THE STATISTICAL RE-EDUCATION OF PSYCHOLOGY ®

Geoff Cumming and Fiona Fidler
La Trobe University
Neil Thomason
University of Melbourne
Australia

Psychology remains addicted to null hypothesis significance testing despite decades of effort by reformers. Extensive changes in statistical understanding and practices are needed. The authors propose a model of reform—the statistical re-education of psychology—by making an analogy with the conceptual change model of learning. Four diverse components of reform are identified, and illustrated by brief examples of research. Reform is especially challenging because many statistics teachers in psychology first need to achieve conceptual change themselves. In relation to a highly desirable increase in use of confidence intervals (CIs), it seems that many psychologists do not understand CIs well, and guidelines for CI use are lacking. The conceptual change model is offered to guide research needed on many aspects of reform, and the important and exciting task of the statistical re-education of psychology.

PSYCHOLOGY MUST MOVE BEYOND FLAWED NHST PRACTICES

Psychology is addicted to null hypothesis significance testing (NHST), despite decades of criticism of the technique, and evidence that NHST is widely misunderstood and has caused great damage. Finch, Thomason, and Cumming (in press) and Nickerson (2000) gave reviews. Important changes may now at last be possible, with the American Psychological Association (APA) giving, via its Taskforce on Statistical Inference (TFSI; Wilkinson & TFSI, 1999) and widely-used *Publication Manual* (APA, 2001), limited official encouragement of reform. The APA position is, however, equivocal, and successful reform of psychologists' inferential practices is far from assured.

Extensive changes to psychologists' practices are needed (Wilkinson & TFSI, 1999), including wider use of exploratory data analysis, effect size measures, meta-analysis, graphical presentations, and a major shift from NHST to interval estimation. We will mainly refer to an increased use of confidence intervals (CIs), especially in graphical form as error bars. Despite decades of advocacy there has been little discussion of how reform might be achieved, although some appreciate that the task is challenging (e.g., Thompson, 2001). Most attention has been paid to the influence of the APA *Publication Manual* and to the power that journal editors possess to state requirements for data analysis.

A MODEL FOR STATISTICAL REFORM IN PSYCHOLOGY

We believe there are many obstacles to reform. For example, Cumming and Finch (2001b) argued that psychologists do not have well-developed practices for CI use and may not understand CIs well, and that there is little research on CI interpretation. The new *Publication Manual* (APA, 2001) advocates CIs, but gives not a single example of CI use and no advice on style for reporting CIs—but continues to give detailed guidance on how to report NHST. Psychology thus appears to be in a strange situation: CI use is being advocated, yet we have little idea how CIs should best be used. For an empirical scientific discipline this seems to be an extraordinary situation. (Cumming and Finch, 2001b, argued also that other disciplines, including medicine in which CIs are now widely used, have little to offer on these issues.).

There are two reasons statistics education is involved. First, reform that achieves lasting success will require drastic revision of statistics courses in psychology. New textbooks and software are needed. Second, we propose a model for statistical reform by making an analogy with the *conceptual change* model of learning (e.g., West & Pines, 1985). We believe our model to be more comprehensive than previous descriptions of reform, but such complexity is required if reform is to be understood. This analogy justifies referring to reform as *the statistical re-education of psychology*. There are surely few greater challenges in statistics education than the re-education of an entire discipline, especially when current practices are, as in psychology, deeply entrenched and poorly understood, and desired new practices are in some cases

underdeveloped. We sketch a conceptual change model of learning, then use the model's components as subheadings to describe corresponding aspects of our view of successful statistical reform in psychology. Under each we give some brief examples from our own research.

A conceptual change model of learning

Imagine a ball flying around on the end of a string. If the string breaks will the ball continue on a path that, viewed from above, is curved? Many people incorrectly answer yes: Their naïve belief that 'turning' continues in the absence of an external force is entrenched and often persists despite physics education. A conceptual change approach views learning not as mere acquisition of new concepts, but as a journey with a start point as well as a destination. Initial erroneous concepts must be confronted and overcome, no matter how deeply-held, then replaced by the correct concepts. The learner must be actively engaged, and must construct for him or herself the new correct understandings. It may be extremely difficult to overcome initial naïve concepts and to acquire the correct concepts so thoroughly that they are owned by the learner and applied intuitively in appropriate situations. A wide range of well-designed learning activities, requiring much initiative and reflection by the learner, may be needed to achieve conceptual change. Working with computer simulations may be especially valuable (White, 1993). Thomason, Cumming and Zangari (1994) used the term *naïve statistics* and adopted such a conceptual change approach for the design rationale of StatPlay, which is interactive multimedia for learning of some basic statistical concepts, and which has proved effective.

The model applied to psychology

We propose that the conceptual change model is appropriate also for the statistical re-education of psychology. Consider: Current statistical practices are deeply entrenched, yet have many flaws; a successful reform outcome requires new attitudes and understandings as well as adoption of new practices; the reform road is proving difficult; and multiple strategies are likely to be necessary. An outline of a typical conceptual change model may include:

1. A description of the learner's initial concepts, practices and attitudes.
2. A description of the learner's initial attitudes towards learning and the goals of learning.
3. A good understanding by the teacher of the desired final concepts and practices.
4. Knowledge, preferably based on research, of how the learner can be supported to make an efficient and successful transition from the initial to the desired state.

Replace 'learner' by 'psychologists', and 'teacher' by 'advocates of reform', and we have four broad components of what reform requires. Cumming and Finch (2001b) argued that, with respect at least to the crucial topic of CIs, psychology currently has good knowledge on *none* of the four. Considerable research is required on all four, and the discipline of statistics education should be able to give invaluable guidance.

However, our analogy underestimates psychology's reform task. Conceptual change models envisage students with naïve beliefs working with teachers who know the target concepts accurately, understand the conceptual changes their students need, and are dedicated to help. In contrast, advocates of psychology reform are motivated and persistent but—as the absence of relevant research results attests—do not themselves understand in full detail exactly what the reform outcomes should be. In addition, much of the reform work must be carried out by large numbers of other psychologists, including statistics teachers, textbook authors, and journal editors and their reviewers. Many of these large groups do not appreciate the necessity for reform, and need to achieve considerable conceptual change themselves before they can work effectively to support, respectively, their students and manuscript authors to reach reformed understandings.

In other words, conceptual change models typically envisage a one-step process: A teacher with good understanding helps students make the learning journey. Psychology, however, may need a two- or more-step process of conceptual change because many who are in teaching or leadership roles have not yet themselves made the journey. In addition, not even the reformers have mapped all the steps of this reform journey. Our conceptual change model may need to be invoked several times to describe fully the statistical re-education of psychology. The four

subsections below examine briefly the four components of reform we identified above. Some of our comments are general and some relate mainly to CIs.

COMPONENTS OF THE STATISTICAL RE-EDUCATION OF PSYCHOLOGY

Psychologists' current practices, and use and understanding of CIs

Our focus on these issues echoes the identification by Batanero (2001) of "Errors and attitudes in the use of statistics by researchers" (p. 392) as a research priority for statistics education. We mention a number of approaches. First, Finch, Cumming and Thomason (2001; henceforth 'the JAP study') analysed 150 papers from the high-status *Journal of Applied Psychology*, 1940 to 1999. They found NHST dominated and CIs were virtually never used. Research methodology in JAP has increased greatly in sophistication over 60 years, but inference practices have shown remarkable and depressing stability. There is little sign that decades of critiques by reformers had by 1999 led to any changes in statistical practices in JAP.

'The Loftus study' (Finch, Cumming, Williams, et al, 2001) will be described below. Here we mention that across a range of psychology journals we found an increase over the last decade from 9% to 22% in empirical papers that reported at least one CI as numerical values, or included a figure with CI or standard error (SE) bars. Such desirable reform practices were infrequent but have shown a modest increase. However, the conclusions in almost all these papers relied on NHST; even when CIs were reported they were rarely used for interpretation.

In two current studies we are investigating researchers' understanding of CIs shown as error bars in figures. Belia, Fidler and Cumming (2001) and Williams, Fidler and Cumming (2001) are inviting via email the authors of recent papers in leading journals to visit a website and complete a quick judgement task. This is proving to be a practical way to tap aspects of the statistical understanding of researchers, and it may prove to be a technique with great potential for research on statistical cognition. We also ask questions about the respondent's current practices and opinions on reform. Some respondents are providing interesting comments.

We sketch here the Belia et al (2001) study. A respondent visiting the site sees a figure somewhat like the left or right panel of Figure 1, but without the dashed horizontal lines. The heavy dots are the means of two independent groups, sizes 36 and 34, each shown with its 95% CI. By clicking the mouse the respondent can move the right hand mean, with its CI, up or down. The task is to position that mean so the two means are just significantly different, by two-tailed t test, $\alpha = .05$. Eyeball estimation is requested, not calculation. (The correct answer, surprisingly to many people, is approximately as shown in the left panel of Figure 1.)

Respondents are randomly allocated to one of five sites, so we can compare, for example, CI with SE bars, and dot with column representation of means. We are studying three groups of researchers, defined by the journals in which they published. The groups are psychology, which uses few error bars; neuroscience, which routinely uses SE error bars; and medicine, which routinely reports CIs but usually in tables and not as error bars in figures.

Responses are revealing very severe misconceptions, and interesting cross-discipline comparisons are emerging. A striking, even scandalous result is that a large proportion of these well-published researchers seem not to appreciate in the repeated measures case that the error bars on the separate pretest and posttest means are virtually irrelevant to the significance of the difference. They respond as if the means are of independent groups, without noting that the variability of the pre-post differences was not provided. This is an error that statistics teachers hope students in their introductory courses will avoid. If confirmed this is a dramatic finding: A large proportion of researchers may make elementary blunders in making inferences from some of the simplest figures with error bars. If use of figures with error bars is to increase, as we believe it should, then research is required to find ways to ensure more accurate inferential interpretations of such figures. Some combination of researcher re-education and improved presentation of figures is required.

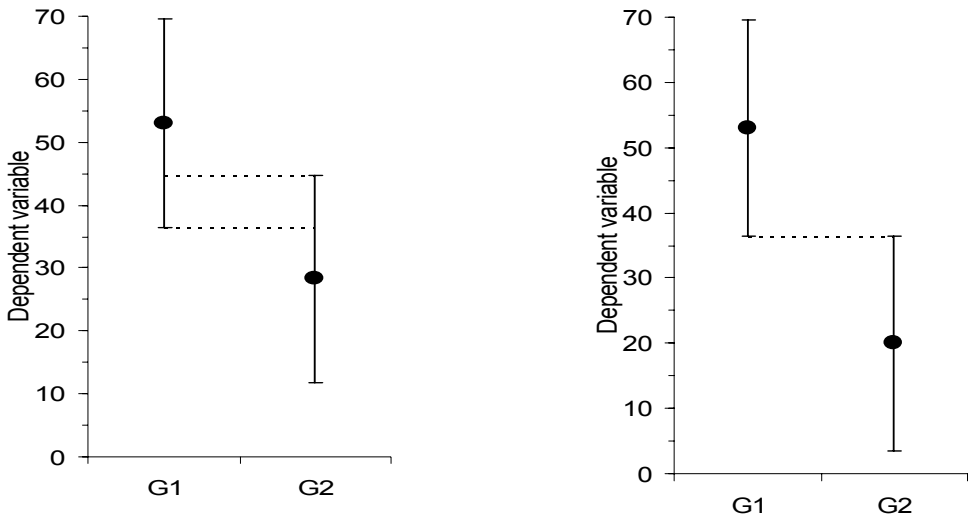

Figure 1. Two configurations of 95% CIs attached to sample means.
The dashed horizontal lines assist estimation of CI overlap, which is zero in the right panel. In the left panel it is .5, meaning the overlap is half of bar width *w*, that being half the full CI width. If G1 and G2 are two independent groups, of similar sizes that are not small, then the two-tail *p* values for the two-sample *t* test between the two means are approximately .04 for the left panel and .006 for the right. Thus the left panel is close to the $\alpha = .05$ significance border, and the right to the $\alpha = .01$ border.

Psychologists' attitudes to reform

In the Loftus study 59 *Memory & Cognition* authors responded to an email survey. They reported almost universal use of NHST, and only around 30% agreed that SEs or CIs were more informative than hypothesis tests. Thus most disagreed with the predominant reform view that in many situations CIs are to be preferred to hypothesis tests. However, a majority of respondents expressed some degree of support for the Loftus reforms.

Fidler et al. (2001) studied papers in the *Journal of Consulting and Clinical Psychology* ('the JCCP study'), and 62 authors responded to an email survey. An interesting discrepancy emerged: Just 30% of recent papers reported standardised effect sizes, but fully 77% of respondents agreed that "standardised effects sizes are appropriate to my research". Respondents may not have been an unbiased sample of authors, but the discrepancy occurred also for other reform items. It seems that at least some authors may be more reform-receptive than their published papers suggest.

How should CIs ideally be presented and used?

If an experimental design has more than one independent variable (IV), especially if there is a repeated measures IV, the display of CIs on means is problematic: A particular mean may have different CIs for the various main effects and comparisons of which it is part (Loftus & Masson, 1994). In the Loftus study we observed a variety of ways that, in such situations, authors of journal papers represented CIs. Some solutions were dreadful, some ingenious, but it is clear that there is no consistency of practice and a great lack of well-founded guidelines.

Cumming and Finch (2001b) analysed the problem and gave some interim recommendations, for example to use the graphics of Figure 1 rather than column graphs, and only to attach to a mean the CI calculated from the data contributing to that mean. Cumming and Finch emphasised that development of guidelines requires not mere opinion but study of how various graphical representations are understood, and identification of representations that best prompt easy, intuitive and correct interpretations. A research program is required.

What policies will give successful reform?

We mention two approaches to this crucial issue: first, assessing attempts by enterprising editors to implement reform in their journals and, second, preparing materials designed to assist researchers understand and use some reform practices. Geoffrey Loftus, editor of *Memory & Cognition* 1994-1997, strongly encouraged use of figures with error bars. In the Loftus study we examined 696 *Memory & Cognition* papers published before, during and after the Loftus editorship. Use of figures with bars increased to 47% under Loftus then declined. However, bars were rarely used for interpretation and NHST remained almost universal. Analysis of 415 papers in other psychology journals confirmed that Loftus' efforts had little influence beyond the papers he published. Loftus' experiment was valuable, but suggests that more than the efforts of any individual will be required.

In the JCCP study we examined papers published in 1988-2001, before and after Kendall's editorial encouragement to report effect sizes. Overall, standardised effect sizes were published in only 23% of possible cases, and there was little variation over years. Other reform practices also showed little or no effect of editorial encouragement. Although not specifically encouraged by Kendall, CI use increased to 24% in 2000-01, in line with a similar small increase noted in the Loftus study, but here again CIs were rarely used to support interpretation.

Turning to the second aspect of work on policies to support reform, we mention the Cumming and Finch (2001a) tutorial on CI use for simple designs. It emphasised the value of CIs for substantive interpretation and for promoting meta-analytic thinking—the combination of findings over studies. In addition, it described how CIs can be calculated for the standardised effect size measure, Cohen's δ or d, which requires use of noncentral t distributions.

Finally, Cumming and Finch (2001b) analysed the relation between CI (and SE) error bars and inference. Belia et al (2001) suggest many researchers use some overlap rule to guide response, but many of these rules are wrong or very conservative. Cumming and Finch, using the title *Inference by eye*, reviewed what little published analysis there is of overlap, and suggested seven *rules of eye* (by analogy with rules of thumb) to guide reading of CI and SE bars. Rule 5, for example, states that for two independent groups, under reasonable assumptions, overlap of the 95% CIs by .5 of w, the half-width of a CI, corresponds to $p < .05$ for the two-tailed t test comparing the means. This is the situation shown in the left panel of Figure 1. Zero overlap, as in the right panel, corresponds to $p < .01$. We hope development of such rules will help overcome the misconceptions found by Belia et al, and allow researchers to inspect a figure with error bars and carry out inference by eye easily and reasonably accurately.

The two CI papers (Cumming & Finch, 2001a, b) were illustrated by figures from interactive graphical simulations that run under Microsoft Excel, which we refer to as *live figures*. The live figure for Figure 1, for example, allows parameters to be changed by mouse click or drag. Numerical displays are given of p values and other values of interest. Users can explore, and use their own data. The live figure idea is described by Cumming (submitted). The set of live figures is intended to support the development of skills of inference by eye; it is termed ESCI ("ess-key"; Exploratory software for confidence intervals; information at: www.latrobe.edu.au/psy/esci).

CONCLUSIONS

The diverse findings mentioned above suggest that psychologists persist with NHST and there has been little reform progress. However, CI use may have increased a little, and there are hints some authors may be somewhat receptive to reform. If it is confirmed that many researchers have severe misconceptions about CIs and the inferences they permit, reform faces an enormous obstacle. Ways must be found to present CIs better and researchers need to be assisted to understand CIs accurately—only then can wider use of CIs be confidently promoted.

Case studies suggest editorial policy will not by itself be sufficient to achieve reform. Reform of psychologists' statistical practices is urgent and vital, but complex and challenging. Many important aspects require study. We believe the disciplines of statistics education and psychology bring together the expertise needed to carry out the necessary research and development work. Reforming psychologists' statistical practices will require attitude change and the acquisition of new understandings and skills by those teaching statistics in psychology, as well as by researchers, practitioners and students. It is therefore even more challenging than

overcoming students' naïve statistics beliefs. Even so, we offer our conceptual change model as a guide for the important and exciting task of the statistical re-education of psychology.

ACKNOWLEDGEMENT

This research was supported by the Australian Research Council.

REFERENCES

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

Batanero, C. (2001). Main research problems in the training of researchers. In C. Batanero (Ed.) *Training researchers in the use of statistics* (pp. 385-396). Voorburg, The Netherlands: International Association for Statistical Education, International Statistical Institute.

Belia, S., Fidler, F., & Cumming, G. (2001). Researchers' interpretation of the width of graphically-presented confidence intervals and other error bars. Manuscript in preparation.

Cumming, G. (2001). *Live figures: Interactive diagrams for statistical understanding.* Paper submitted to the Sixth International Conference on Teaching Statistics.

Cumming, G., & Finch, S. (2001a). 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. (2001b). Inference by eye: Confidence intervals, overlap, and how to read pictures of data. Manuscript in preparation.

Fidler, F., Edmonds, H., Fyffe, P., Harrington, C., Pannuzzo, D., Schmitt, R., Smith, J., & Cumming, G. (2001). Editorial influence on statistical practices in the Journal of Consulting and Clinical Psychology. Manuscript in preparation.

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., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., Anderson, J., & Goodman, O. (2001). Reform of statistical inference in psychology: The case of Memory & Cognition. Manuscript in preparation.

Finch, S., Thomason, N., & Cumming, G. (in press). Past and future APA guidelines for statistical practice. *Theory & Psychology*.

Loftus, G.R., & Masson, M. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*, 476-490.

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

Thomason, N. R., 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.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 70*, 80-93.

West, L.H.T., & Pines, A.L. (1985). *Cognitive structure and conceptual change*. Orlando, FL: Acdemic Press.

White, B.Y. (1993). ThinkerTools: Causal models, conceptual change, and science education. *Cognition & Instruction, 10*, 1-101.

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

Williams, J., Fidler, F., & Cumming, G. (2001). Researchers' interpretation of graphically-presented confidence intervals and other error bars: Implications for replication. Manuscript in preparation.