® INTERVAL ESTIMATES FOR STATISTICAL COMMUNICATION: PROBLEMS AND POSSIBLE SOLUTIONS

CUMMING, Geoff
La Trobe University
Australia

FIDLER, Fiona
La Trobe University and University of Melbourne
Australia

In many behavioural and social sciences reformers are urging wider use of interval estimates. We believe confidence intervals can improve research communication markedly, but several problems are raised by our empirical studies of how people understand and misunderstand intervals. We describe three of these problems: an incorrect belief about confidence interval overlap and its relation to statistical significance; failure to distinguish between confidence intervals and standard error bars; and finally, neglect of the importance of research design in applying and interpreting intervals. Our suggested solution is better guidelines, or 'rules of eye', and improved graphical presentations to assist with confidence interval presentation and interpretation. The rules of eye are also pedagogic tools, for teaching deeper understanding of interval estimates. By confronting existing misconceptions, these guidelines should facilitate conceptual change in thinking not only about interval estimates themselves, but also the often misunderstood concept of statistical significance.

CONFIDENCE INTERVALS: ADVANTAGES AND MISCONCEPTIONS

Interval estimates (such as confidence intervals, CIs) have several advantages over traditional null hypothesis significance tests (NHST) for statistical communication. They focus attention on effect size, and interval width offers a guide to precision. Because of the focus on effect size, they have the potential to facilitate meta-analytic thinking, or thinking 'across studies' (Cumming & Finch, 2001). These attributes are vital in the dissemination of statistical data, and in choosing the representations that best communicate research findings. Interval estimates may also have pedagogic advantages when used to teach null hypothesis testing. They can be used to directly confront misconceptions typically associated with NHST, facilitating conceptual change.

Yet in many social and life sciences, including psychology, CIs and other interval estimates are rarely used. Because NHST has dominated these sciences for around half a century, methods for calculating CIs (in some cases) and guidelines for their interpretation are relatively underdeveloped. Perhaps consequently, interval estimates are sometimes also misinterpreted by students and researchers. Misconceptions associated with NHST are well documented and studied (e.g. Tversky & Kahneman, 1971; Oakes, 1986; Schmidt & Hunter, 1997). The same cannot be said of interval estimates. Here we present a series of statistical cognition studies that focus on researchers' understanding of interval estimates. The results highlight problems in how interval estimates are interpreted and therefore point to the need for improved statistical education, and better guidelines for researchers. Our general approach has three steps.

First, we conduct statistical cognition experiments to investigate what misconceptions researchers may have, and to describe any misconception we find. We do this across three disciplines - Psychology (Psy), Medicine (Med) and Behavioural Neuroscience (BN)- because these three disciplinary communities of researchers have very different customs for use of interval estimates (Cumming, Williams & Fidler, 2004). In medicine, CIs are routinely reported, but appear as text or in tables. In behavioural neuroscience, standard errors are often displayed in figures. In psychology, both are rare—whether in text, tables or figures.

Second, we develop graphical representations, and guidelines (or 'rules of eye') that may help researchers to confront, understand, and overcome the misconception. Finally, we consider how best to use these guidelines in teaching, so that students can avoid developing the misconceptions, or overcome them if already established. Directly confronting misconceptions is an integral part of conceptual change (e.g. White, 1993). It requires engaging with the concepts in a meaningful context, and providing students with vivid and memorable representations that support correct understanding. For us, this involves building interactive graphical simulations so students can enter their own data and display these in various ways. The software runs under Microsoft Excel and is called ESCI ("ess-key", Exploratory Software for Confidence Intervals).

95% CONFIDENCE INTERVALS ON TWO INDEPENDENT MEANS

Schenker and Gentleman (2001) pointed out a widely believed, but incorrect rule, often used for interpreting interval estimates in medical and health science literature. The incorrect rule is that 'just touching' 95% CIs (i.e., CIs that just do not overlap) are equivalent to a statistically significant difference (at $p<.05$) between point estimates. In fact, 95% CIs on independent groups means will overlap by approximately a quarter of the total CI length when $p=.05$ (Cumming & Finch, 2005).

Belia, Fidler and Cumming (2004) investigated the extent of this misconception. We emailed researchers published in leading journals in medicine, psychology and behavioural neuroscience. Researchers replied by following a link to one of our websites, where they saw a display of two independent groups means with error bars. 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-tail). A participant's response was the position of the adjustable Group 2 mean.

Responses were very widely spread, and 34% of respondents set CIs to just touch. Participants tended to set the means too far apart, not realizing that the .05 statistical significance borderline requires overlap. Their mean response corresponds to $p = .009$ (Psy .017, BN .008, Med .006) rather than the target .05. The distributions of responses in each discipline were similar, despite the different statistical reporting cultures.

Figure 1 shows CIs on two independent means, demonstrating the extent of overlap when the means are just statistically significantly different (p is approximately .05). For Group 1, $n_1=32$, $M_1=83.0$, $w_1=28.1$, where w_1 is the margin of error, that is the half-width of the 95% CI. For Group 2, $n_2=35$, $M_2=40.6$, $w_2=31.3$. The proportion overlap, which is the distance between the dashed lines, is expressed as a proportion of the average margin of error, that is, the average of w_1 and w_2. Proportion overlap is .57, and the two-tailed p value, for the independent groups t test that compares the two means, is .046. Figure 1 thus shows the configuration of means and CIs that is close to the border of statistical significance, for the traditional .05 level.

Cumming and Finch (2005) expressed this as a rule of eye, useful for interpreting research findings, and in teaching:

> ***Rule of eye (Two means with 95% CIs)*** *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 .5 or less.*
>
> *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.*
>
> *These relationships are sufficiently accurate when both samples sizes are at least 10, and when the margins of error do not differ by more than a factor of 2.*

Cumming and Finch (2005) investigated how the p value varies for a wide variety of sample sizes, and CI lengths. They used Welch-Satterthwaite methods, and so did not need to assume homogeneity of variance. Their conclusion is that the stated rule holds reasonably, or is a little conservative, under the surprisingly general conditions stated.

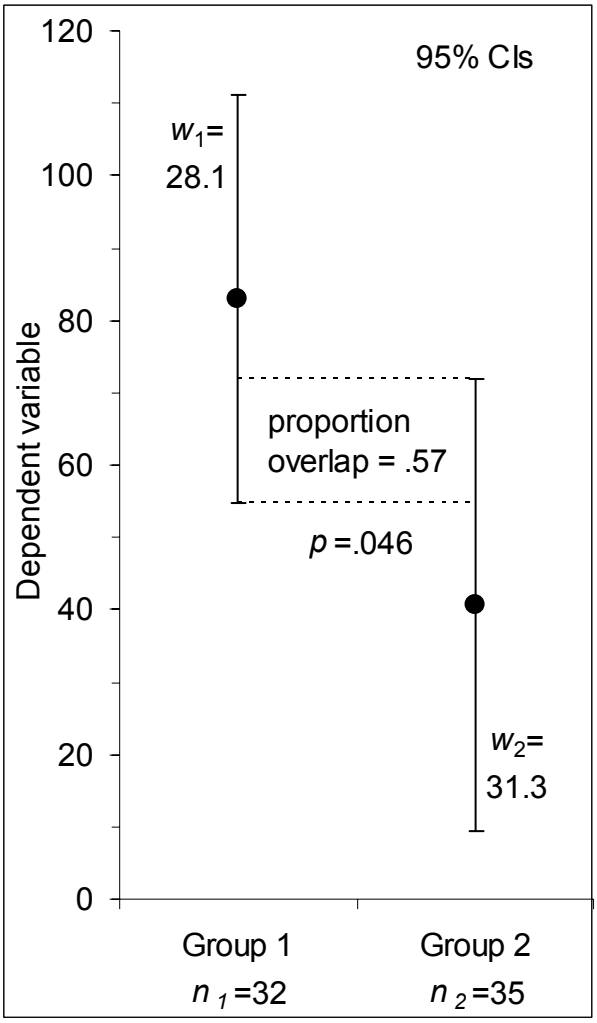

Figure 1. An example with two independent groups (Group 1 and Group 2), showing the sample means (filled dots) and 95% CIs. The margin of error (w_1, w_2) is half the total length of the CI. The proportion overlap is the distance between the two dashed lines, expressed as a proportion of the average of w_1 and w_2, and here is .57. The two-tailed p value for the difference between the two means is .046, so the configuration illustrates the rule of eye for two 95% CIs.

CONFIDENCE INTERVALS AND STANDARD ERRORS ARE NOT SUFFICIENTLY DISTINGUISHED

In an alarming number of cases, authors do not state what intervals or error bars they have used (e.g. standard deviation, standard error, CI). For example, in a recent unpublished study of 10 leading psychology journals, we found that 31% of articles that include a figure with error bars failed to identify the error bars in figure captions—despite the APA *Publication Manual* (2001) requiring such identification.

In the Belia et al. (2004) study, there was little difference in the distributions of responses, for just statistically significant differences, for 95% CIs, and SE bars. The (incorrect) rule that intervals should just touch, was used about as often with SE bars (30%) as with 95% CIs (34%)! Of course, the rule is even more inaccurate when applied to SE bars. 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 the population mean μ.

Figure 2 shows the same example as Figure 1, but with SE bars. Here w_1=13.8 is the SE for Group 1, and w_2=15.4. The gap is the distance between the dashed lines, expressed as

a proportion of the average SE, that is, the average of w_1 and w_2. The proportion gap is .91, and the two-tailed p value is of course the same as before (.046). Figure 2 thus shows the configuration of means and SE bars that is close to the border of statistical significance at the .05 level. Cumming and Finch (2005) expressed this as a rule of eye.

> ***Rule of eye (Two means with SE bars)*** *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.*
> *In addition,* $p \leq .01$ *when the proportion gap is about 2 or more.*
> *These relationships are sufficiently accurate when both samples sizes are at least 10, and when the SEs of the two groups do not differ by more than a factor of 2.*

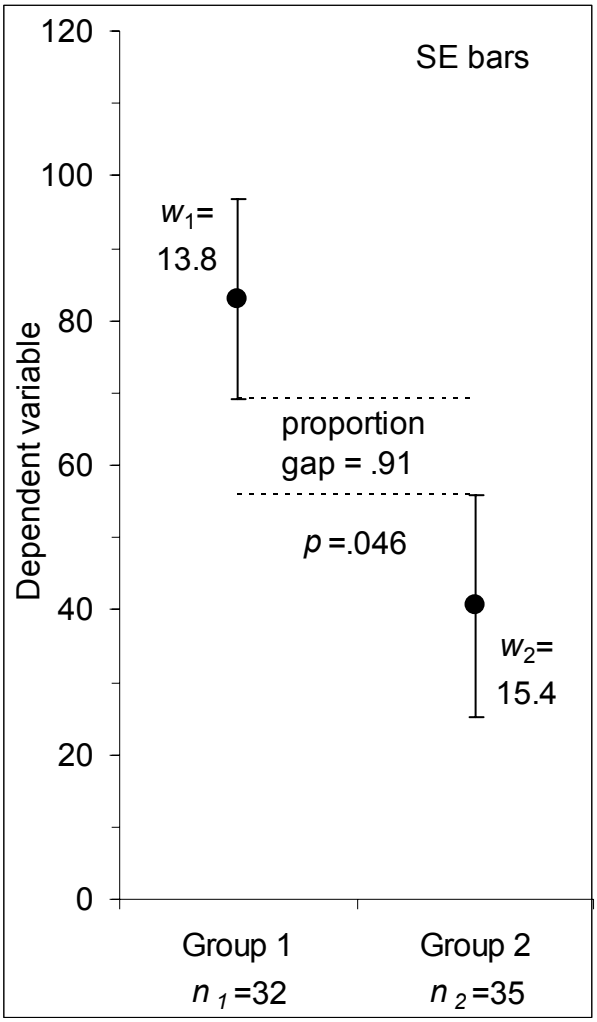

Figure 2. The same example as shown in Figure 1, but showing the sample means and SE bars. The proportion gap is the distance between the two dashed lines, expressed as a proportion of the average of w_1 and w_2, and here is .91. The two-tailed p value is unchanged at .046, so the configuration illustrates the rule of eye for SE bars on two independent means.

IMPORTANCE OF EXPERIMENTAL DESIGN: INDEPENDENT MEANS, OR REPEATED MEASURE?

In the Belia et al. (2004) results, the vital aspect of experimental design was often overlooked. The two rules of eye already mentioned, relating to overlap of CIs and SE bars, apply only to independent groups. For repeated measure designs, interval estimates around individual means provide no direct information about the statistical significance of the difference. Only 11% of participants demonstrated recognition of this, and indicated that the task could therefore not be completed.

Figures 3 and 4 are part screen images from ESCI pages that are designed to distinguish independent groups and paired data. Figure 3 shows an independent-groups example in which proportion overlap is .40, which indicates that $p<.05$ for the difference between the two independent means. Values are displayed for descriptive statistics for the two groups, and for the 95% CIs. The value of overlap is shown, but the exact p value is not displayed, because the introductory course for which this software was designed discussed inference only in terms of pictures of CIs, with no mention of statistical significance or p values. An adapted rule of eye was used in which proportion overlap of about .5 or less is taken as 'reasonable evidence that there is a difference between the two population means', and proportion overlap of 0, or a gap, is taken as 'quite strong evidence' of a true difference.

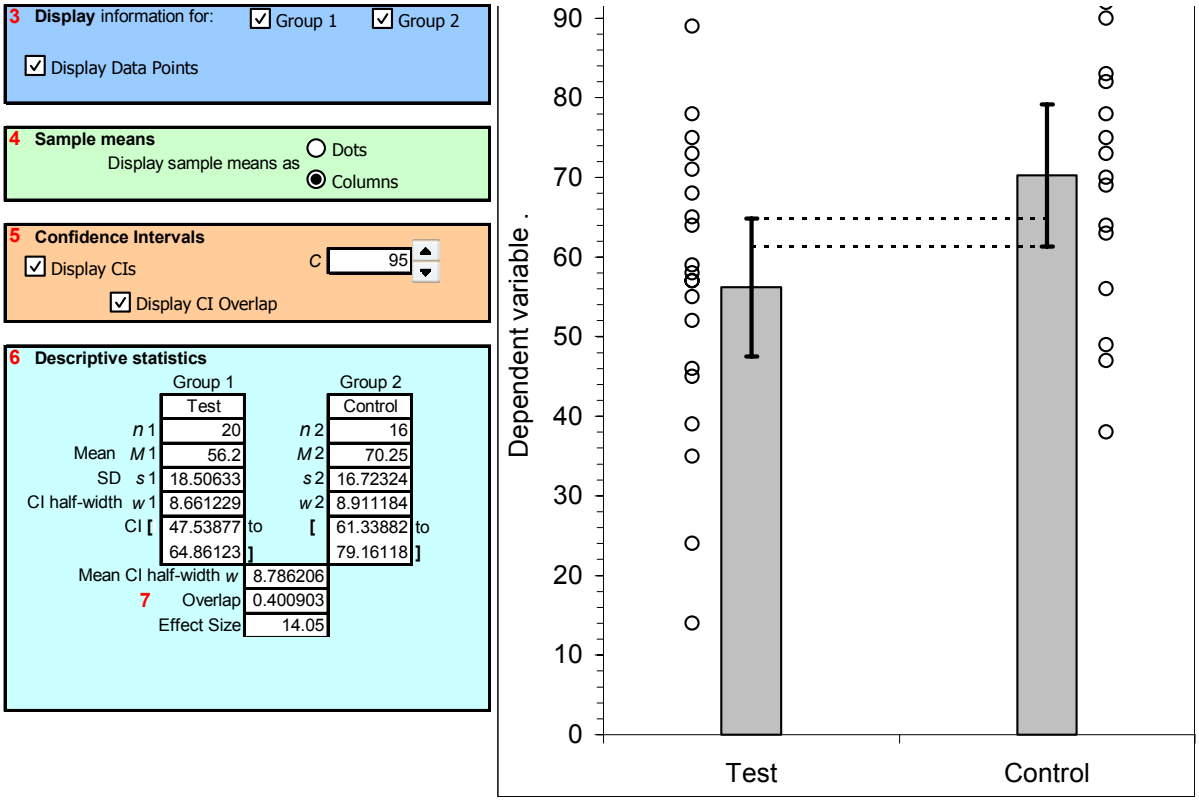

Figure 3. Part screen image of an ESCI page that allows entry of your own data for two independent groups, then display of data points (small circles), means (columns), CIs, and overlap (dashed lines). Numerical values are displayed also. Overlap here is .40, so $p<.05$, although this software is used in an introductory course in which there is no mention of statistical significance or p values, and so there is no display of the exact p value.

A paired-data example is shown in Figure 4. CIs on the pretest and posttest are wide, and have high proportion overlap, but these intervals, and this overlap, are *irrelevant* for assessing the difference between the means. The difference is plotted on a floating differences axis, and the CI on the difference does not include zero, indicating that $p<.05$ for the difference. The CI on the difference is so short because the pretest and posttest scores are highly positively correlated. Because statistical significance and p values are not mentioned in this introductory course, the difference is assessed simply by interpreting the relevant CI, which is the CI on the difference. Any value outside the CI is relatively implausible as the true value of the population parameter. Therefore, noting that zero lies outside the CI justifies a conclusion that there is reasonable evidence that there is an increase from pretest to posttest in the population.

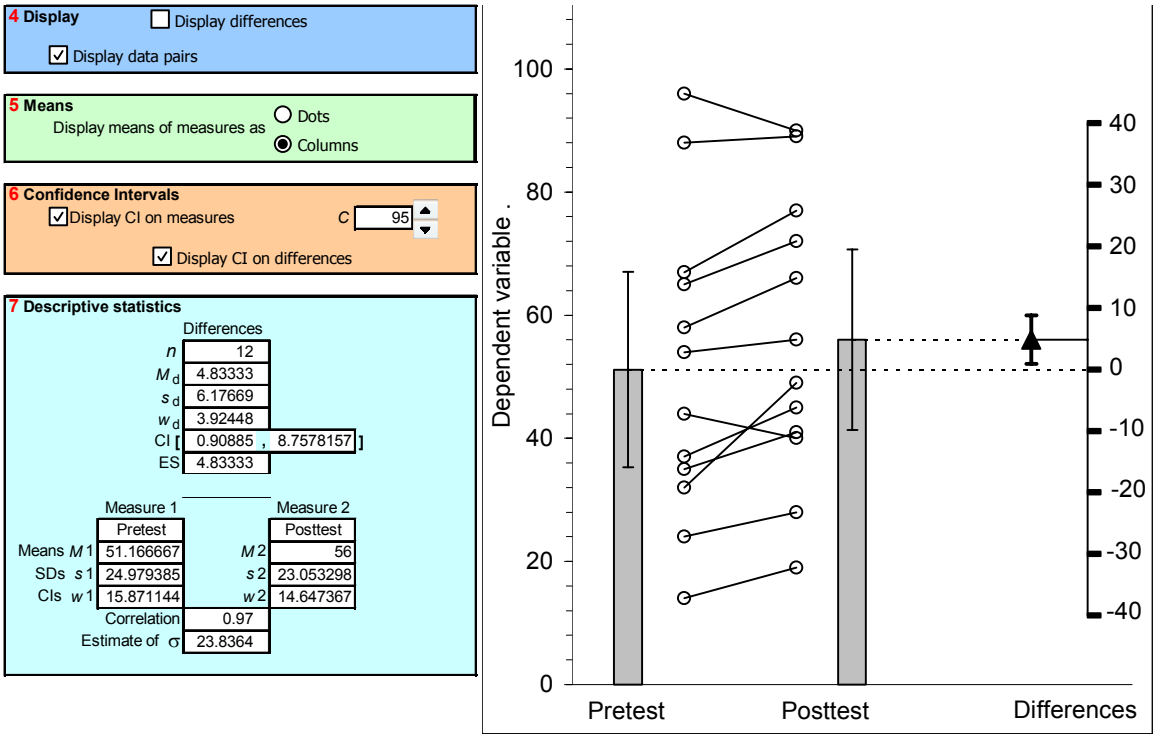

Figure 4. Part screen image of an ESCI page that allows entry of your own data for paired data, then display of pairs of data points (small circles), means (columns), CIs, and the difference (filled triangle) between the means, on a floating differences axis. Numerical values are displayed also. The CIs on the Pretest and Posttest are *irrelevant* for assessing the difference between the two means. The CI on the difference is short, because of the correlation between the two measures, and does not include zero, indicating that the *p* value for the difference is <.05.

The design of the graphs in Figures 3 and 4 were developed to make the independent groups and paired designs as distinctive as possible. Note that conventional graphical practice is to make no distinction: A simple figure of two means with two CIs may represent either design. All too often even the figure caption does not make clear what the design is.

Data points for the two groups are displayed individually in Figure 3, but must be connected by lines in Figure 4 to make the pairing salient. In addition the differences axis in Figure 4, and the CI on the difference, makes clear the impact of the correlation between measures in the paired design.

EDUCATIONAL IMPLICATIONS

Calls for statistical reform of social and life sciences continue to grow. Along with many others, the APA Task Force on Statistical Inference (TFSI) has called for increased use of interval estimates, particularly CIs (Wilkinson et al., 1999). Statistics curricula in these disciplines must now respond to these calls. Part of the challenge of reform is to instigate conceptual change (Thomason, Cumming & Zangari, 1994), with guidance from cognitive research that identifies the misconceptions commonly held by students and researchers. Our conceptual change strategy is to provide multiple representations, dynamically linked on screen, so a student or other user can see the diagram, and also the data and relevant descriptive and inferential statistics, and to see how these different representations are linked. This allows them to see how the displays change if some aspect of the data is changed.

We are currently using the ESCI software tools, from which Figures 3 and 4 are derived, in a course for first year psychology students in which during the first semester there is no mention of statistical significance or *p* values, but inference is based on pictures that include CIs, and these pictures are interpreted using rules of eye (of which a selection have been given here). Our experience with this course is that students respond well and can think

about important issues of design and inferential interpretation, without the traditional complexities of statistical significance techniques.

There is now the need to assess the effectiveness of these rules of eye, and displays, in helping students avoid, or overcome, the misconceptions held by most researchers. If they are effective with students, then presumably they can also be used effectively as the basis for professional development to assist researchers overcome their misconceptions. Ideally, they should be able to contribute to statistical education for researchers across several disciplines, and that may be one of the biggest challenges for achieving improved statistical communication!

REFERENCES

American Psychological Association (2001). *Publication Manual of the American Psychological Association (5th edn.).* 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 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., Cumming, G., Wilson, S. et al. (2004). Statistical reporting practices in psychology (1998-2004): Responses to the APA Task Force on Statistical Inference. *Manuscript in preparation*.

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

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., & Hunter, J.E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L.L. Harlow, S.A. Mulaik, and J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.

Thomason, N., Cumming, G., & Zangari, M. (1994). Understanding central concepts of statistics and experimental design in the social sciences. In K. Beattie, C. McNaught, and S. Wills (Eds.), *Interactive multimedia in university education: Designing for change in teaching & learning* (pp. 59-81). Amsterdam: Elsevier.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 2,* 105-110.

White, B.Y. (1993). ThinkerTools: Causal models and science education. *Cognition & Instruction, 10*, 1-102.

Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals. *American Psychologist, 54,* 594-604.