

Running head: REPLICATION AND CONFIDENCE INTERVALS

(A copy-edited version of this article appears as:

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

Replication, and Researchers' Understanding of Confidence Intervals and  
Standard Error Bars

Geoff Cumming, Jennifer Williams, and Fiona Fidler

La Trobe University

Melbourne, Victoria, Australia

Contact author:

Geoff Cumming

Mail:

School of Psychological Science, La Trobe University, Victoria, Australia 3086

Tel: +61 3 9479 2820 Fax: +61 3 9479 1956

Email: G.Cumming@latrobe.edu.au

## Abstract

Confidence intervals (CIs) and standard error bars give information about replication, but do researchers have an accurate appreciation of that information? Authors of journal articles in psychology, behavioural neuroscience, and medicine were invited by email to visit a website and indicate on a figure where they judged replication means would plausibly fall. Responses from 263 researchers suggest that many leading researchers in the three disciplines under-estimate the extent that future replications will vary. A 95% CI will on average capture 83.4% of future replication means. A majority of respondents, however, hold the confidence level misconception (CLM) that a 95% CI will on average capture 95% of replication means. Better understanding of CIs is needed if they are to be successfully used more widely in psychology.

## Replication, and Researchers' Understanding of Confidence Intervals and Standard Error Bars

Given a sample mean and 95% confidence interval (CI), what is the probability a replication of the experiment would give a sample mean within the original CI? This is readily calculated (Estes, 1997), but seldom mentioned in psychology textbooks. Further, what do researchers believe this probability to be? We report data on this question.

Statistical reformers argue that psychology's over-reliance on null hypothesis significance testing (NHST) is damaging (Schmidt, 1996), and that wider use of CIs, among other techniques, would improve research communication (Cumming & Finch, 2001; Harlow, Mulaik, & Steiger, 1997; Wilkinson & Task Force on Statistical Inference, 1999). Finch, Thomason, and Cumming (2002) and Nickerson (2000) provide reviews. A pithy statement of some key features of CIs is given in the American Psychological Association (APA) *Publication Manual*: "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).

Part of the case for statistical reform is that researchers have severe misconceptions about NHST (Oakes, 1986). Although cogent reasons are given for preferring CIs, there is little evidence that they escape misconception. Can psychologists understand and use CIs successfully, and thus realise the advantages claimed for them? Psychology insists that its professional practice

should be evidence-based, and so it is unfortunate that statistical reformers provide little evidence of effectiveness to support the sweeping changes they recommend. We seek to contribute to the currently sparse knowledge of statistical cognition of CIs by investigating researchers' understanding of error bars in relation to replication. Our focus is not so much on judging researchers, but on assessing the cognitions prompted by CIs shown conventionally in a figure. We study graphically-presented CIs, and standard error (SE) bars which are  $\pm 1$  SE of the mean.

Figure 1 about here

### *Error Bars and Replication*

Figure 1 shows, to the left of each panel, the population mean  $\mu$  and SD  $\sigma$ . Next is the mean and 95% CI for our original sample of size  $n = 36$  from this population. To the right is the sampling distribution of sample means expected if the population is normally distributed and we took many replications. This is a normal distribution with mean  $\mu$  and SD of  $\sigma / \sqrt{n}$ , so 95% of replication means will fall within  $1.96\sigma / \sqrt{n}$  either side of  $\mu$ . In addition we expect, in the long run, 95% of CIs calculated for repeated samples will include  $\mu$ . Figure 1 illustrates these two types of variability: First, the sampling distribution summarises how replication means vary about  $\mu$  and, second, the two examples of original means show how successive CIs may capture or miss  $\mu$ .

Our opening question, ‘what is the probability the next replication mean will fall within the original 95% CI?’ might prompt the answer ‘.95, of course’. However Figure 1 illustrates how the two sources of variability combine to make the answer less, on average, than .95: The original CI will typically be centred a little distance from  $\mu$  and, as well, replication means vary around  $\mu$ . The probability will be close to .95 only when the original mean happens to fall very close to  $\mu$ , and the capture rate will drop as it falls further from  $\mu$ . Different original CIs will differ greatly in the percentage of replication means they include. For many samples the mean will be fairly close to  $\mu$ , and so the probability will be only a little less than .95. For some samples, however, the mean will happen to fall some distance from  $\mu$ , and the probability will be considerably smaller than .95. For example, 5% of original CIs will lie entirely above or below  $\mu$ , and so include less than 50% of replication means, as the right panel of Figure 1 illustrates.

Figure 2 about here

Appendix (1) explains the formula for the probability that, on average, a replication mean falls within an original CI. Using large sample methods, the probability is .834, meaning just 5 out of 6 replication means will, on average, fall within a 95% CI. Figure 2 shows how this probability varies with  $C$ , the level of confidence. It illustrates the extent to which the probability is less than  $C/100$  because of the combination of the two sources of variability illustrated in Figure 1. For SE bars, which for large samples correspond to CIs with  $C = 68.3$ , the chance is close to 50-50. Note again that these probabilities are

averages, and conceal great variation in capture rates for different original CIs, as illustrated in Figure 1.

### *Researchers' Understanding*

We conducted an internet-based investigation of researchers' understanding of CIs, SE bars, and replication. We sent emails to authors of journal articles in psychology (PSY), behavioural neuroscience (BN), and medicine (MED). We chose these disciplines because, as Belia, Fidler, Williams and Cumming (2004) reported, the three have different practices for use of CIs and SE bars. In MED about two-thirds of empirical articles include CIs as numerical values in tables, but CI or SE bars are seldom shown in figures. In BN almost half of empirical articles show SE bars in a figure, but few report CIs. In PSY, relatively small proportions of articles include CIs in any form, or SE bars in a figure. Disciplines overlap but, broadly, we studied three groups of researchers with different experiences with interval estimates.

### Method

We selected 20 PSY, 9 BN and 5 MED predominantly empirical journals that have high impact factors and were accessible to us, and used author email addresses from articles in alternate issues. In 2001 and early 2002 we started with the most recent available issue, then worked backwards as we needed further addresses, but not earlier than 1998. We discarded any email address appearing in more than one discipline, and used any address only once. We sent email invitations to 3,848 authors. Our emails made no mention of any discipline. Our return email address gave no clue of our psychology affiliation.

Likewise the URLs for our experimental websites 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.

Figure 3 about here

Those who responded clicked on a link to a website where they saw a display such as that in Figure 3, but without the horizontal lines on the right. There were six websites: The CI site for each discipline showed, as in Figure 3, the mean dot at 750, and error bars each of width 300. (We follow the convention of referring to the 'width' of each bar that makes a CI.) The SE site for each discipline was the same except the bar widths were each 150, and the text described the bars as showing "+/- 1 SE (standard error of the mean)". Any participant visited only one site and so completed only one task. At a CI site, the caption was "Figure 1. Mean reaction time (ms) and 95% Confidence Interval, for a single group (n=36)." The instructions were

The large dot represents the mean of the sample of  $n = 36$  participants. The bars show a 95% confidence interval around the mean. Of course if the experiment were replicated, the mean obtained would almost certainly not be exactly the same as the mean shown. Please plot ten lines that you think **could plausibly be a set of means for ten further independent samples of the same size** taken from the same population...

An applet allowed the respondent to click to position horizontal lines such as those shown in Figure 3. An undo button allowed repositioning of lines. After placing ten lines the participant clicked for the next screen, completed some questions, then submitted his or her response.

### Results

After allowing for undeliverable emails, 8.7% of authors we approached submitted responses; a further 16.3% visited the website but elected not to complete the task, or, in a minority of cases, found the applet non-functional. These percentages were similar over discipline and task. All comments from participants suggest they took the task seriously. The computer logs gave no evidence of multiple responding.

#### *Characteristics of Replication Means*

A set of 10 replication means is a random sample from the sampling distribution shown in each panel of Figure 1. The mean of the 10 means in a set is  $M$ . Some of the properties of such sets are:

1. Most  $M$ s will fall above, near, or below the original mean, depending on where that original mean falls in relation to  $\mu$ . Averaged over all original samples,  $M$  is expected to fall at the original mean, 750 in our case, as in Figure 3 **a** and **c**.
2. The SD of the values in a set about their own mean  $M$  is  $SD_M$ .

Appendix (2) explains why  $SD_M$  will average around 153 for original CI bars, and 150 for original SE bars, as in Figure 3 **a** and **b**.

3. The SD of the values about the original mean of 750 (rather than about  $M$ ) is  $SD_O$ . Appendix (3) explains why  $SD_O$  will average around 216 for the CI case, and 212 for the SE case, as in Figure 3 **b** and **c**. These values are larger than those for  $SD_M$  because they encompass both sources of variation illustrated in Figure 1.
4. The ten mean values in a set are likely to be bunched, meaning that central values tend to be relatively close together and the highest and lowest values more widely spaced. We define a bunching index as the average of the gap between the two highest means, and that between the two lowest means, divided by the average of the 3 middle gaps. If the means are, from lowest to highest,  $M_1, M_2, \dots, M_{10}$ , the index is

$$\{[(M_{10} - M_9) + (M_2 - M_1)]/2\} / [(M_7 - M_4)/3].$$

To estimate the expected degree of bunching, we calculated the bunching index for each of 1,000,000 independent random samples of size 10 from a normal distribution. The mean was 3.0.

5. All 4 characteristics described above are averages, and there will be considerable set-to-set variation about these averages.

No set can have all of properties 1, 2, and 3. Figure 3 shows three possible responses to the CI task: Set **a** has properties 1 and 2, but not 3; **b** has 2 and 3, not 1; and **c** has 1 and 3, not 2. Each has a bunching index of 3.0.

Figure 1 shows that set **b** is the most typical pattern for replication means, where  $M$  for the set has some offset from the original mean. For both  $SD_M$  and  $SD_O$  to have their expected average values, the offset of  $M$  from 750 is

161 (for the SE task, 157). Set **a** would be the most typical only for an original mean very close to  $\mu$ . Set **c** would not typically be obtained, but shows the spreading of means necessary for a set with  $M = 750$  to have  $SD_O$  equal its expected average.

We asked participants to mark a *plausible* set of replication means. Because *any* set of 10 values is a *possible* set of means, we cannot score responses as correct or incorrect, but can assess them against the properties described above, and the three patterns illustrated in Figure 3.

Figures 4 and 5 about here

#### *Data Analysis*

Figure 4 shows that  $M$  was on average positioned close to 750 by respondents in all groups, and that only one participant had  $M$  offset as far from 750 as  $M$  is in **b** (dotted lines with diamonds). Figure 4 also shows the  $N$  for each of the 6 groups. The filled squares in Figure 5 show that  $SD_M$  was on average close to, or a little larger than, the property 2 expected values (dotted lines with open squares); only for the Med CI group was the expected value just outside the 95% CI. Average  $SD_O$  values (filled triangles in Figure 5) were close to the  $SD_M$  values, and these means and their CIs are all clearly below the property 3 expected values (dotted lines with open triangles). The data thus indicate that respondents tended to centre their 10 means near 750, with SD little more than 150; they generally gave their sets properties 1 and 2, but not 3. The variability about the original mean (overall average  $SD_O$  164) was generally much too small to represent how, on average, replication means will

actually vary (average  $SD_o$  for property 3 is 214). However,  $SD_o$  varied considerably over participants, and 20.9% had  $SD_o$  at least as large as expected for property 3. Overall, the majority of responses were, for all groups, generally similar to Figure 3 **a**, and generally not like **b** or **c**.

Figure 6 about here

An alternative analysis of variability is to consider how many means in a set were placed within the original bars. For the CI task, the average number of means within the bars is expected to be 9.50 for the assumptions of Figure 3 **a** (properties 1 and 2), 8.18 for **b**, and 8.34 for **c**. For the SE task the averages expected are 6.83 for **a**, 4.60 for **b**, and 5.21 for **c**. Figure 6 gives the frequencies for the numbers of means placed within the bars for each group. In the CI condition, 54.5% of respondents included all 10 means within the bars; 78.4% placed 9 or 10 within the bars. For the SE condition, 16.3% included all 10 within the original bars, and 69.0% included from 6 to 10. The CI average was 9.18 means within bars, and the SE average 6.67. This analysis supports our conclusion that **a**, rather than **b** or **c**, generally typifies our participants' responses.

In other words, a large proportion of responses are generally consistent with what we call the *confidence level misconception* (CLM), which is the belief that about  $C\%$  of replication means will fall within an original  $C\%$  CI. Respondents have the CLM if they believe that about 95% of future means will fall within the original CI, rather than the 83% described earlier (see also Figure

2 and the Appendix), or that about 68% will fall within SE bars, rather than 52%.

In all, 94.3% (248/263) of respondents typed comments in response to our request to “please explain how you approached the task”. Comments were coded independently by two of the authors, and the 4.3% of codings that had initial disagreement were resolved by consultation. Fully 26.6% (66/248; 27.0% for the CI task, 26.2% for SE) of comments included an explicit statement of the CLM, such as, for the CI task: “In 10 replications, 9.5 of them should be within the confidence interval”, and for the SE task, “I assumed that 68 percent of sample means should be within one SE of the first mean”. (For SE, we took from 60 to 75% as indicating the CLM.) A further 19.0% (47/248; 19.0% for the CI task, 18.9% for SE) of comments included a more extreme quantitative statement, to the effect, for example, that all the means should be within the original CI, or 95% within the original SE bars. Thus 45.6% (113/248) of comments included a quantitative statement of the CLM or a more extreme misconception. Many more comments included non-quantitative statements consistent with the CLM, such as “means should be rather similar and very close”, or “I assumed nearly all of the means would fall within the SEM shown”.

Perhaps, however, participants recognized that  $SD_o$  should be as high as property 3 requires, but, being unable to represent all three properties in a single set, gave higher priority to properties 1 and 2 and thus gave a response like **a**. This would require the additional assumption that the original mean had fallen

at or very close to  $\mu$ , but only 8.8% (10/113) of the quantitative statements of misconception included any comment about assuming the original mean was close to, or an estimate of  $\mu$ .

The important insight, presumably incompatible with holding the CLM, is that there are two sources of variability, as Figure 1 shows: the original mean from  $\mu$ , and amongst the replication means. Only 7 comments (2.8%) avoided a statement of the CLM and gave any recognition of both these sources of variability.

For every response we calculated the bunching index, to assess the extent to which respondents gave their sets property 4. The overall mean was 2.8, a value close to the 3.0 expected for a normal distribution.

We asked how many years ago the respondent had published their first journal article. Responses ranged from 0 to 54 (median 11), but there was no sign of a relation with any response measure.

In all cases above for which we report overall data, the values for the CI and SE tasks, and for the three disciplines were similar.

### Discussion

Our most striking result is the extent to which researchers, when examining a conventional figure with error bars, underestimate the variability of replication means, and hold the CLM. Respondents generally centred their replication sets close to the original mean, and gave them  $SD_M$  and  $SD_O$  values whose averages were broadly consistent with the CLM, and generally

inconsistent with the high values of  $SD_o$  that replication means would, on average, actually show.

Why did we ask researchers to carry out the possibly strange task of giving a set of plausible replication means, rather than simply asking them what proportion of replication means they thought would fall within the original CI? We chose the task because we wanted to tap implicit understanding, and reduce the likelihood that a respondent might analyse or calculate before responding, as the direct question may have prompted. We then asked the neutral “please explain how...” question. It is remarkable that fully 46% of respondents who commented made a spontaneous quantitative statement of the CLM, or more extreme, and very few qualified their comment by stating they assumed the original mean was equal to  $\mu$ . Many more made non-quantitative comments consistent with the CLM. Very few comments mentioned the two sources of variability illustrated in Figure 1 that cause the expected  $SD_o$  value to be so high. We thus have converging evidence, from the sets of means selected and from the spontaneous comments, to support our main conclusion.

The data on placement of replication means within the original error bars give an additional and particularly strong indication that the CLM may often have guided respondents: Fully 74% of respondents placed 9 or more CI means, or 6 or more SE means within the original bars. We cannot give a quantitative estimate of the prevalence of the CLM, but the strength and consistency of the three different indications summarized above justify the

conclusion that conventional presentation of error bars in a figure prompts this misconception in a majority of our respondents, and possibly a large majority.

On the other hand, the responses revealed good appreciation of several aspects of replication means. Most basically, there was a realization that error bars give guidance about where replication means are likely to fall. This realization had a quantitative aspect: Recall that the original bars were twice the width for the CI as for the SE task. Figure 5 shows, however, that the variability within sets was similar for the CI and SE tasks. Any participant saw only one task, but this similarity indicates that respondents generally understood the difference between the two types of error bars and understood, at least implicitly and roughly, that SE bars correspond to a 68% CI. This result contrasts with that of Belia et al. (2004), who found that researchers make insufficient distinction between CI and SE bars when making judgments about statistical significance. The current study presented a single mean with error bars, whereas in the Belia et al. study respondents needed to consider the configuration of two independent means each with error bars. We speculate that the difference in complexity of the displays may be one reason for the difference in findings of the two studies; and that another reason may be the use by some Belia et al. respondents of mistaken heuristics about interval overlap that did not distinguish CI and SE bars.

A further aspect of accurate understanding concerned bunching: Many respondents made a comment about a normal distribution, or about means tending to cluster, with a few further away. The average bunching index was

very close to that expected for a normal distribution. Respondents thus generally seemed to appreciate that replication means can, given the central limit theorem, reasonably be expected to have a normal distribution, and respondents could on average give their replication means an appropriate amount of bunching.

No clear differences among disciplines emerged on any measure, despite the very different customs in relation to CIs and SE bars in psychology, behavioural neuroscience, and medicine. Belia et al. (2004) also found no differences among disciplines.

Our response rate was very low, despite our efforts to make the task accessible and brief, and this low response rate must qualify our conclusions. However our respondents presumably were, if anything, more statistically confident and competent than non-respondents, and so our results may underestimate the prevalence of the CLM. We conclude that a conventional figure with error bars elicits the CLM in a majority, and quite possibly a large majority, of world-leading researchers across psychology, behavioural neuroscience, and medicine. The belief that about 95% of future replication means will fall within the 95% CI of an initial experiment amounts to severe underestimation of the extent to which replications will vary from the result of an initial experiment. It corresponds to a belief that only about 5% of future means will fall outside the original CI, whereas in fact on average about 17% (i.e.,  $100 - 83.4$ ) will do so. This finding is consistent with the long-established Law of Small Numbers (Tversky & Kahneman, 1971), which holds that

researchers tend to underestimate sampling variability, and therefore believe replications will be unjustifiably close to a first result.

There are strong reform reasons why emphasis should swing from NHST to statistical estimation but, to reap the benefits of CIs, the graphical formats of presentation, and guidelines for CI interpretation must prompt accurate understanding by researchers. Researchers need, in particular, a more accurate appreciate of replication and of the chance that an original CI will include future means. We hope that Figure 1, which is intended to highlight the two sources of variability that contribute to the large variation over replication, may lead to more accurate understanding. Researchers need to bear in mind that, on average, just 5 out of 6 replication means (83.4%) will fall within an original 95% CI. Occasional CIs happen to be extreme, and have a much lower probability of capturing replication means.

#### References

- 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*. Submitted for publication.
- 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.

- Estes, W. K. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review*, 4, 330-341.
- Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory and Psychology*, 12, 825-853.
- Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.) (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
- 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: Wiley.
- Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1, 115-129.
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 92,105-110.
- Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

## Appendix

1. We follow Estes (1997) to find the average value of the probability an original  $C\%$  CI captures a replication mean. Assume  $\sigma$  is known, or  $n$  is sufficiently large—traditionally, at least 30—that it is reasonable to use the sample estimate as  $\sigma$ . The sampling distribution of the sample mean  $\bar{X}$  has mean  $\mu$  and variance  $\sigma^2/n$ . Consider the difference between the means of two independent samples of size  $n$  from the same population. The sampling distribution of this difference has mean zero and variance  $2\sigma^2/n$ , because variance is additive when independent variables (the two sample means) are subtracted. The CI on the first mean has bars of width  $w = z_C \times \sigma / \sqrt{n}$ , where  $z_C$  is the standard normal deviate corresponding to the central  $C\%$  area (1.96 for  $C = 95$ ). An interval of  $\pm w$ , centred in a normal distribution with variance  $2\sigma^2/n$ , defines under that distribution an area that is the probability we seek, assuming the population is normal, or sufficiently close to normal for the Central Limit Theorem to justify our assuming the sampling distribution is normal. This area corresponds to the standard normal deviate  $z = z_C / \sqrt{2}$ . For  $C = 95$ ,  $z_C = 1.960$ ,  $z = 1.386$ , and the probability is .834. Figure 2 shows how this probability varies with  $C$ .
2.  $SD_M$  is the SD of the values about their own mean  $M$ , and is calculated using as divisor the degrees of freedom (number of means in the set  $- 1$ )  $= 9$ .  $SD_M$  is a good estimate of  $\sigma / \sqrt{n}$ , the SE of the sample mean. For

the CI task, the bar width  $w = 300$  and  $SE = 300/1.96 = 153.1$ , because we are using large sample methods. For the SE task,  $w = 150$  and so  $SE = 150$ . We therefore expect  $SD_M$ , the SD of a set of replication means, to average around 153 for the CI task, and 150 for the SE task.

3. The sampling distribution, in 1 above, of the difference between two means implies that replication means will be normally distributed with mean equal to the original mean, and SD of  $\sqrt{2}\sigma / \sqrt{n}$ . For a set of 10 replication means we expect the root mean square deviation from our original mean 750 (rather than from the set mean  $M$ ) to average around  $\sqrt{2}\sigma / \sqrt{n}$ . This SD of the set about 750 is  $SD_O$ , and is calculated using the degrees of freedom 10 as divisor. We expect  $SD_O$  to average around  $\sqrt{2} \times 153.1 = 216.5$  for the CI task, and  $\sqrt{2} \times 150 = 212.1$  for the SE task.

### Author Note

Address correspondence to Geoff Cumming, School of Psychological Science, La Trobe University, Australia 3086; email:

[G.Cumming@latrobe.edu.au](mailto:G.Cumming@latrobe.edu.au)

Figures 1 and 2 are derived from live figures that make up one component of ESCI (“ess-key”; Exploratory Software for Confidence Intervals), which runs under Microsoft Excel. This component of ESCI may be downloaded, for personal use without cost, from [www.latrobe.edu.au/psy/esci](http://www.latrobe.edu.au/psy/esci)

### Acknowledgements

We thank the researchers who responded, Bradley Dean for website development, and Robert Maillardet and Thomas Matyas for comments. This research was supported by the Australian Research Council.

## Figure Captions

*Figure 1.* Two cases of a CI and replication means. To the left in each panel is the mean  $\mu$  and SD  $\sigma$  of the population, assumed normal. Then is the mean (filled dot) and 95% CI for our original sample, of size  $n = 36$ . To the right is a dot histogram of means for 300 independent replications, and the theoretical sampling distribution for replication means. The light horizontal lines and shading indicate the proportion of replication means that would fall within the original CI. In the left panel 85% would do so, whereas in the right the original mean happens to not capture  $\mu$ , so less than 50% of replication means would fall within this CI. The figure illustrates two sources of variability: variation of replication means about  $\mu$ , and variation of original means about  $\mu$ . These illustrations are of simulations, in which  $\mu$  and  $\sigma$  are known. In practice, of course, and in our experimental tasks,  $\mu$  and  $\sigma$  are not known, and only the original mean and its CI can be displayed.

*Figure 2.* The percentage of replication means that will, on average, fall within a CI with confidence level  $C$ . Samples are assumed to be large. Numerical values are shown for 99%, 95%, and 90% CIs, and for SE bars, which for large samples correspond to 68.3% CIs.

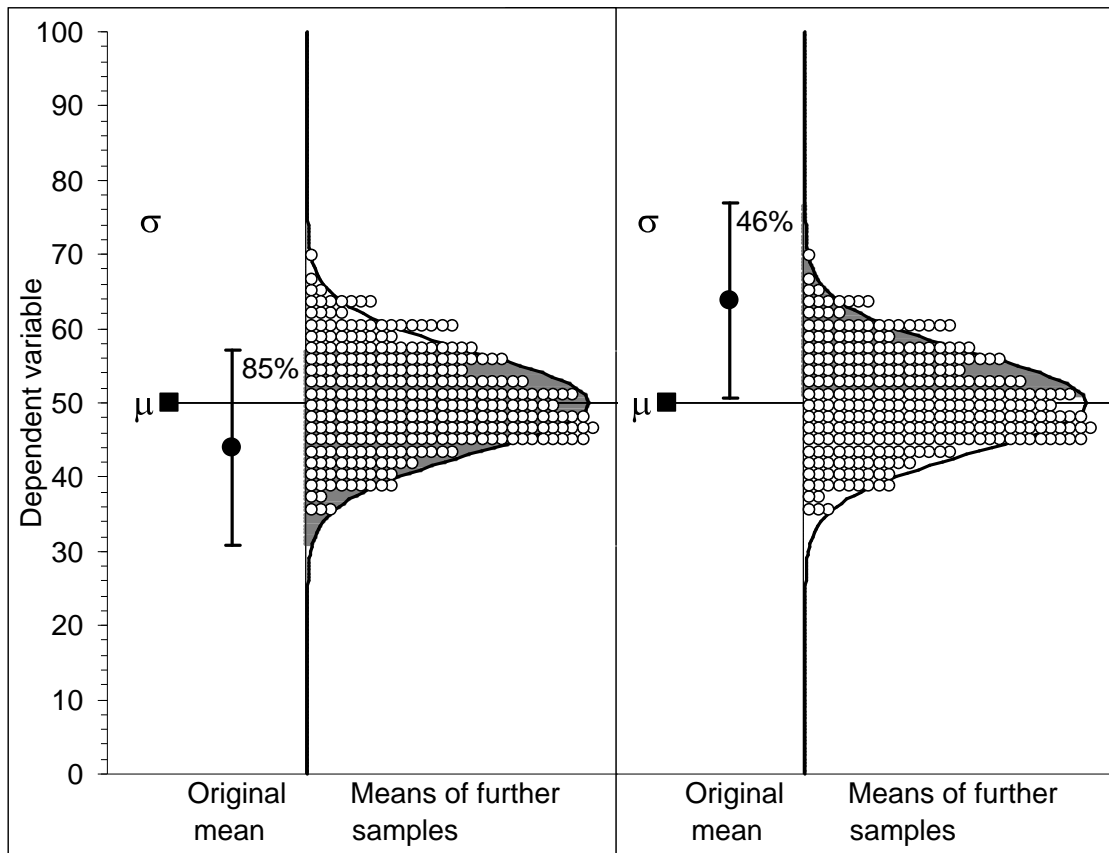
*Figure 3.* On entering an experimental website a participant saw instructions, and the figure at left but without the horizontal lines. Repeated clicking within the figure caused a single set of up to 10 horizontal lines to appear, at the vertical positions of the clicks. The participant clicked to form a plausible set of replication means. Three example sets are shown: Sets **a** and **c** have property

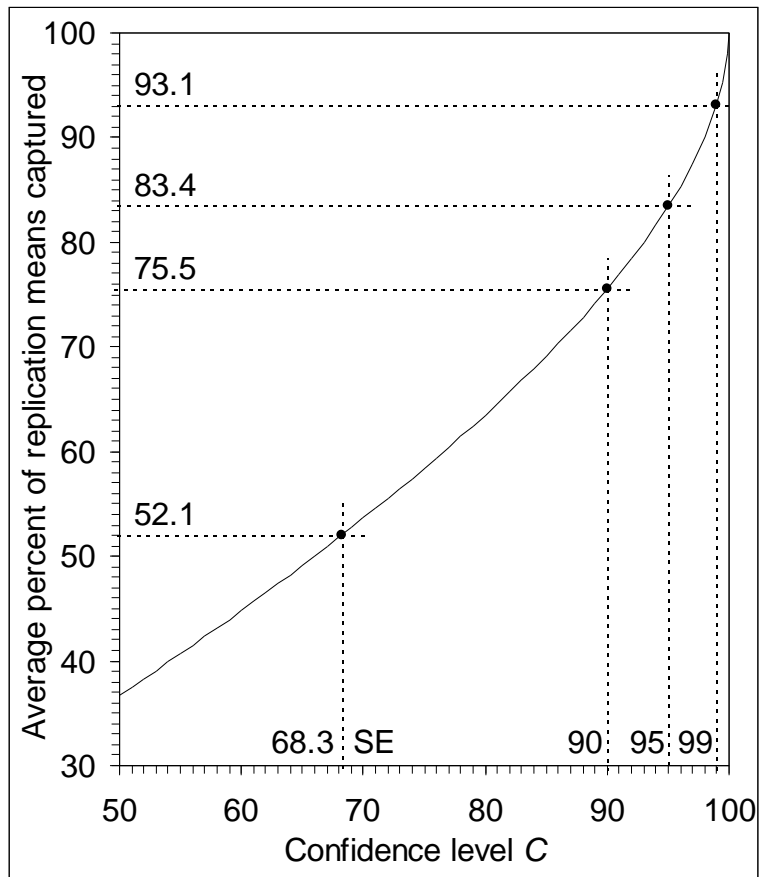
1 (see text) by having set mean  $M = 750$ ; sets **a** and **b** have property 2 by having  $SD_M = 153$ ; and **b** and **c** have property 3 by having  $SD_O = 216$ .

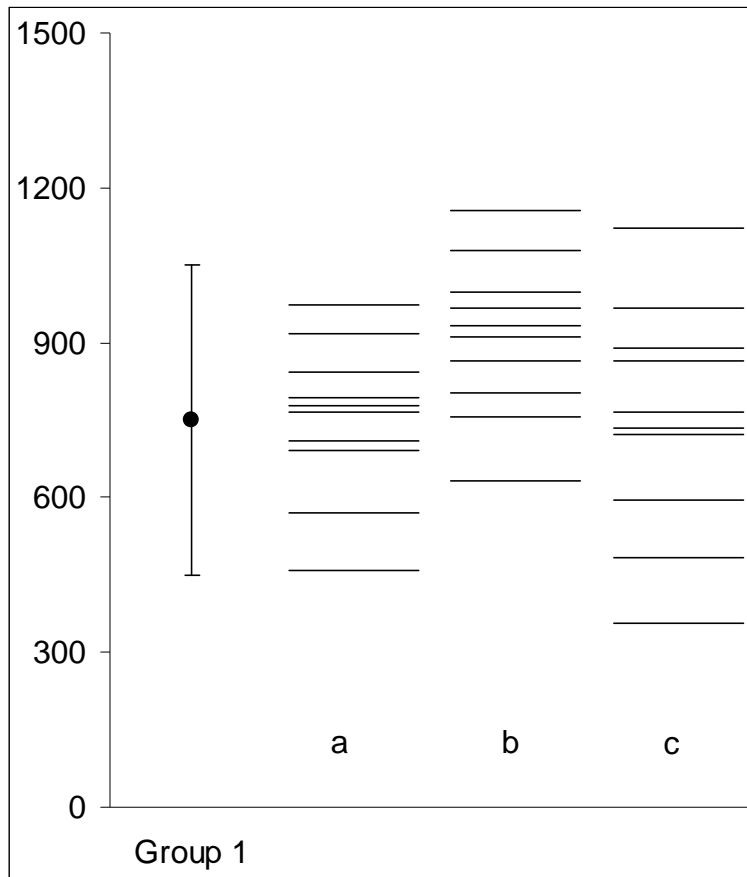
*Figure 4.* Small open circles are replication set means,  $M$ , for individual participants. Large filled circles are means of  $M$  for different groups. Error bars are 95% CIs, based on inter-participant variation within a group. The numbers of participants in each group ( $N$ ) is also shown. The dotted line marked with an open circle is at 750, the original mean, and those marked with open diamonds are the expected average values of  $M$  if properties 2 and 3 (see text) are both to hold, as in example set **b** in Figure 3. Note the vertical scales is different from that in Figure 1.

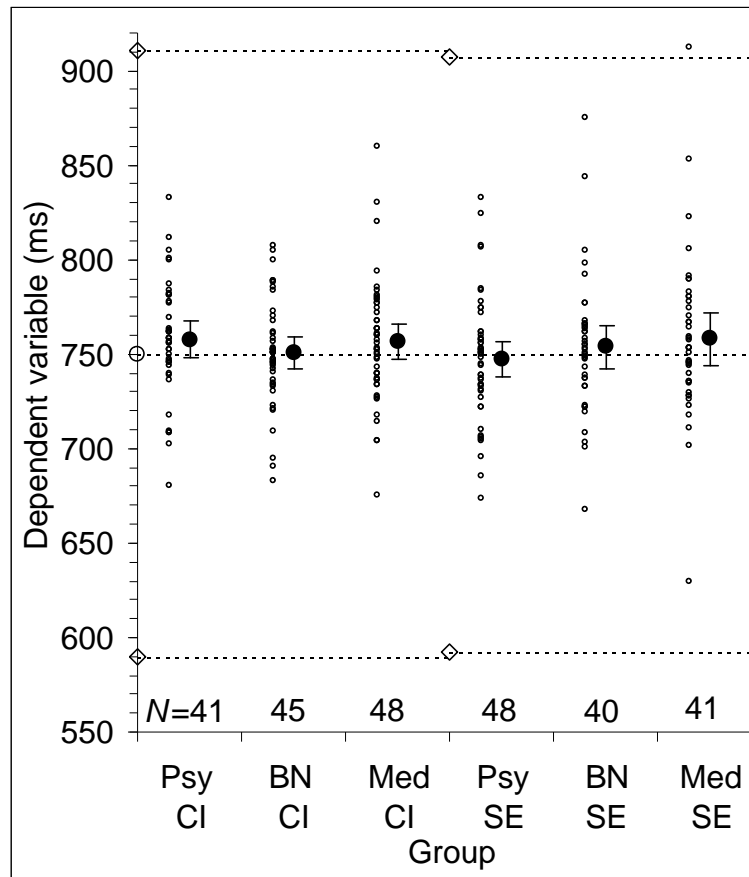
*Figure 5.* Mean of  $SD_M$  (filled squares) and  $SD_O$  (filled triangles), for each group.  $SD_M$  measures variation of replication means within a set about their own mean  $M$ , and  $SD_O$  measures their variation about the original mean 750. Error bars are 95% CIs, based on inter-participant variation within a group. The dotted lines marked with open squares indicate the average values expected for  $SD_M$  under property 2 (see text), and with open triangles those for  $SD_O$  under property 3 (see text).

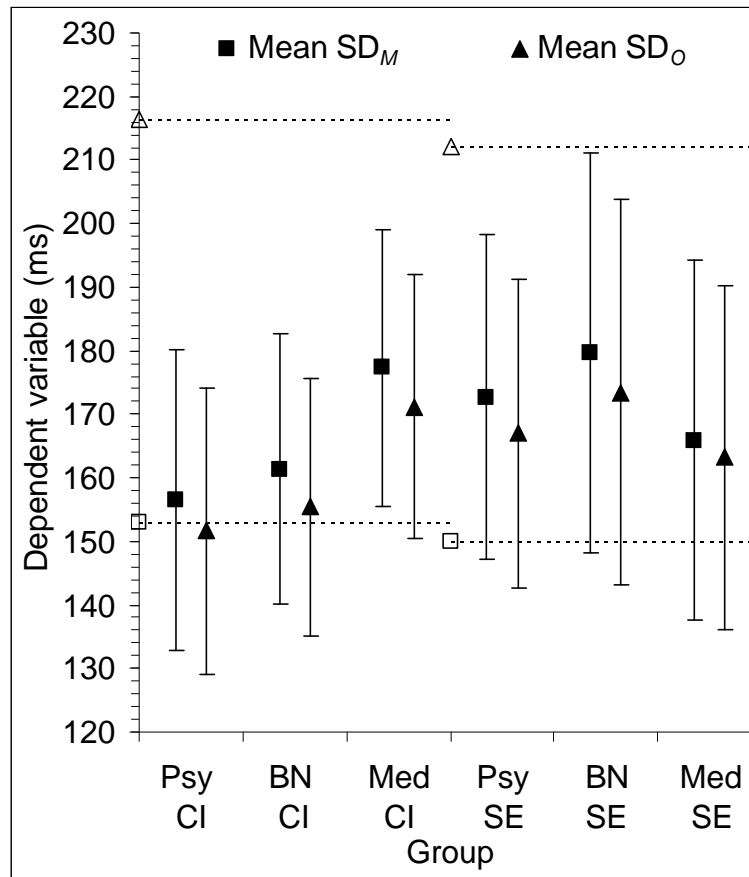
*Figure 6.* Frequencies with which participants in the 6 groups placed various numbers of replication means (maximum 10) within the CI or SE bars shown on the original mean.

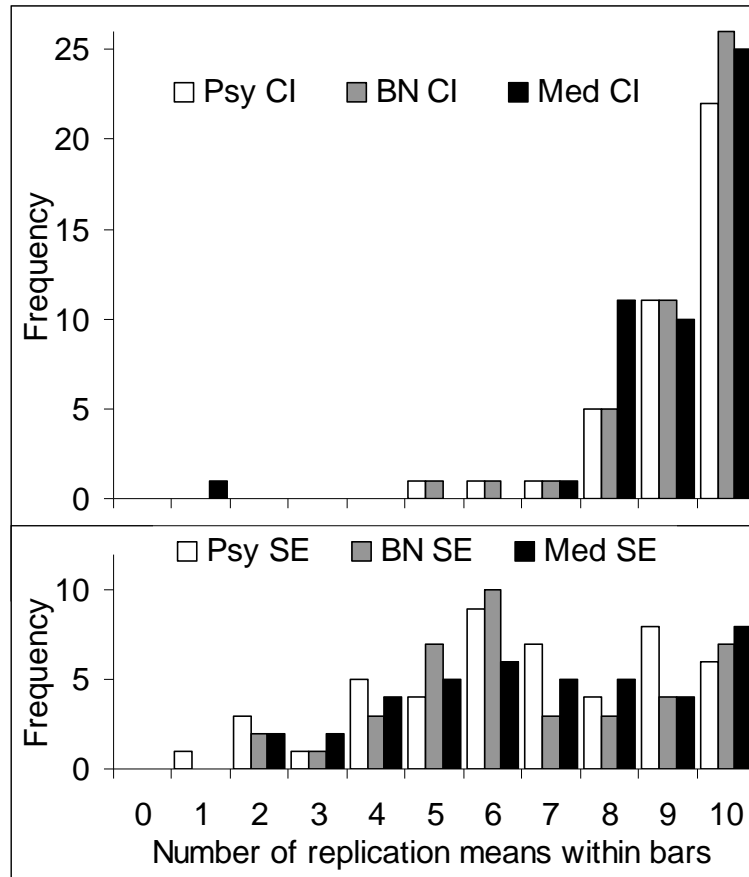












This document was created with Win2PDF available at <http://www.daneprairie.com>.  
The unregistered version of Win2PDF is for evaluation or non-commercial use only.