META-ANALYSIS: PICTURES THAT EXPLAIN HOW EXPERIMENTAL FINDINGS CAN BE INTEGRATED

Geoff Cumming
La Trobe University, Australia
g.cumming@latrobe.edu.au

Meta-analysis (MA) is the quantitative integration of empirical studies that address the same or similar issues. It provides overall estimates of effect size, and can thus guide practical application of research findings. It can also identify moderating variables, and thus contribute to theory-building and research planning. It overcomes many of the disadvantages of null hypothesis significance testing. MA is a highly valuable way to review and summarise a research literature, and is now widely used in medicine and the social sciences. It is scarcely mentioned, however, in introductory statistics textbooks. I argue that MA should appear in the introductory statistics course, and I explain how software that provides diagrams based on confidence intervals can make many of the key concepts of MA readily accessible to beginning students.

META-ANALYSIS TO INTEGRATE AND SUMMARISE A BODY OF RESEARCH

Meta-analysis (MA) is a set of techniques to integrate empirical studies on the same or similar issues. MA burst onto the scene in 1976 when Gene Glass gave his presidential address to the American Educational Research Association and reported the results of his massive MA of studies that evaluated the effect of psychotherapy. Glass located more than 1,000 reports, but just 375 reported sufficient information to be included, and these gave an overall estimated effect size of 0.68 SD units. At a time when some were arguing psychotherapy was useless, this was a stunning outcome: Any single study may have flaws, but the body of literature, when integrated by MA, gave a clear message that, on average, psychotherapy has a large and clinically important benefit. At the same time Hunter and Schmidt (2004) were also developing MA techniques.

MA is now widespread in medicine and social science. *Psychological Bulletin*, the premier review journal in psychology, publishes many MAs. Statistical techniques and software have been developed. Very few, however, of even the latest introductory statistics texts give MA more than a cursory mention. I believe the ideas of MA are so fundamental to research, and to how data contributes to progress in science, that even beginning students should encounter MA. I present pictures that I have found make it easy for students to explore MA concepts.

Figure 1 shows fictitious data for 9 studies that asked whether a healthy breakfast before coming to school decreases anxiety. Each study obtained anxiety scores for children after healthy breakfasts. I assume 20 is a well-established population mean score for children, on the measure used in all the studies. For each study, type in the mean, SD, and *n*, then the study is summarised by a 95% confidence interval (CI). If the CI includes 20, the mean is not statistically significantly different from 20, at the .05 level, two-tailed—as for 6 studies in Figure 1. Three studies gave means statistically significantly less than 20. Figure 1 shows the result of the MA of the 9 studies, as a single CI: 18.8 is the overall estimate of anxiety after a healthy breakfast, and the precision of this estimate is quite high—the CI is narrow—reflecting the value of combining studies. The CI does just include 20, and so the overall *p* value is just greater than .05. (This example is simplified: Realistic MAs involve studies that include a control comparison, and standardisation of measures so that studies using different original measures can be included in the MA.)

META-ANALYSIS OVERCOMES PROBLEMS OF NHST

Null hypothesis significance testing (NHST) has been strongly criticised for decades (Cohen, 1994; Finch, Thomason, and Cumming, 2002). It is widely misunderstood and misused, and leads to poor research decision-making. Telling critiques of NHST were made by advocates of MA: Frank Schmidt (1996) explained how MA is superior to traditional ways of reviewing. In traditional reviews a large number of studies are found to support a finding but, typically, another large collection of studies failed to obtain the effect: They found 'no statistically significant effect'. The traditional reviewer concludes there are conflicting findings in the literature, no firm conclusion can be reach, and 'further research is required'. The typically low power of studies in

the social sciences means many studies inevitably will not find a statistically significant effect, even when there is a real effect, perhaps of a large and important size. MA, by contrast, ignores the statistical significance of any study, takes the point estimate from every study, combines these and gives the best overall point estimate, together with a measure of precision of that estimate.

Schmidt and others have described cases in which whole disciplines have been misled by traditional reviews, until more justifiable conclusions have been reached via MA. The erroneous belief in the lack of generality of job aptitude tests is one well-documented example.

META-ANALYSIS SHOULD CHANGE STATISTICAL PRACTICES

It would seem a simple expectation that any study reports its best estimate of the effect of interest, and the precision of that estimate. The appropriate CI is sufficient. Early meta-analysts found, however, that numerous studies did not report even that. Preoccupation with NHST was part of the problem: Researchers thought it sufficient to report they had obtained 'a highly significant improvement'. Without a value for how large the improvement was, that study would be useless for a MA, and therefore can contribute little or nothing to the cumulative discipline.

It is important that researchers 'think meta-analytically' (Cumming and Finch, 2001), meaning that they think of past research in MA terms, think of their own study in the context of a MA of past research, and report their own study in a way that facilitates future MAs. They need to report estimates of effect size, for all findings, together with precision of those estimates, for example by reporting CIs. The American Psychological Association (APA) *Publication Manual* (APA, 2001), which is used as a guide by more than 1,000 journals, advises "it is almost always necessary to include some index of effect size or strength of relationship…" (p. 25). One of main reasons statistical reformers advocate use of CIs, and much less reliance on NHST, is the need to support MA and meta-analytic thinking (Cumming and Finch, 2001, 2005).

Figure 1 shows the result of our (fictitious) study, then the final MA that includes our study. It is sobering to compare the two MA results, with and without our study: Our efforts served to shift the overall best estimate from 18.8 to 18.7, and to narrow the CI minutely! (The final CI happens to just miss 20, so our efforts have also resulted in a final p value less than .05.) Figure 1 helps us to think meta-analytically about previous research, and our own work.

META-ANALYSIS IN SOCIAL SCIENCE, MEDICINE, EVEN PHYSICS!

Hunt (1997) told the story of the rise of MA, and explained how MA has spread rapidly across medicine, and the social and behavioural sciences. Much of Hunt's book reads like a novel, and he makes the ideas of MA easily accessible. He argued that MA probably saved the funding of social science research, because legislators were getting tired of spending many millions only to be always told that 'findings conflict, more research is required'! MA may be the only realistic chance we have of finding evidence-based answers to numerous serious social issues. Hunter and Schmidt (2004) have also been making cogent arguments for MA, over almost 30 years.

One of Hunt's dramatic examples was from medicine. He presented a diagram like Figure 1—called in medicine a Forest plot—to summarise 33 studies examining whether streptokinase is beneficial for acute myocardial infarction (AMI). The studies involved some 37,000 patients, approximately half of whom received inactive placebo rather than the drug. Only 6 of the 33 studies showed a statistically significant benefit of streptokinase for AMI. Hunt, however, also showed a cumulative MA figure, in which the studies were arranged in date order, from 1959 to 1988, then MAs were conducted, on the first 2 studies, then the first 3, etc. Successive CIs summarised the state of knowledge after each successive study. By 1997, after 15 studies and some 4,300 patients, the p value for the MA result was less than .001. For practical purposes the issue was settled: Streptokinase is beneficial in AMI. Had MA been used back then, it would have been unethical to conduct any of the subsequent studies, and more than 32,000 patients would not have been asked to participate—many to be given placebo and some to die because of that.

The Cochrane Collaboration (www.cochrane.org) is a public database of reports—mostly MAs—that give up-to-date summaries of research evidence on numerous questions of medical importance. Cochrane also includes reports about some psychological treatments, for example the efficacy of cognitive-behavioural therapy for a range of conditions. Towards the end of most reports are several or many Forest plots that summarise the MA of all available studies on the

issue. Increasingly, funding bodies and ethics committees require that research results be made available to Cochran, or other appropriate public database, as a condition for funding and ethics approval. Medical research is already thinking meta-analytically. It is definitely time that the social and behavioural sciences caught up.

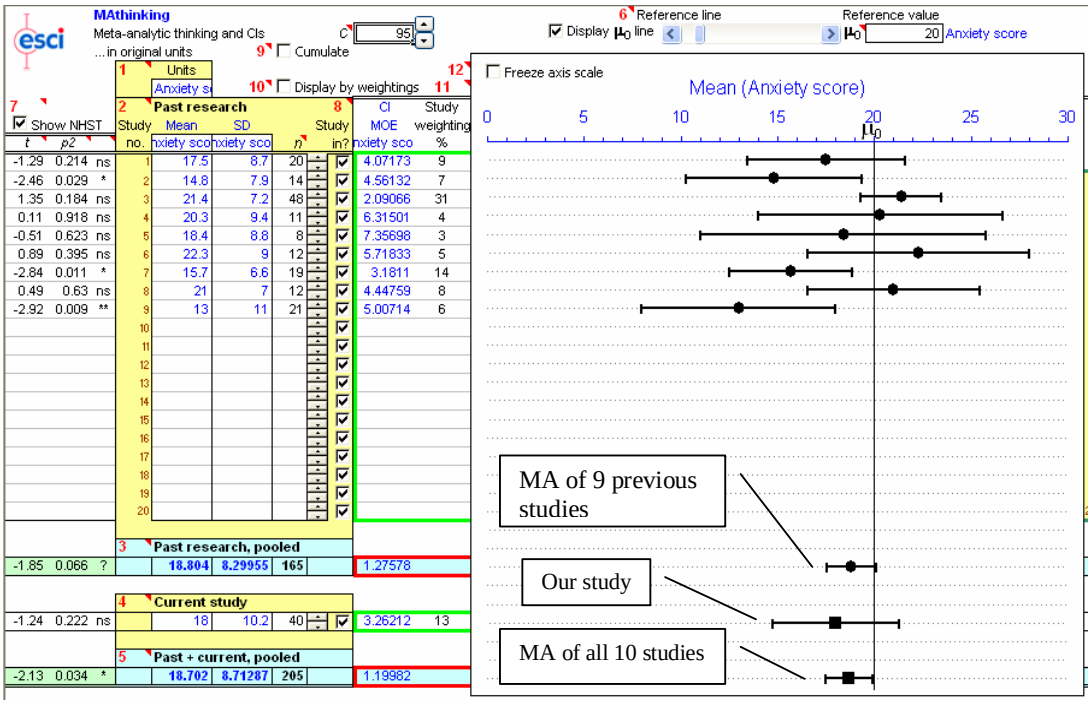

Figure 1: An image from ESCI (Exploratory Software for Confidence Intervals, which runs under *Excel*) that uses CIs to illustrate MA.

Hedges, a pioneer of MA, studied how particle physics copes with disagreement among results of experiments in different laboratories. He reported (Hedges, 1987) that physics uses MA: The statistical techniques to assess and combine data were very similar to those coming into use in social science, even if the terminology was different. Further, Hedges compared the extent of variation across studies in particle physics and in social science and argued it was quite similar in the two fields. MA suggested that hard science is no harder than soft science!

It can be an enormous task to undertake a major MA, partly because great effort is needed to locate every possible study on the question, whether published or not. MA requires many decisions, for example to define the questions to be analysed, and to specify requirements that studies must meet to be included in the MA. There is also the 'apples and oranges' problem: How similar must studies be to warrant inclusion? Critical analysis of the body of research, and of the MA itself is also required: MA does not licence any decrease in critical standards!

UNDERSTANDING AND TEACHING META-ANALYSIS THROUGH PICTURES

Examination of Figure 1 illustrates some important aspects of MA. Students can use the software's interactive features to explore many further aspects. Vary the mean, SD, *n* of one or more studies to see how this changes the CIs for those studies, and the MA result. Click to remove a study from the MA and thus see the influence of any single study. For example, remove one or more studies that individually do not show statistical significance (the CI includes 20), and see the change in the MA result—this illustrates the importance of 'the file drawer effect', the under-reporting of studies that fail to obtain statistical significance.

Long CIs naturally attract the eye, but indicate low precision. Weighting of studies in MA is inversely proportional to study variance. Equivalently, study weighting is inversely related to CI length. In Figure 1, study weighting is shown in percent, in the column just to the left of the

CIs. Figure 2 is a display that indicates weightings approximately by size of the dot, and thickness of the CI arms. Larger and heavier graphics attract the eye, and indicate a more heavily-weighted study. It is also possible to show cumulative MA, as Hunt described.

MA provides more than simply an overall point estimate based on all contributing studies. Studies are coded for various features considered important, for example the age of the participants, whether or not blind scoring was used, and the delay between treatment and final measurement. If the MA includes sufficiently many studies, multiple regression can be used to assess the influence such coding variables have on the effect sizes observed in different studies. A coding variable with a large influence is identified as a likely moderating variable. This procedure is one illustration of how MA can contribute to theoretical advance, by identifying important relations among variables, possibly including variables not manipulated in any single study.

My experience with first and third year psychology students suggests lectures using the software illustrated here, and students' own work with this software, are effective in helping them understand MA and its importance. I plan to evaluate the effectiveness of this approach. I look forward to the arrival of MA in introductory courses in statistics and research methods. MA has such a central role in modern research it is essential that students appreciate its importance from the start, and software based on good pictures can help them do that.

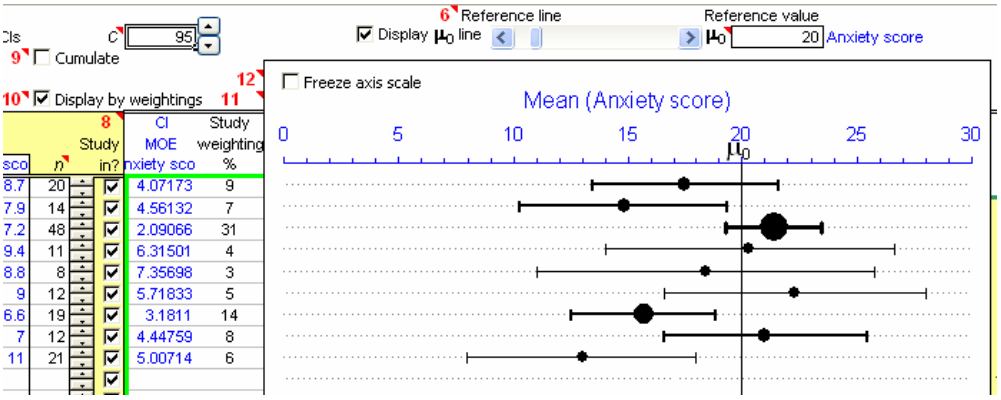

Figure 2: Study weightings are shown in percent at left, and by CI dot size and line thickness

ACKNOWLEDGEMENT

I thank Fiona Fidler for comments on a draft.

REFERENCES

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

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

Cumming, G. and Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61, 530–572.

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

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

Hedges, L. (1987). How hard is hard science, how soft is soft science? *American Psychologist*, 42, 443-455.

Hunt, M. (1997). *How Science Takes Stock. The Story of Meta-Analysis*. New York: Sage.

Hunter, J. E. and Schmidt, F. L. (2004). *Methods of Meta-Analysis: Correcting Error and Bias In Research Findings* (2nd edition). Thousand Oaks, CA: Sage.

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