2 = 34$, and margins of error are $w_1 = 28.0$ and $w_2 = 30.8$. The p value for the difference between the means is .047, so that difference is close to the conventional statistical significance borderline ($\alpha = .05$). The dotted horizontals help estimation of proportion overlap. The CIs overlap by a little over half the average margin of error; proportion overlap is actually .58 (see text). Compare this with the Rule 4 criterion of .50.

Our analysis is based on the method of Welch (1938) and Satterthwaite (1946), which pools error variances for the denominator of an independent-groups t statistic without requiring the assumption of equal variance in the two underlying populations.

We define *proportion overlap* as the vertical distance between the dotted horizontals in Figure 4, expressed as a proportion of the average margin of error. The groups have sizes $n_1 = 30$ and $n_2 = 34$. The margins of error are $w_1 = 28.0$ and $w_2 = 30.8$, so the average margin of error is 29.4. The overlap is 77.0 – 60.0 = 17.0, in the units of the dependent variable, so proportion overlap is 17.0/29.4 = .58. Inspection of Figure 4 confirms that proportion overlap of the CIs is a little more than .50: The intervals overlap by

[1] A component of the Exploratory Software for Confidence Intervals (ESCI; "ess-key"), which runs under Microsoft Excel, can be used to generate figures similar to those in this article and to adjust C. This component of ESCI may be downloaded, for personal use without cost, from www.latrobe.edu.au/psy/esci.

a little more than half the average of w_1 and w_2. The p value for the difference between the two independent means is .047, and so Figure 4 illustrates the configuration of means and error bars when the difference between the means is near the border of statistical significance, using the most common criterion of $\alpha = .05$. This configuration gives the basis for our fourth rule. Figure 5 illustrates this rule when group sizes are equal and not small and $w_1 = w_2$.

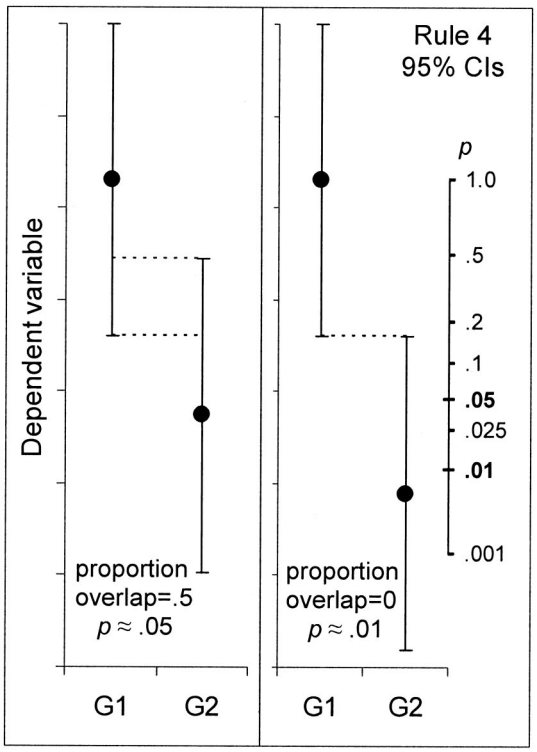
Rule of Eye 4: For a comparison of two independent means, $p \leq .05$ when the overlap of the 95% CIs is no more than about half the average margin of error, that is, when proportion overlap is about .50 or less (see Figure 4; Figure 5, left panel). In addition, $p \leq .01$ when the two CIs do not overlap, that is, when proportion overlap is about 0 or there is a positive gap (see Figure 5, right panel). These relationships are sufficiently accurate when both sample sizes are at least 10, and the margins of error do not differ by more than a factor of 2.

Note that, other things being equal, a greater difference between the means implies a smaller overlap or greater gap and a smaller p value for the t test comparison of the means. We explored the relation between proportion overlap and p value for a wide variety of sample sizes and margins of error. We concluded that overlap, expressed as a proportion of average margin of error, is the best way to summarize the complex relationships underlying inference from CIs for two independent means (Cumming, 2004).

In the majority of cases that have proportion overlap of .50 and meet the stated conditions—sample sizes of at least 10 and w_1 and w_2 not differing by more than a factor of 2—the p value is between .04 and .05 and, in virtually every case that meets the conditions, p is between .03 and .05 (Cumming, 2004). Therefore, the rule is generally a little conservative in the sense that the true p value is usually a little less than the upper bound for p stated in the rule. It is striking that the rule holds even when group sizes are quite different, providing that each is at least 10 and the margins of error do not differ by more than a factor of 2.

When group sizes are equal and not small, margins of error are equal, and proportion overlap = .50, then p = .038 (rather less than .05), and when proportion overlap is zero, p = .006 (considerably less than .01). Schenker and Gentleman (2001) reported that many medical researchers take zero overlap of CIs as equivalent to statistical significance at the .05 level. This equivalence is, however, a misconception (Saville, 2003; Wolfe & Hanley, 2002): When 95% CIs just touch end to end, the p value is about .006, very much less than .05. Figure 5 includes a p value scale, which we suggest may be useful for pedagogy rather than for the routine reporting of results. It gives a general insight into the way p varies with separation between the means and with proportion overlap.

Figure 5
Rule of Eye 4 for Two Independent Groups, Both of Size 50 and With Equal Margins of Error



Note. Proportion overlap of the 95% confidence intervals (CIs) is .50 in the left panel and .00 in the right, and the corresponding approximate p values are .05 and .01, respectively. The vertical scale on the right shows the p value as a function of the value of the Group 2 (G2) mean, all other aspects remaining the same: Hold the Group 1 (G1) mean fixed, move the G2 mean to any position, then the point on the scale that is aligned horizontally with the G2 mean gives the two-tailed p value for that difference between means. This scale shows that the p values stated are a little conservative—in the left panel the G2 mean is a little below .05 on the scale, and in the right panel a little below .01.

Paired Data

In the two-independent-groups case, if we have Figure 3a, we can interpret the single CI on the difference between means; if not, we can apply Rule 4 to the CIs on the two independent means A and B. In stark contrast, however, with paired data we do not have the second option: The CIs on M_A and M_B, the two score means, are irrelevant for inference about the mean difference. For inference with paired data, we need the CI on the mean of the differences. This CI is shown on the right in Figure 3b, and its margin of error w_d is sensitive to the correlation between the two scores. The relation is $w_d^2 = w_A^2 + w_B^2 - 2rw_Aw_B$, where r is the Pearson product–moment correlation of the two scores, calculated for the data. When, as is common, the correlation is positive, then the larger the correlation, the smaller is w_d. With zero correlation, Figure 3b becomes similar to Figure 3a. In practice, the correlation is rarely negative, but it could be, in which case w_d can be as much as twice w_A or w_B. In summary, w_d may in principle have any value from practically zero up to about twice w_A or w_B, depending on the correlation between the two scores. Therefore, knowing w_A and w_B gives not even the roughest of guides for inference. In the paired data case, Figure 1 cannot support inference, and it is not possible to formulate a rule of eye based on overlap of separate CIs.

This conclusion applies to any repeated-measure situation: Error bars on separate means are simply irrelevant for any inference on the repeated-measure IV. They may be highly misleading for a reader who is not alert to the presence of repeated-measure IVs and what that implies for inference.

Rule of Eye 5: For paired data, focus on and interpret the mean of the differences and the CI on this mean (see Figure 3b). Noting whether the CI on the mean of the differences captures 0 is a test of the null hypothesis of no difference between the means. The CIs for the two separate scores (e.g., pretest and posttest) are irrelevant for inferences about the mean of the differences. In general, beware of separate error bars for a repeated-measure IV: They are irrelevant for the inferences likely to be of interest.

Interpreting p Values

Two further aspects of *p* values deserve consideration in relation to our rules. Earlier we noted that reporting exact *p* values encourages a move from dichotomous NHST to the consideration of *p* values as useful input to interpretation. We suggest a similar attitude should be taken when interpreting CIs. Just as *p* values of .04 and .07 are unlikely to justify dramatically different conclusions, so we should not make a big issue of whether a CI just includes or excludes a value of interest. If independent means are compared, Rule 4 can be used to assess *p* values against .05 and .01 but without undue concern for precise accuracy, or for justifying borderline dichotomous decisions. (Recall that we advocate inference by eye for the appreciation, broadly, of the inferential import of figures. It is not intended to replace *p* value calculations if these are desired.)

Second, by discussing single inferences, we have taken a decisionwise approach and have not mentioned any adjustment of *C*, or *p* values, to allow for multiple inferences, as required by an experimentwise approach. For inference by eye, as for any inference, the author and reader should decide what approach is best for a given research situation and interpretive aims. Our simple decisionwise approach is probably reasonable if there are few inferences and they were identified in advance. When many inferences are examined, or if selection is post hoc from a large number of possible inferences, some small *p* values can easily be produced by sampling variability. Rather than proposing an experimentwise version of Rule 4, or the use of CIs with unfamiliar *C* values (Tryon, 2001), we suggest it will suffice if the issue is borne in mind. If more than a handful of inferences are being considered, a researcher may choose to adopt a more conservative approach to *p* values and to interpreting CIs.

SE Bars

Cleveland (1994) regarded the practice of showing *SE* bars in figures as "a naïve translation of the convention for numerical reporting of sample-to-sample variation" (p. 218). Referring to *SE* bars he wrote the following:

> The difficulty . . . is that we are visually locked into what is shown by the error bars; it is hard to multiply the bars visually by some constant to get a desired visual confidence interval on the graph. Another difficulty, of course, is that confidence intervals are not always based on standard errors. (p. 219)

The constant Cleveland refers to is $t_{(n-1),C}$, a critical *t* value, which varies with *n* and *C*. When *n* is not small, *SE* bars correspond approximately to a 68% CI, and the critical *t* value is about 2 so the *SE* bars need to be doubled in length to give a 95% CI. When *n* is very small—less than about 10—the critical *t* value increases above 2, and *C* for the equivalent CI drops below 68. One example figure in the *Publication Manual* shows *SE* bars for groups of size $n = 2$ and $n = 4$ (APA, 2001, p. 182), which correspond to 50% and 61% CIs, respectively. In such cases, *SE* bars are likely to be a misleading basis for inference.

Cleveland's (1994) emphasis on inference is justified, and he makes a good case for preferring CIs over *SE* bars. It is CIs that have been the focus of statistical reform, and CIs are also preferred in medicine (International Committee of Medical Journal Editors, 1997). However, despite the *Publication Manual*'s advocacy of CIs (APA, 2001, p. 22), the only error bars shown in its example figures (pp. 180, 182), and in Nicol and Pexman (2003), are *SE* bars. In addition, figures published in some psychology and many behavioral neuroscience journals often include *SE* bars, and so we include below two rules of eye for *SE* bars.

Cohen (1994) suggested that one reason psychologists seldom report CIs may be that their CIs are often embarrassingly large. If researchers prefer to publish *SE* bars merely because they are shorter, they are capitalizing on their readers' presumed lack of understanding of *SE* bars, 95% CIs, and the relation between the two.

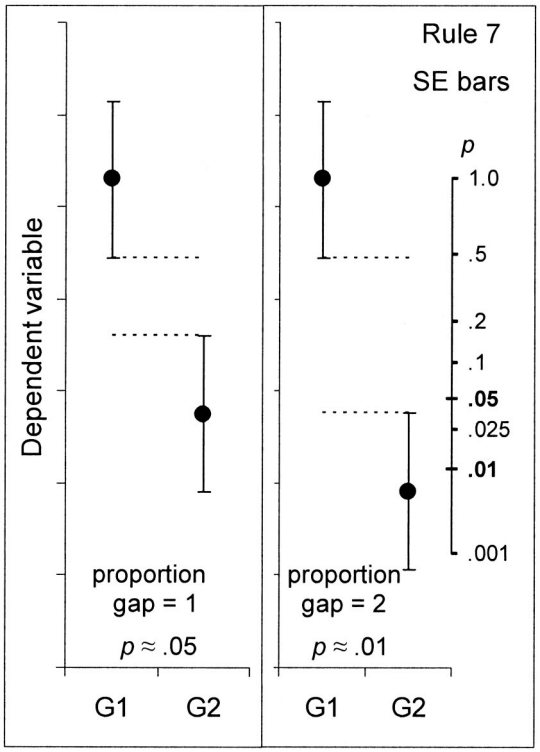
A problem arising from the first ambiguity of Figure 1—bars may be *SE* or CIs, or even *SD*—is that it is easy to find journal articles that show error bars in figures but do not state what they represent (Finch et al., 2004; Vaux, 2004). It is extremely unfortunate that the ambiguous graphic of Figure 1 has several different meanings.

Rule of Eye 6: For n at least 10, SE bars can be doubled in length to get, approximately, the 95% CI; and the SE bars themselves give approximately a 68% CI, so in about two thirds of cases SE bars capture μ. Thinking of either of these CIs allows Rule 3 to be applied. For n less than 10, SE bars give a CI with C distinctly less than 68.

For two independent groups, halving the intervals in Rule 4 and Figure 5 gives the corresponding Rule 7 and Figure 6 for *SE* bars. Note that for Figure 6 the sample sizes, positions of the means, and amount of variability in each group are all the same as in Figure 5. Therefore, the *p* values for the differences between the means have not changed from Figure 5 to Figure 6, and the same *p* value scale can be displayed in both figures.

Rule of Eye 7: For a comparison of two independent means, $p \leq .05$ when the gap between the SE bars is at least about the size of the average SE, that is, when the proportion gap is about 1 or greater (see Figure 6, left panel). In addition, $p \leq .01$ when the proportion gap is about 2 or more (see Figure 6, right panel). These relationships are sufficiently accurate when both sample

Figure 6
Rule of Eye 7 for Two Independent Groups, Both of Size 50 and With Equal SEs



Note. The proportion gap between the standard error (*SE*) bars is 1 in the left panel and 2 in the right, and the corresponding approximate *p* values are .05 and .01, respectively. As in Figure 5, the vertical scale on the right shows the *p* value as a function of the value of the Group 2 (G2) mean, all other aspects remaining the same. This scale, which applies to both panels, shows that the *p* values stated are a little conservative. G1 = Group 1.

sizes are at least 10, and the SEs of the two groups do not differ by more than a factor of 2.

Outstanding Issues

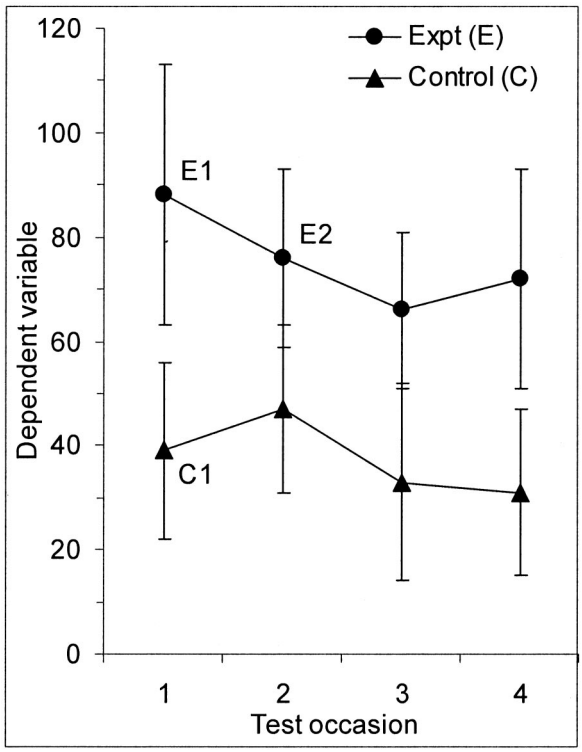
Complex Experimental Designs

Figure 7 shows cell means with error bars for a two-way design, with one repeated measure. This is a common design, and both examples in the *Publication Manual* of figures with bars are of this design (APA, 2001, pp. 180, 182). The trouble is that the bars shown may legitimately be used to assess between-subjects comparisons but may not be used to assess any within-subjects effect. Rules 4 (for 95% CIs) or 7 (for *SE* bars) may be used to assess between-subjects comparisons such as Experimental 1 with Control 1, but the error bars needed to assess a within-subjects comparison such as Experimental 1 with Experimental 2 are not provided in the figure. Error bars on cell means are irrelevant for any within-subjects effect. Belia et al. (2004) found that a large majority of researchers tend to overlook statements identifying a repeated-measure IV and to interpret error bars erroneously, as if the means were independent. We suspect, therefore, that few researchers have the crucial distinction between within-subjects and between-subjects effects clearly in mind when examining figures such as Figure 7.

The problem of how the CIs needed for effects involving within-subjects IVs can be represented in figures has been discussed by Estes (1997), Loftus (2002), Loftus and Masson (1994), and Masson and Loftus (2003). Various generalizations of Figures 3a and 3b could be considered, and any additional means displayed may be of main effects, interactions, or any contrasts of interest. Different contrasts are likely to have CIs of different widths, and it may be enlightening to see these displayed in a single figure. However, if more than a few effects are of interest, the graphical challenge is very great, and no convincing and proven graphical designs have yet emerged. Investigation is needed of the extent to which CIs can be effectively used with complex experimental designs.

The problem is illustrated by a figure in the *Publication Manual* (APA, 2001, p. 181), which shows the means

Figure 7
Means With Error Bars for a Two-Way Design With One Repeated Measure



Note. Fictitious data. Experimental (Expt; E)–control (C) is a between-subjects independent variable, and test occasion is the within-subjects independent variable. The error bars shown on the cell means, whether confidence intervals or standard error bars, may be used to assess between-subjects comparisons such as E1 with C1 but are irrelevant for any within-subjects comparison, such as E1 with E2.

for a two-way design with one within-subjects IV. A line segment is shown, with the notation "If a difference is this big, it is significant at the .05 level." The problem is that different differences need to be specified for each of the two main effects, for simple main effects on either IV, and for any other contrast or interaction of interest. Showing a single interval with such a general notation cannot be correct.

Statistical Cognition

Statistical cognition is the study of how people think about statistical concepts and representations. There is scant cognitive evidence that CIs, and other techniques recommended by reformers, can in fact give the improved understanding and communication that is claimed for them. With its research skills and knowledge of perception and cognition, as well as statistics, psychology is uniquely placed to help develop interval estimation practice that is evidence based. What representations and guidelines will prompt easy intuitive understanding and minimize misconception? Such cognitive research is needed to guide reform of statistical practices. Belia et al. (2004) and Cumming et al. (2004) are examples of one type of statistical cognition study.

Inference by Eye: Broadening the Scope

Note the limitations of our discussion. We have focused on two-sided CIs on the mean, with an underlying normally distributed population. Rules 4 and 7 are for independent groups, with sample sizes of at least 10 and error bars that do not differ in length by more than a factor of 2. We have considered single inferences, with no adjustment to account for inflated error rates with multiple inferences.

There are important generalizations to be explored, including one-sided intervals (corresponding to one-tailed hypothesis tests); parameters with restricted ranges that generally have nonsymmetric CIs, including proportions and correlations; a wide range of other parameters beyond the mean; ordinal measurement (McGill, Tukey, & Larsen, 1978); relaxation of the assumption of normality and use of robust methods (Wilcox, 1998, 2003); and CIs generated in different ways, such as via resampling (Edgington, 1995). Our Rules 1, 2, and 3 are sufficiently fundamental that, with minor changes of wording (e.g., *point estimate* for *mean*), they will largely apply across most generalizations. Similarly, the caution of Rule 5 about repeated measures applies generally. Rule 6 is not so general because the relation between *SE* and CI bars may be different in different situations, as it is for very small *n*. The breadth of applicability of Rules 4 and 7 requires investigation. We speculate that they hold approximately when CIs are close to symmetric and the underlying populations do not depart drastically from normal. Also, in such situations it may make little difference whether the CIs are derived by conventional calculation or, for example, via resampling.

Conclusions

Wider use of interval estimation has the potential to improve research communication and, more fundamentally, to encourage more sophisticated theorizing and testing of theories in our discipline. Achieving routine use of intervals will, however, be a very substantial change, requiring changed attitudes and practices on the part of researchers, editors and their consultants, and those involved in statistics education within psychology. Difficulties include the unfamiliarity of CIs to many psychologists, widespread misconceptions, ambiguities in common graphical designs, and the lack of guidelines for the reporting and interpretation of CIs.

The more we work with CIs, the more we realize how little their representation and understanding seem to have been investigated. We regard the suggestions in this article as an early step and look forward to further development of inference by eye. We are optimistic, however, that discovering how to exploit statistical estimation could be very fruitful for our discipline.

REFERENCES

Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (2000). *Statistics with confidence: Confidence intervals and statistical guidelines* (2nd ed.). London: British Medical Journal Books.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2004). *Researchers misunderstand confidence intervals and standard error bars.* Manuscript submitted for publication.

Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement, 62,* 197–226.

Cleveland, W. S. (1994). *The elements of graphing data.* Summit, NJ: Hobart Press.

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

Cohen, J. (1995). The earth is round ($p < .05$): Rejoinder. *American Psychologist, 50,* 1103.

Cumming, G. (2004). *Inference by eye: Confidence intervals, p values, and overlap.* Manuscript in progress.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61,* 530–572.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3,* 299–311.

Edgington, E. S. (1995). *Randomization tests* (3rd ed.). New York: Dekker.

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

Fidler, F. (2002). The fifth edition of the APA *Publication Manual*: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement, 62,* 749–770.

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement, 61,* 575–604.

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., et al. (2004). Reform of statistical inference in psychology: The case of *Memory & Cognition. Behavior Research Methods, Instruments, & Computers, 36,* 312–324.

Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory and Psychology, 12,* 825–853.

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than *p* values: Estimation rather than hypothesis testing. *British Medical Journal, 292,* 746–750.

Hays, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart & Winston.
Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.
International Committee of Medical Journal Editors. (1997). Uniform requirements for manuscripts submitted to biomedical journals. *Annals of Internal Medicine, 126,* 36–47.
Kendall, P. C. (1999). Clinical significance. *Journal of Consulting and Clinical Psychology, 67,* 283–284.
Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.
Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.
Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 44,* 1372–1381.
Light, R., Singer, J., & Willett, J. (1994). The visual presentation and interpretation of meta-analysis. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 439–453). New York: Sage.
Loftus, G. R. (2002). Analysis, interpretation, and visual presentation of experimental data. In J. Wixted & H. Pashler (Eds.), *Stevens' handbook of experimental psychology*: *Vol. 4. Methodology in experimental psychology* (3rd ed., pp. 339–390). New York: Wiley.
Loftus, G. R., & Masson, M. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1,* 476–490.
Masson, M., & Loftus, G. R. (2003). Using confidence intervals for graphically based interpretation. *Canadian Journal of Experimental Psychology, 57,* 203–220.
McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *The American Statistician, 32,* 12–16.
Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.
Nicol, A. A. M., & Pexman, P. M. (2003). *Displaying your findings: A practical guide for creating figures, posters, and presentations.* Washington, DC: American Psychological Association.
Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* Chichester, England: Wiley.
Rossi, J. S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Muliak, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 175–197). Mahwah, NJ: Erlbaum.
Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2,* 110–114.
Saville, D. J. (2003). Basic statistics and the inconsistency of multiple comparison procedures. *Canadian Journal of Experimental Psychology, 57,* 167–175.
Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician, 55,* 182–186.
Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.
Smithson, M. (2000). *Statistics with confidence.* London: Sage.
Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61,* 605–632.
Smithson, M. (2002). *Confidence intervals.* Thousand Oaks, CA: Sage.
Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9,* 164–182.
Steiger, J. H., & Fouladi, T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. Harlow, S. Muliak, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Mahwah, NJ: Erlbaum.
Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31,* 25–32.
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.
Vaux, D. L. (2004, April 22). Error message. *Nature, 428,* 799.
Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29,* 350–362.
Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300–314.
Wilcox, R. R. (2003). *Applying contemporary statistical techniques.* New York: Academic Press.
Wilkinson, L., & Task Force on Statistical Inference, American Psychological Association. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.
Wolfe, R., & Hanley, J. (2002). If we're so different why do we keep overlapping? When 1 plus 1 doesn't make 2. *Canadian Medical Association Journal, 166,* 65–66.

Appendix
Abbreviated Statements of Rules of Eye for Simple Figures Showing Means With Error Bars

1. Identify what the means and error bars represent. Do bars show confidence intervals (CIs) or standard errors (*SE*s)? What is the experimental design?

2. Make a substantive interpretation of the means.

3. Make a substantive interpretation of the CIs or other error bars.

4. For a comparison of two independent means, $p \leq .05$ when proportion overlap of the 95% CIs is about .50 or less. (Proportion overlap is expressed as a proportion of the average margin of error for the two groups.) In addition, $p \leq .01$ when the proportion overlap is about 0 or there is a positive gap (see Figure 5). (Rule 4 applies when both sample sizes are at least 10 and the two margins of error do not differ by more than a factor of 2.)

5. For paired data, interpret the mean of the differences and error bars for this mean. In general, beware of bars on separate means for a repeated-measure independent variable: They are irrelevant for inferences about differences.

6. *SE* bars are about half the size of 95% CI bars and give approximately a 68% CI, when *n* is at least 10.

7. For a comparison of two independent means, $p \leq .05$ when the proportion gap between *SE* bars is at least about 1. (Proportion gap is expressed as a proportion of the average *SE* for the two groups.) In addition, $p \leq .01$ when the proportion gap is at least about 2 (see Figure 6). (Rule 7 applies when both sample sizes are at least 10 and the two *SE*s do not differ by more than a factor of 2.)