Psychological Methods
2005, Vol. 10, No. 4, 389–396

Copyright 2005 by the American Psychological Association
1082-989X/05/$12.00 DOI: 10.1037/1082-989X.10.4.389

Researchers Misunderstand Confidence Intervals and Standard Error Bars

Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming
La Trobe University

Little is known about researchers' understanding of confidence intervals (CIs) and standard error (*SE*) bars. Authors of journal articles in psychology, behavioral neuroscience, and medicine were invited to visit a Web site where they adjusted a figure until they judged 2 means, with error bars, to be just statistically significantly different ($p < .05$). Results from 473 respondents suggest that many leading researchers have severe misconceptions about how error bars relate to statistical significance, do not adequately distinguish CIs and *SE* bars, and do not appreciate the importance of whether the 2 means are independent or come from a repeated measures design. Better guidelines for researchers and less ambiguous graphical conventions are needed before the advantages of CIs for research communication can be realized.

Keywords: confidence intervals, statistical cognition, standard error, error bars, statistical reform

Null hypothesis significance testing (NHST) and the use of *p* values is, across many disciplines, the most common statistical technique, but it is widely misunderstood (Finch, Thomason, & Cumming, 2002; Nickerson, 2000; Oakes, 1986) and can prompt poor research decision making (Schmidt, 1992, 1996). Statistical reformers seek to reduce reliance on NHST and *p* values (Cohen, 1990, 1994; Thompson, 1996), especially by use of confidence intervals (CIs; Harlow, Mulaik, & Steiger, 1997), which have the advantage of providing information about precision as well as statistical significance (Cumming & Finch, 2001, 2005). Following advocacy by reformers, CIs came into routine use in medical research during the 1980s (Altman, Machin, Bryant, & Gardner, 2000; Fidler, Thomason, Cumming, Finch, & Leeman, 2004). CIs have been seldom used in psychology (Finch, Cumming, & Thomason, 2001; Thompson, 1999), but the latest *Publication Manual of the American Psychological Association* (APA) states "because confidence intervals combine information on location and precision and can often be directly used to infer significance levels, they are, in general, the best reporting strategy. The use of confidence intervals is therefore strongly recommended" (APA, 2001, p. 22). More than 1,000 journals across many disciplines use the *Publication Manual* (APA, 2001, p. xxi), so its advocacy of CIs has the potential to influence statistical practice very widely.

Little, however, is known about how well researchers understand CIs and standard error (*SE*) bars. Statistical reformers cite evidence of cognitive misconception to support criticism of NHST but can cite little evidence about whether CIs are understood well and can be interpreted appropriately by researchers. This is unfortunate, because without such evidence there cannot be evidence-based reform of statistical practices in psychology. Of particular interest is the extent of researchers' knowledge of how error bars can justifiably be used for inference.

Our aim was to study some aspects of the understanding of graphically presented CIs and *SE* bars by authors of articles published in international journals. We investigated the interpretation of error bars in relation to statistical significance. This may not be the best way to think of error bars (Cumming & Finch, 2001, 2005), but is worthy of study because of the current dominance of NHST and *p* values and because greater interpretive use of error bars is unlikely unless the relationship with *p* values is understood. We also investigated researchers' appreciation, when comparing two means, of the importance of experimental design—in particular whether the independent variable is a between-subjects variable or a repeated measure.

We elected to study graphically presented intervals be-

Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming, School of Psychological Science, La Trobe University, Melbourne, Victoria, Australia.

This research was supported by the Australian Research Council. We thank the researchers who participated, Bradley Dean for Web site construction, and Mark Burgman, Graeme Galloway, and Richard Platt for comments.

Correspondence concerning this article should be addressed to Geoff Cumming, School of Psychological Science, La Trobe University, Melbourne, Victoria 3086, Australia. E-mail: g.cumming@latrobe.edu.au

cause we believe that pictorial representation is often valuable and can "convey at a quick glance an overall pattern of results" (APA, 2001, p. 176) and because we agree with the advice of the APA Task Force on Statistical Inference: "In all figures, include graphical representations of interval estimates whenever possible" (Wilkinson et al., 1999, p. 601). By "error bars" we refer to the ambiguous graphic, two of which are shown in Figure 1, that marks an interval and may represent a CI, *SE* bars, or even standard deviation (*SD*) bars. All error bars we use are centered on means (*M*s), and n is the number of data values contributing to a mean. CIs are calculated as $M \pm t_C \times SE$, where $SE = SD/\sqrt{n}$, and t_C is the critical value of t for $(n - 1)$ degrees of freedom for the chosen level of confidence, C. For us, this is 95%, implying that t_C is close to 2. In all cases, *SE* bars are $M \pm SE$.

To make an inferential assessment of a difference between two means, it may be best for researchers to consider a single interval on the difference itself (Cumming & Finch, 2005). For two reasons, however, we chose to study a comparison of intervals on the two separate means. First, it is common in journals to see figures showing separate cell means, sometimes with error bars, and with such figures, assessing any difference requires consideration of intervals shown on the separate means. The *Publication Manual* includes two examples of figures of this type (APA, 2001, pp. 180, 182). Second, Schenker and Gentleman (2001) reported that in medicine and health science, a rule of thumb is sometimes used for interpreting CIs on two separate means. The rule maintains that nonoverlap of two 95% CIs on independent means implies a significant difference at the .05 level between the means and that overlap of the two CIs implies there is no significant difference. This rule is widely believed, but incorrect, and refers to CIs on separate means. In fact, nonoverlap of the two CIs does imply a significant difference, but with p distinctly less than .05, an overlap does not necessarily imply there is no statistically significant difference at the .05 level.

We sought participation by researchers who had published in journals in psychology (Psy), behavioral neuroscience (BN), or medicine (Med). (Psy was psychology other than BN.) These disciplines are of interest because they have very different customs for use of interval estimates. Using published author contact e-mail addresses, we sent invitations for authors to visit an experimental Web site where they were asked to adjust a simple figure of two

Please imagine that you see the figure below published in a journal article.

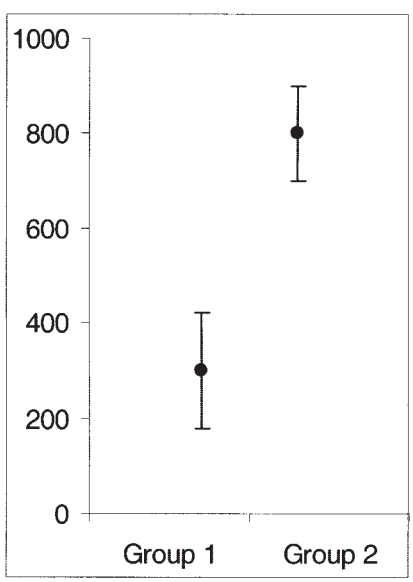

Figure 1. Mean reaction time (ms) and 95% Confidence Intervals for Group 1 (n=36) and Group 2 (n=34).

Please click a little above or below the mean on the right: You should see the mean move up or down to your click (the first time it may take a few seconds to respond). Please keep clicking to move this mean until you judge that the two means **are just significantly different** (by conventional t-test, two-tailed, p < .05). (I'm not asking for calculations, just your approximate eyeballing.)

Figure 1. The applet and instructions seen by a participant in the confidence interval (CI) task. The Group 1 mean was always fixed at 300. After clicking to shift the Group 2 mean until it was judged just statistically significantly different from the Group 1 mean, the participant clicked a button to see the next screen, completed some questions, and then submitted his or her response. For the standard error (*SE*) task, the two means were similarly labeled *Group 1* and *Group 2,* and the figure caption read "Mean reaction time (ms) and *SE* bars for Group 1 (n=36) and Group 2 (n=34)." For the repeated measures task, the two means were labeled *Pre Test* and *Post Test,* and the figure caption read "Mean reaction time (ms) and *SE* bars for one group (n=36): pre-test scores and post-test scores." Instructions were otherwise the same for all tasks. Each participant saw only one task.

means, with error bars, until they judged the two means to be just statistically significantly different.

Method

Before undertaking the main study, to assess use of CIs and *SE* bars in the three disciplines, we examined 978 empirical articles published in 1999–2001 in 33 leading journals across Psy, BN, and Med. The percentage of articles that reported CIs as numerical values, CIs as error bars in a figure, or *SE* bars in a figure was in each case and for each discipline 12% or less, except that 64% of Med articles reported CIs as numerical values, and 44% of BN articles included a figure with *SE* bars. Broadly speaking, researchers in Psy have relatively little exposure to CIs or *SE* bars; in Med, CIs are routinely reported in tables, but error bars are seldom shown in figures; and in BN, CIs are rarely used, but *SE* bars are often shown in figures. Disciplines overlap, but broadly, we studied three communities of researchers who have had markedly different experience with interval estimates.

For our main study, we sent e-mails to 3,944 authors of research articles published in leading journals in the three disciplines. We selected 21 Psy, 6 BN, and 5 Med predominantly empirical journals that have high impact factors and were accessible to us, and we used author e-mail addresses from articles in every second issue. Working in 2001 and early 2002 we started with the most recent available issue and then worked backward. The earliest issues used were from 1998. We discarded any e-mail address appearing in more than one discipline and used any address only once. In our e-mail invitations we made no mention of any discipline or that we were studying more than one discipline. Our return e-mail address gave no clue of our psychology affiliation. Likewise, the URLs for our experimental Web sites contained no clue to our discipline or to the discipline of the participants directed to a particular site. We asked participants not to respond more than once and not to pass our invitation to anyone else.

Researchers who agreed to respond followed a link to one of our Web sites, where they saw a display such as that in Figure 1. An applet allowed the respondent to click to move the Group 2 mean with attached 95% CI up or down, until the two means were judged to be just statistically significantly different ($p < .05$, two-tailed). Careful clicking could move the mean by as little as about three units of the vertical scale, so quite precise positioning was possible. A participant's response was the position of the adjustable mean, on the scale shown on the vertical axis, when he or she clicked to leave the main display and proceed to the next screen. On the next screen were some questions and a button to click to submit responses.

Approximately one third of the respondents in each discipline saw the CI task described in Figure 1. Another third saw the same display, except the figure caption read "Mean reaction time (ms) and *SE* bars for Group 1 (n=36) and Group 2 (n=34)." (We assigned the groups different sizes to reinforce the message that the means were of independent groups.) The remaining third saw a display with *SE* bars in which the two means were labeled in the figure *Pre Test* and *Post Test,* rather than *Group 1* and *Group 2,* and the figure caption read "Mean reaction time (ms) and *SE* bars for one group (n=36): pre-test scores and post-test scores." We refer to these three tasks as the CI, *SE,* and repeated measures (RM) tasks, respectively. Except for variations in the figure caption, the instructions were the same for all tasks. We designed the tasks and instructions to be as simple and brief as possible, in order to discourage calculations and to maximize the likelihood that potential respondents would complete the task and submit a response. Each participant saw only one task.

Pilot testing revealed an anchoring effect, in that the initial position of the Group 2 mean influenced a participant's response, that is, where he or she positioned the Group 2 mean. Therefore, use of any single initial position would give results contaminated to an unknown extent by the anchoring influence of that initial position. We thus chose to use two initial positions: Approximately half the participants in each discipline–task combination saw the Group 2 mean positioned initially at 800, as in Figure 1. The other half saw it initially at 300.

Results and Discussion

After allowing for undeliverable e-mails, we found 15.2% (473 of 3,122) of authors we approached submitted usable responses; a further 22.1% visited the Web site but elected not to complete the task or, in a minority of cases, found the applet nonfunctional. There was little variation in these percentages over discipline or task, although the group sizes varied somewhat (see Figure 2, and the text below for the RM task). All comments from participants suggest they took the task seriously. The computer logs gave no evidence of multiple responding.

As expected, an anchoring effect was observed, in that for every discipline–task combination the average response was higher for the 800 initial position than for the 300. For each combination, we calculated the difference between these two averages. We then adjusted all responses by subtracting half this difference from each response with an 800 initial position and adding half for each response with 300. Figure 2 is based on the adjusted responses. The wide spread of responses shown in Figure 2 thus cannot be attributed primarily to an anchoring effect. The overall mean differ-

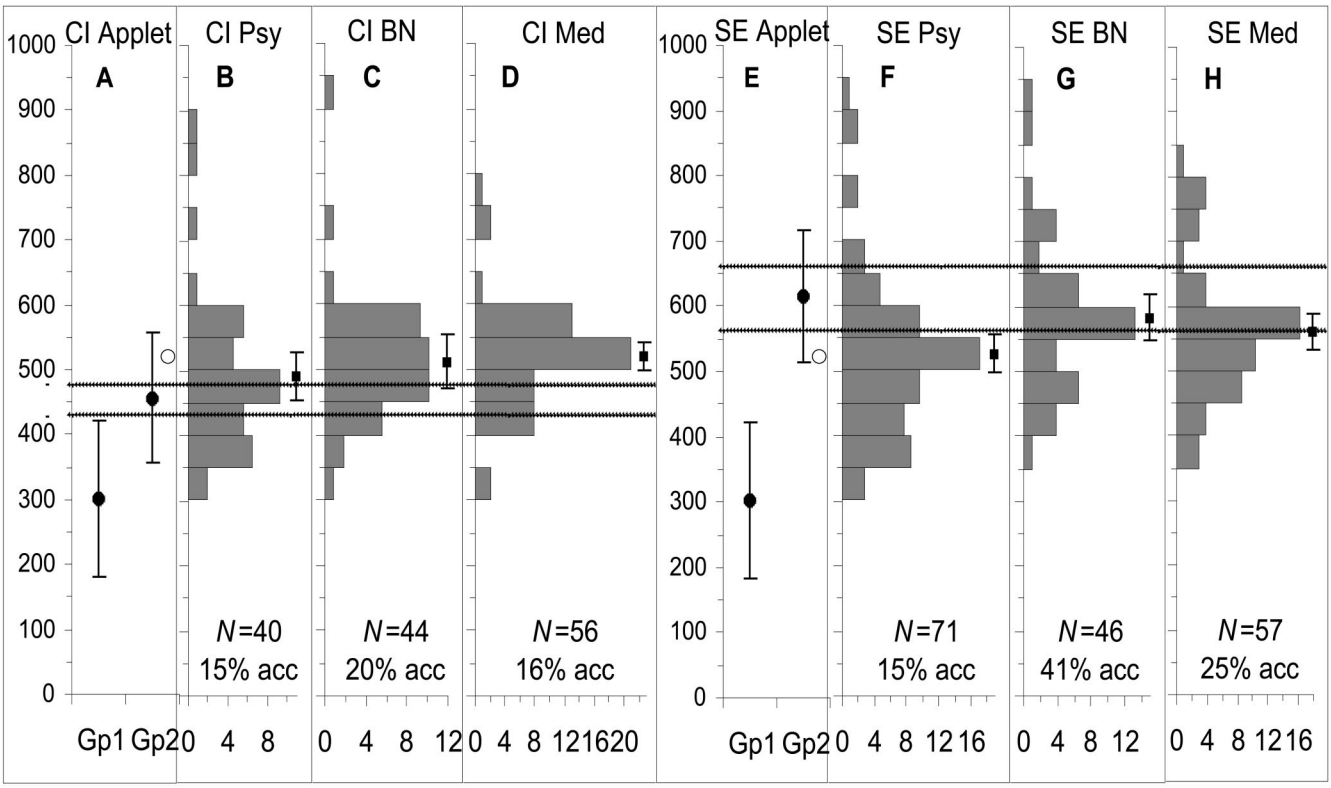

Figure 2. The judgment task and results. A: The correct configuration for the confidence interval (CI) task. The p value for the difference between the Group 1 (Gp1) and Group 2 (Gp2) means is .05. Horizontal lines show the position of the Group 2 mean for $p = .025$ (upper line) and $p = .10$ (lower line). The open circle is the position of this mean if intervals are set to just touch. B–D: Frequency histograms for Group 2 mean positions set by psychology (Psy), behavioral neuroscience (BN), and medicine (Med) respondents, respectively. Filled squares are mean responses, with 95% confidence intervals. For each discipline, the number of respondents (N) is shown and also the percentage of responses considered accurate (acc) because they fall between the horizontal lines. E–H: The corresponding information for the standard error (*SE*) task.

ence between the averages for the two initial positions was 53. This sizable anchoring effect suggests that many respondents may not have been very confident in their ability to accurately carry out the task.

Figure 2A shows for the CI task, the correct positioning of the Group 2 mean at 454 for a p value of .05 (Cumming & Finch, 2005; Wolfe & Hanley, 2002), calculated with the conventional t test for independent groups. The horizontal lines represent a reasonable range for answers to be regarded as accurate, in that the p value would be within a factor of 2 of the target .05.

The configuration of means and CIs in Figure 2A indicates that $p = .05$ corresponds to 95% CIs on independent means that overlap by a little more than one quarter of the full width of either interval, at least for our example. Cumming and Finch (2005) investigated how overlap and p values relate as sample sizes and CI widths vary, without assuming homogeneity of variance. They found a simple and useful relationship: 95% CIs that overlap by one quarter of the average length of the two intervals yield p values very close to, or a little less than, .05, provided only that both sample sizes are at least 10 and the two CIs do not differ in width by a factor of more than 2. They also found a simple relationship for *SE* bars, as illustrated in Figure 2E: *SE* bars that have a gap equal to the average of the two *SE*s yield p values close to .05, with the same provisos.

Figures 2B–2D show for the CI task and for each discipline the frequency histogram of responses; also shown is the mean response with the 95% CI. Responses were extremely varied, with only a small proportion being accurate; the percentages of responses falling within our accuracy range are shown for each discipline. Overall, the three disciplines did not differ greatly. Figure 2E shows the correct configuration for the *SE* task, with the Group 2 mean

at 614; Figures 2F–2H show that, for this task also, responses were extremely varied, few were accurate, and the disciplines did not differ substantially. The variation in responses is striking: Averaged over all disciplines and both tasks, the absolute error in positioning the Group 2 mean was for 65% of respondents greater than 25% of the correct gap between means and for 33% was greater than half the correct gap.

Respondents to our CI task were generally too strict: They tended to set the means too far apart, not realizing that the .05 statistical significance borderline requires overlap as illustrated in Figure 2A. Their mean response corresponds to $p = .009$ (Psy, $p = .017$; BN, $p = .008$; Med, $p = .006$) rather than the target .05. By contrast, *SE* task respondents were generally too lax: They tended to set the means too close together, not realizing that a gap is required for .05 statistical significance, as shown in Figure 2E. Their mean response corresponds to $p = .109$ (Psy, $p = .158$; BN, $p = .078$; Med, $p = .104$). (The p values for median responses were generally very similar to those for means.) BN researchers, who often see *SE* bars in their journals, may take little comfort in their greater accuracy for the *SE* task, because only a minority (41%, Figure 2G) positioned the mean within our accuracy range.

No respondent saw both tasks, but overall, the researchers did not sufficiently distinguish CI and *SE* bars: The correct response for the *SE* task (614) was 160 higher than that for the CI task (454), but the observed difference was, on average, only 48 (Psy, difference = 35; BN, difference = 71; Med, difference = 39). The full interval marked by *SE* bars is about half the width of a 95% CI and gives a 68% CI, unless sample size is very small. It is seriously unfortunate that an identical graphic, the error bar, can have two such different meanings.

After setting the mean, the respondent clicked to move to a second screen where we asked two general, open-ended questions. The first asked how the respondent approached the task, and the second invited any further comments. To maximize the response rate, we kept questions to a minimum. To avoid prompting an overly analytic attitude or the seeking of statistical advice, we did not ask probing questions about the respondent's statistical understanding, even though it would be interesting to have more information about respondents, their areas of research specialization, and their understanding of error bars and inference.

Overall, 59% of respondents to the CI and *SE* tasks typed comments about how they did the task. These comments were very diverse, often brief, and resisted any useful analysis, except that we noted 61% included some statement that was clearly or probably statistically incorrect, such as "I positioned the means to be 2 *SE*s apart" (about 3 is correct), "I moved the mean so it was just outside the other error bars," or a statement that seemed to confuse *SD* and *SE*. We also asked how many years ago the respondent published his or her first journal article. Responses ranged from 0 to 48 years (*Mdn* = 10), but there was no sign of any relation with accuracy of response.

Error Bars That Just Touch

Schenker and Gentleman (2001) pointed out how severely erroneous is the rule of thumb widely believed in medicine and health sciences research that statistical significance corresponds to 95% CIs that just touch. In fact, when sample sizes are similar and not small and CI widths are similar, if 95% CIs on independent means just touch, the two-tailed p value is about .006 (Cumming & Finch, 2005; Payton, Greenstone, & Schenker, 2003), not .05 as many believe. When intervals representing *SE* bars on independent means just touch, the p value is about .16, again for similar sample sizes that are not small and similar *SE*s. To investigate whether respondents tended to set error bars to just touch, we examined a version of Figure 2 without the adjustment for anchoring and with narrower histogram bins. The unadjusted histograms were similar to those shown in Figure 2, but for all disciplines and for both CI and *SE* tasks there was a distinct peak around 520, which corresponds to error bars just touching. A bin centered on 520 and of width 50 (as used in Figure 2) captured the peaks best, so for each discipline–task combination, we calculated the percentage of respondents who set the Group 2 mean within 25 of 520. Figure 3 shows these percentages. As the widths of the 95% CIs in Figure 3 suggest, the groups were not sufficiently

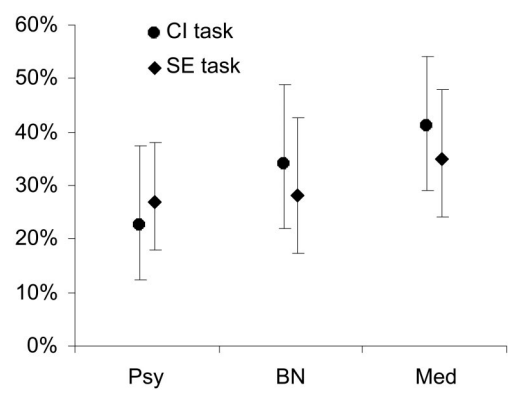

Figure 3. Percentage of respondents to the confidence interval (CI) and standard error (*SE*) tasks who positioned the Group 2 mean so the two intervals just touched. For each task, the accurate value for just touching was 520, which is marked with an open circle in Figures 2A and 2E. A response within 25 of 520 was defined as intervals just touching. Results are shown for six separate groups of participants, for the six discipline–task combinations. Error bars are 95% CIs (Altman, Machin, Bryant, & Gardner, 2000, chap. 6). Psy = psychology; BN = behavioral neuroscience; Med = medicine.

large to justify detailed inference, but overall, respondents in Med (38.1%, 43/113) may have used the rule somewhat more frequently than respondents in BN (31.1%, 28/90) or Psy (25.2%, 28/111). In addition, it is striking that overall, the (incorrect) rule of thumb was misapplied to *SE* bars (29.9%, 52/174) about as often as it was applied to CIs (33.6%, 47/140). Once again many respondents did not seem to appreciate the distinction between CI and *SE* bars.

Designs With a Repeated Measure

Figure 4 shows fictitious means and *SE* bars for a two-way design with one repeated measure. We chose this figure to discuss the importance of type of independent variable because it allows a contrast of the two main types: independent groups (between-subjects) and repeated measures (within-subjects) variables. In addition, such figures are common in BN journals, and one appears in the *Publication Manual* (APA, 2001, p. 180) as an example of good practice. For between-subjects comparisons, such as a1 with b1, the *SE* bars can be used to guide inference: If they are further apart than those in Figure 2E, the *p* value for the single comparison is less than .05. For repeated measures comparisons, however, such as b1 with b2, the *SE* bars shown are quite irrelevant, because they take no account of correlation between the measures (Cumming & Finch, 2005). The appropriate interval is the *SE* bars (or CI) on the mean of the differences between paired b1 and b2 scores, and the width of this interval varies markedly with the correlation. Even assuming the correlation to be positive, as is usually the case in practice, the *SE* bars on the difference could be anywhere from wider than the *SE* bars in Figure 3

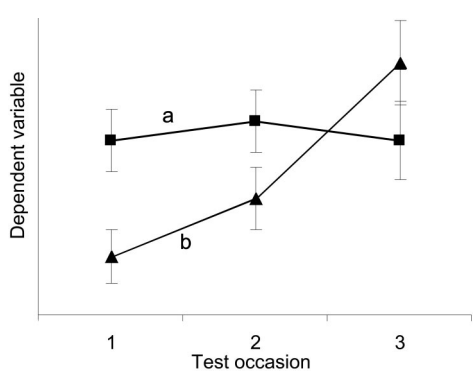

Figure 4. A two-way design with one repeated measure. Two independent groups, *a* and *b,* were each measured at Test Occasions 1, 2, and 3. Standard error bars are shown for each mean. Data are fictitious. Error bars on individual cell means, as shown here, may legitimately be used to assess the statistical significance of between-subjects comparisons, for example a1 with b1. They may not, however, be used to assess within-subjects comparisons, for example b1 with b2.

down to practically zero. Respondents to our RM task saw the equivalent of b1 and b2 and so had insufficient information for an accurate response.

In designing our RM task we chose pre- and posttests as measures that would most naturally signal a repeated measures independent variable. The labels *Pre Test* and *Post Test* appeared at the bottom of the figure, and as described earlier, there was a statement that the data were for a single group. Respondents had more, and clearer, signals of a repeated measure than is often the case in figures in journal articles. Our interest was whether respondents would realize the implication for inference. On the second screen we deliberately did not ask whether they considered the task to be possible, because we wished to avoid an overly analytic attitude; after all, a journal article does not print a warning beside a figure that making the natural comparison may not be justifiable!

We reasoned that a potential respondent who recognized that the RM task could not be solved would either decide not to submit a response or would mention the problem in an answer to our open-ended questions. There was no sign, however, of a lower response rate to the RM task. In analyzing the typed answers, we classified any mention of the issue or any doubt expressed about the task as recognition of the problem, whether or not the respondent had moved the Post Test mean. Only 11% of RM respondents gave any such recognition; otherwise, their responses resembled those for the *SE* task. There were no clear differences among disciplines: For Psy, 3/51 (6%) recognized the problem; for BN, 8/47 (17%) did; and for Med, 7/61 (11%).

In figures like Figure 4, with both between- and within-subjects factors, we suspect few researchers appreciate that error bars of this kind may be used only for between-subjects comparisons. It is a serious problem that the usual graphical conventions, as in Figure 4, do not make salient whether a factor is a between- or within-subjects factor.

Conclusions

We studied understanding of CIs and *SE* bars in relation to statistical significance. We hope, however, that researchers will use interval estimates much more broadly than merely as indicators of *p* values. Cumming and Finch (2001, 2005), for example, emphasized the value of CIs as giving point and interval estimates in measurement units that should be readily comprehensible in the research situation, also as helping to combine evidence over experiments and thus supporting meta-analysis and meta-analytic thinking, and further as giving information about precision that may be more useful than a calculation of statistical power.

Our response rate was low, but we think it reasonable to assume that nonrespondents, including those who visited the site and elected not to complete the task, would if

anything be less statistically confident and competent than respondents. If so, our findings are underestimates of the severity and prevalence of misconception among researchers in the three disciplines.

We conclude that very many researchers whose articles have appeared in leading journals in psychology, behavioral neuroscience, and medicine have fundamental and severe misconceptions about how CIs and *SE* bars can justifiably be used to support inferences from data. Misunderstanding seems little influenced by experience with different disciplinary customs for error bar use: Misconceptions are severe in some disciplines with established error bar or CI use (BN, Med) but also in Psy, in which use of CIs is small but growing. Our findings are consistent with those of Cumming, Williams, and Fidler (2004), who described misconceptions held by many researchers about the relation between error bars and replication and who also found no differences among the three disciplines.

We identified four different problems. First, responses were very widely spread and inaccurate: Only 22% of respondents set the means so the p value was between .025 and .10. Second, respondents overall did not adequately distinguish CIs and *SE* bars, as if they did not sufficiently recognize that a single graphic is used for two very different indicators of precision. Third, many respondents (overall 31.5%, 99/314) used the incorrect rule that error bars, whether a 95% CI or *SE* bars, just touch when means are just statistically significantly different ($p < .05$). Finally, for the RM task, a large majority in every discipline apparently overlooked the clear statements that the means they saw were from a repeated measures or paired design. They did not appreciate the crucial role of experimental design in the interpretation of intervals.

Each of these four misconceptions is a major problem for accurate inferential use of CIs and *SE* bars, as they are commonly reported on cell means. The fact that all four misconceptions appear to be widespread among both early career and established researchers, across three disciplines, is a major impediment to more effective use of statistical estimation to improve research communication.

Advocates of statistical reform put forward strong reasons why emphasis should swing from NHST to statistical estimation (Finch et al., 2002; Kline, 2004), but the misconceptions we have identified threaten such reform. Our results do, however, point to developments necessary if the benefits of CIs are to be realized. Better training and guidance are needed for researchers (Grissom & Kim, 2005; Kline, 2004; Smithson, 2002), and this should include accurate advice about when and how overlap of CIs or *SE* bars can be used to assess a difference (Cumming & Finch, 2005). In addition, we need better graphical conventions for displaying interval estimates that reduce ambiguity, make the status of independent variables salient, and signal more clearly how intervals may be used for data interpretation (Cumming & Finch, 2005; Masson & Loftus, 2003).

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.

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

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

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., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170–180.

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

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science, 15,* 119–126.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology:* Little evidence of reform. *Educational and Psychological Measurement, 61,* 181–210.

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

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach.* Mahwah, NJ: Erlbaum.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Kline, R. B. (2004). *Beyond significance testing. Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57,* 203–220.

Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* Chichester, England: Wiley.

Payton, M. E., Greenstone, M. H., & Schenker, N. (2003). Overlapping confidence intervals or standard error intervals: What do

they mean in terms of statistical significance? *Journal of Insect Science, 3,* Article 34. Retrieved May 16, 2005, from www.insectscience.org/3.34

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. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181.

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

Smithson, M. (2002). *Confidence intervals.* Thousand Oaks, CA: Sage.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25,* 26–30.

Thompson, B. (1999). Why "encouraging" effect size reporting is not working: The etiology of researcher resistance to changing practices. *Journal of Psychology, 133,* 133–140.

Wilkinson, L., & the Task Force on Statistical Inference. (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.

Received August 9, 2004
Revision received December 14, 2004
Accepted April 21, 2005 ■

Call for Nominations

The Publications and Communications (P&C) Board has opened nominations for the editorships of **Behavioral Neuroscience, JEP: Applied, JEP: General, Neuropsychology, Psychological Methods,** and **Psychology and Aging** for the years 2008–2013. John F. Disterhoft, PhD; Phillip L. Ackerman, PhD; D. Stephen Lindsay, PhD; James T. Becker, PhD; Stephen G. West, PhD; and Rose T. Zacks, PhD, respectively, are the incumbent editors.

Candidates should be members of APA and should be available to start receiving manuscripts in early 2007 to prepare for issues published in 2008. Please note that the P&C Board encourages participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. Self-nominations also are encouraged.

Search chairs have been appointed as follows:

- **Behavioral Neuroscience:** Linda P. Spear, PhD, and J. Gilbert Benedict, PhD
- **JEP: Applied:** William C. Howell, PhD
- **JEP: General:** Peter A. Ornstein, PhD
- **Neuropsychology:** Susan H. McDaniel, PhD, and Robert G. Frank, PhD
- **Psychological Methods:** Mark Appelbaum, PhD
- **Psychology and Aging:** David C. Funder, PhD, and Leah L. Light, PhD

Candidates should be nominated by accessing APA's EditorQuest site on the Web. Using your Web browser, go to http://editorquest.apa.org. On the Home menu on the left, find Guests. Next, click on the link "Submit a Nomination," enter your nominee's information, and click "Submit."

Prepared statements of one page or less in support of a nominee can also be submitted by e-mail to Karen Sellman, P&C Board Search Liaison, at ksellman@apa.org.

Deadline for accepting nominations is **January 20, 2006,** when reviews will begin.