STATISTICS EDUCATION IN THE SOCIAL AND BEHAVIOURAL SCIENCES: FROM DICHOTOMOUS THINKING TO ESTIMATION THINKING AND META-ANALYTIC THINKING

Geoff Cumming
School of Psychological Science, La Trobe University, Australia
g.cumming@latrobe.edu.au

Null hypothesis significant testing (NHST) dominates in the social and behavioural sciences, despite strong evidence of its disadvantages. Worst may be its reinforcement of dichotomous thinking *(DT), which focuses on impoverished reject or don't-reject decisions. By contrast,* estimation thinking *(ET) and* meta-analytic thinking *(MAT) focus on sizes of effects, and cumulation of evidence to increase precision. A shift from DT to ET and MAT is highly desirable. Statistics education for ET and MAT will emphasise effect sizes, confidence intervals and meta-analysis, starting with the introductory course. Interpretation of confidence intervals should emphasise estimation, not NHST. Students and researchers, when specifying research goals, or discussing and interpreting findings, should use language that reflects ET and MAT. The outcome should be more quantitative theories, more sophisticated disciplines and better research progress.*

Of the many problems of null hypothesis significance testing (NHST), the most basic is its promotion of dichotomous thinking (DT) by its focus on a binary choice between rejecting or not rejecting a null hypothesis. Reform of statistical practice needs to overcome DT, and the way to do this is to adopt and teach estimation thinking (ET) and meta-analytic thinking (MAT). To support these contentions I first summarise the criticism of NHST and recommendations for better replacement techniques. Statistical practices shape fundamentally how researchers think about their experimental work and even the types of theory they develop. Conversely, adopting language that expresses richer ways of thinking should naturally encourage use of improved statistical techniques, especially estimation and meta-analysis. I will suggest that adopting estimation language is important for the statistical education needed for reformed statistical practice.

ADVOCACY OF STATISTICAL REFORM

For more than half a century NHST has been subjected to severe criticism of its conceptual foundations, the way researchers understand it, and the way it is used. Carver (1978) termed it "a corrupt form of the scientific method" (p. 378). Loftus (1991) wrote: "I find it difficult to imagine a less insightful means of transiting from data to conclusions" (p. 103). Paul Meehl (1978) declared: "reliance on merely refuting the null hypothesis… is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology" (p. 817). Kline (2004, Chapter 3) provided an excellent summary, and recommended a marked shift of emphasis away from NHST, towards estimation of effect sizes (ESs), the substantive interpretation of ES estimates and their confidence intervals (CIs), and replication.

The most basic flaw of NHST is that its impoverished outcome—a decision merely to reject or not reject the null—leads researchers to think in terms of merely testing such point null hypotheses. Research questions are framed as "Is there an effect?", and theories are conceived also in such simplistic terms. Such DT, and reliance on NHST are mutually reinforcing. The famous attack on NHST by Paul Meehl (1978) identified impoverished theorising as a dreadful outcome of reliance on NHST. He blamed "the Fisherian tradition [NHST], … [which] has inhibited our search for stronger tests, so we have thrown in the sponge and abandoned hope of concocting substantive theories that will generate stronger consequences than merely 'the Xs differ from the Ys'" (p. 824). We should, instead, develop quantitative theories that allow us to "generate numerical point predictions (the ideal case found in the exact sciences)" (p. 824). Gigerenzer (1998) agreed, and argued that, in psychology at least, NHST and poor theorising continued two decades later, and that reform needs to tackle both together. Like Meehl, he advocated improved theorising and better statistical techniques to build more quantitative, sophisticated and cumulative disciplines.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

ESCIMAT AND WAYS OF THINKING

Cumming (2010) summarised the immediate goal of statistical reform with the slogan ESCIMAT ("ESS-key-mat"), which combines ESs, CIs, and MAT. This goal is similar to Kline's (2004) and recognises that the main aim of most experiments is to find point and interval estimates (CIs) of the quantities of most interest. In practice, CIs are usually disappointingly wide, although the intervals are just reporting accurately the uncertainty in the data. Seeing wide intervals should encourage us to aim, wherever possible, for greater precision in future experiments. It should also lead us to think of our study in the context of previous studies on similar questions, and possible future studies: That is MAT (Cumming & Finch, 2001). Rarely can a single study settle a question; almost always we need cumulation of evidence over studies, by meta-analysis.

DT suggests asking "Is there an effect?", but ESCIMAT and ET instead ask "How large is the effect"? ET naturally focuses on sizes of effects of interest, and precision of estimates of those effects. ET is the key to expanding conceptions of theory beyond mere dichotomies to Meehl's substantive theories that are rich enough to generate quantitative predictions.

My first main contention is that improvement in statistical practices—a shift from NHST to ESs, CIs and meta-analysis—needs to be accompanied by changes in thinking. Replacing NHST as much as possible by estimation should naturally go with a shift from DT, to ET and MAT. Conversely, encouraging researchers to build quantitative theories, and plan research that asks estimation questions, should naturally lead to use of ESs and CIs, rather than tests of mere dichotomous hypotheses. Advocates of statistical reform need to discuss not only the statistical techniques researchers are advised to use, but the ways researchers can most effectively think about their theories, research planning, and interpretation of results. The desirable shift is from DT, to ET and MAT; from binary thinking to quantitative thinking.

NEW LANGUAGE TO PROMOTE NEW THINKING

Hoekstra, Finch, Kiers, and Johnson (2006) reported evidence from one journal that DT correlates with use of NHST. I sought further initial evidence about DT and alternatives by scanning the first ten articles in a recent issue of *Psychological Science*, a leading empirical journal in psychology. All 10 articles used NHST and eight primarily used language suggesting DT to express aims and conclusions. For example in one case the aim was "to determine whether proud individuals would… be viewed positively by their partners", and the conclusion: "individuals experiencing pride were more liked by their partners… $p = .02$."

Just two of the 10 articles used language suggesting ET: One "measured the degree to which…", and the aim of the other was "estimating the financial value of…." Analysis and interpretation focussed on, respectively, the degree to which…, and the estimated financial value.

Investigation of researchers' thinking by examining published articles requires analysis of language, especially the language used to express aims and conclusions. It seems easy to identify language that indicates underlying DT or ET, as in the brief examples above. Conversely, we can expect that deliberately adopting *estimation language* should encourage ET, and go naturally with use of ESs and CIs. My second main contention is that a useful strategy towards statistical reform is to encourage deliberate adoption of estimation language, and then *meta-analytic language*.

A thoroughgoing shift from DT to ET will require changes at every stage of research, including theorising, planning of experiments, analysing and reporting data, and discussing and interpreting findings. I suggest there are three points where, in many cases, it may be effective to promote change in behaviour. The first is to deliberately use estimation language to express theoretical expectations and experimental aims; this should influence theoretical conceptualisation, research planning, and selection of statistical techniques. The second step of deliberate behaviour change is to report ESs and CIs, and the third is to use estimation and meta-analytic language to interpret the ES estimates and their CIs, and to state conclusions.

The most important change point may be the first, because if theoretical predictions and experimental aims are expressed in terms of estimation, all else may follow. We should therefore encourage researchers not to state binary hypotheses, but to ask "To what extent is the novel procedure effective…?", or "How many…?", or "How strong is the relation between…?". The aim should be "to estimate the extent of…", or "to evaluate how closely the data fit these predictions…".

The medical journal *The Lancet* requires any report of a clinical trial to make "direct reference to an existing systematic review and meta-analysis. When a systematic review or meta-analysis does not exist, authors are encouraged to do their own" (Young & Horton, 2005, p. 107). This requirement reflects MAT, and encourages researchers to think of cumulation of evidence over studies, to achieve more precisely estimated ESs. Such thinking should also lead researchers to report and discuss their results in ways that make it easy to include them in future meta-analyses.

RESEARCH NEEDED

Evidence-based practice (EBP) is expected of practitioners of medicine, psychology and other disciplines. Beyth-Marom, Fidler, and Cumming (2008) argued that statistical practice should also be evidence-based. Statistical reform should be guided by relevant evidence, including evidence from statistical cognition (Beyth-Marom et al.), which is the study of how people understand statistical concepts and interpret statistical messages. If reformed statistical practice is to be EBP, and to support my argument, research is needed on several issues. First, does DT dominate, as I claim, and accompany use of NHST? The only relevant study I have found is that of Hoekstra et al. (2006), who found evidence that NHST does go with DT.

More generally, all my claims above need to be investigated empirically. What language do researchers use, and what are the relations among language, the underlying thinking, and choice of statistical technique? Journal studies and interviews with researchers could illuminate these relations. We also need intervention experiments in which researchers are assigned particular types of language to use, for example to express experimental aims. Then their choices of statistical techniques and the way they express conclusions would be analysed. We should also study how readers interpret findings, as a function of how they are expressed.

A major research aim would be to test my two contentions by examining how effective it could be to persuade researchers to change their language, as a helpful step towards choosing better statistical techniques. To what extent can deliberate use of estimation and meta-analytic language lead to improved choices of statistical methods and better reporting of conclusions?

ESCIMAT reform of statistical practices is needed across many social and behavioural science disciplines, and also to some extent in biological and medical disciplines—any discipline that relies heavily on NHST. However it is researchers in psychology and education that have the skills to conduct the statistical cognition and learning research that is needed to guide reform in all those disciplines. This research is an important and urgent task for psychology and education.

NEW LANGUAGE IN STATISTICS EDUCATION

The *Publication Manual* of the American Psychological Association (APA) is highly influential, being used by more than 1,000 journals across numerous disciplines. The latest edition (APA, 2010) includes NHST, but also makes strong recommendations for use of ESs and CIs. It is the first edition to give guidance and many examples to support ES and CI recommendations.

This edition (APA, 2010) says "State hypotheses and their correspondence to research design" (p. 28), and "After presenting the results, you are in a position to… interpret their implications, especially with respect to your original hypotheses" (p. 35). Such statements focus on testing hypotheses, which most researchers would take to mean NHST, and so the statements betray DT. There are, however, signs of change. It also states: "After you have introduced the problem… [state] your hypotheses or specific question…. Explain how the research design permits the inferences needed to examine the hypothesis or provide estimates in answer to the question" (p. 28). This covers ET as well as NHST. There are an encouraging number of references to investigating a problem or question, and making estimates. These are welcome signs in the language used by the *Manual* that DT is not always taken for granted and ET has a place.

Aron, Aron and Coups (2008), which we use in our introductory course, introduces NHST by referring to its "opposite-to-what-you-predict, roundabout reasoning… something like a double negative" (p. 147). By contrast, estimation is much more natural. If a chemist studies the melting point of a new plastic, even the beginning student should find it natural to see an answer like "132.5±0.2ºC". Similarly, if a psychologist studies a new therapy, a beginning student should need little explanation to understand an answer like "7.5 points on the anxiety scale, 95% CI [4.5,

10.5]", especially if the CI is shown as error bars in a figure. There is no need to warn about anything "opposite-to-what-you-predict"!

I propose that researchers examine the language of their own manuscripts, starting with the statements of aims. They should translate these where necessary into estimation language, then investigate what consequential changes are observed in statistical analysis, and the presentation and interpretation of results. To what extent is the result an improvement?

I propose that teachers of statistics examine their textbooks, notes, examples, exercises, and exam questions, and where necessary translate from dichotomous to estimation language. They should then follow through with consequential changes to statistical techniques, and language used to present and interpret findings. They should consider including meta-analysis, even in the introductory course (Cumming, 2006). Major reorganisation of material may be needed, but the result should be a course that equips students for an ESCIMAT world.

I propose that students should similarly scrutinise their teaching materials, and experiment with translations into estimation language. To what extent do these help understanding, and lead to more informative data analyses and improved planned experiments?

My argument, and these last specific proposals, require evaluation. Together they map out research that should help guide research practice from NHST towards ESCIMAT, with the potential for building more quantitative and cumulative—and thus more progressive—disciplines.

ACKNOWLEDGEMENTS

This research was supported by the Australian Research Council. I thank Pav Kalinowski for valuable comments.

REFERENCES

Aron, A., Aron, E. N., & Coups, E. J. (2008). *Statistics for the behavioural and social sciences: A brief course* (4th ed.). Upper Saddle River, NJ: Pearson Education.

American Psychological Association. (2010). *Publication manual of the APA* (6th ed.). Washington, DC: Author.

Beyth-Marom, R., Fidler, F., & Cumming, G. (2008). Statistical cognition: Towards evidence-based practice in statistics and statistics education. *Statistics Education Research Journal, 7*, 20-39.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Cumming, G. (2006). Meta-analysis: Pictures that explain how experimental findings can be integrated. Paper presented at The 7th International Conference on Teaching Statistics (ICOTS-7). Salvador, Brazil, July. In A. Rossman & B. Chance (Eds.), *ICOTS-7 Proceedings*. Online: www.stat.auckland.ac.nz/~iase/publications/17/C105.pdf.

Cumming, G. (2010). *p* values versus confidence intervals as warrants for conclusions that results will replicate. In B. Thompson & R. Subotnik (Eds.), *Research Methodologies for Conducting Research on Giftedness* (pp. 53-69). Washington, DC: APA Books.

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*, 532-574.

Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology, 8*, 195-204.

Hoekstra, R., Finch, S., Kiers, H., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of *p*-values. *Psychonomic Bulletin & Review, 13*, 1033-1037.

Kline, R. B. (2004). *Beyond significance testing. Reforming data analysis methods in behavioral research*. Washington, DC: APA Books. Chapter 3. Online: www.apastyle.org/manual/related/kline-2004.pdf.

Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology, 36*, 102-105.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Young, C., & Horton, R. (2005). Putting clinical trials into context. *The Lancet, 366*, 107.