The demerits of meta-analysis in science

Ranjan Bhattacharyya and Avijit Ghosh

Meta-analysis now offers the opportunity to critically evaluate and statistically combine results of comparable studies or trials. Its major purposes are to increase the number of observations and the statistical power, and to improve estimates of the effect size of an intervention or an association. Meta-analysis can be used for the following purposes: to establish statistical significance with studies that have conflicting results; to develop a more correct estimate of effect magnitude, and to provide a more complex analysis of harms, safety data and benefits. A meta-analysis is a safer starting point than a single study, but it would not be necessarily more reliable. A meta-analysis is usually part of a systematic review. It is a heavy-duty effort and is often described as the ultimate study outweighing all others.

However, meta-analysis has many demerits. The main problem is the potential for publication bias and skewed data. Research generating results that do not reject null hypotheses may remain unpublished, or may not be entered into a database. Meta-analysis does the pooled estimate of effect, allows for an objective appraisal of evidence and may reduce the probability of false-negative results. The heterogeneity between study results may be explained and it avoids Simpson's paradox. A meta-analysis is a statistical procedure for combining numerical data from multiple separate studies. It should only be performed in the context of a systematic review. Its detailed scientific demerits are mentioned below.

Reviews of meta-analysis in agriculture include those of Philibert *et al.*[1] and Brandt *et al.*[2], who suggest that the methodological quality and application of meta-analytical techniques are highly variable. Most meta-analyses in agronomy focus on crop yield response to experimental manipulation[1]. Yield is however only one criterion by which the performance of cropping systems can be judged: yield stability and resilience, nutritional yield and environmental and economic performance are additionally relevant but less-studied indicators.

This simplification – which also involves subjective decision-making to include or exclude treatments and management regimes – is not inherent to meta-analysis alone. Rather, these issues influence the design and administration of agronomic experiments in general. Yet, the problems of standardization and simplification appear to be amplified by meta-analyses, at times reducing their value for agricultural policy or improving farmer practice.

The framing of meta-analysis is an important yet politically contested topic. Most meta-analyses reviewed highlighted the size and comprehensiveness of their datasets, while implying a capability to answer questions of regional or 'global' significance for food production, food security or environmental challenges. The potential of these analyses to achieve unifying conclusions that have global as well as local relevance, however, ironically appears to be undermined by the large geographical scale at which the results tend to be presented. The presentation of 'global' average results decoupled from the context-specific and diverse qualities of farming systems is unlikely to inform policy and investment decisions meaningfully nor inform ways to improve farmer practice. Small-plot and research station-based experiments may not represent whole-field or whole-farm functioning and may inadequately reflect cropping system dynamics, and the economic and resource allocation choices made by farmers outside the experimental setting. This problem is inherent to the organization of agronomic research, and appears to be amplified in meta-analysis that combines multiple field trials to generate more comprehensive results.

Meta-analysis has not been able to reduce controversies within agriculture – in some cases, it does just the opposite. While meta-analysis is increasingly popular and is of general scientific interest, we suggest that its use to appraise and prioritize agricultural research and development investments should be carefully employed by considering the analytical limitations of the method. Scientists and policy-makers evaluating the results of future meta-analyses should consider how treatments are defined and constructed, and how papers and data are collected, screened and analysed. Although most assessments of the value of meta-analysis focus on quantitative methods, the ways in which researchers justify, frame and position their research questions are also important, as these factors can condition the ways in which statistical analyses are interpreted and discussed. In addition, a critical evaluation of how researchers interpret data derived from plot-scale experiments and discuss their results in farming systems at a regional or global scale is needed. Lastly, when meta-analysis is applied to highly politicized topics, as is the case with organic and conservation agriculture, a more cautious interpretation of results that recognize the socially and politically embedded nature of agricultural research is required.

Most of the criticisms of meta-analysis are related to the potential error and bias that can result from combining studies. Error and bias in a meta-analysis can stem from a series of interrelated issues.

First, with the mixed studies used in a meta-analysis; differences between studies have been referred to as the apples and oranges problem. For example, it may not be appropriate to employ meta-analysis studies that use different methodologies (e.g. surveys versus experiments), sampling designs, and/or variable measurements. Such methodological variables, however, can be coded and used to test for their impact on the overall findings. In response to the apples and oranges criticism, Glass[3] stated, 'Of course, it mixes apples and oranges; in the study of fruit nothing else is sensible; comparing apples and oranges is the only endeavour worthy of true scientists; comparing apples to apples is trivial.'

Second, the concern with mixing studies also includes the problems associated with methodological quality (garbage in–garbage out). If studies with poor methodological quality are included, the meta-analysis results may be biased. A judgment on the quality of each study by the researcher minimizes this problem. For example, each study can simply be rated as high or low quality[4]. Other methods such as rating threats to internal and external validity[5,6] and evaluation of the methodology of each study have been proposed[7]. These methods are relatively easy to accomplish and can be done as part of setting the standards for study inclusion and exclusion in the early steps of meta-analysis.

Third, the inclusion or exclusion of specific studies can influence error and bias. Glass and coworkers[8,9] consistently suggest that a priori considerations of study quality and differences are not necessary for meta-analysis. Others have suggested means for limiting inclusion/exclusion bias[10,11].

Fourth, a related concern has been referred to as the file-drawer problem, which involves the issue of publication bias. This problem arises when a meta-analysis attempts to review all significant and non-significant findings to provide a complete perspective. Non-significant findings, however, are often not published and the file-drawer problem becomes an important issue[12]. Methods have been developed to help detect and minimize publication bias[13], such as the funnel plot (i.e. a scatterplot of sample size versus estimated effect size for a group of studies[14,15], checking the file drawer of unpublished studies and estimating their numbers[16,17], trim-and-fill estimates to determine the number of missing studies[18,19], and weighted estimation methods[20,21]. Alternatively, the file-drawer problem can be solved by specifying the standards for inclusion of studies. By limiting the studies to those with adequate power which the researchers also consider would pass high standards of peer-review for publication or funding[22], he/she can eliminate publication bias since the sample is not intended to cover all the studies.

Fifth, using multiple findings from the same study can be a source of bias in meta-analysis because the corresponding effect sizes may not be independent of one another[9,23]. Statistical methods are often used to account for effect sizes that are not independent. In the comparative analyses presented earlier, more than one evaluation context could come from a single study. For example, respondents in a single study could rate their perceptions of crowding at the trailhead, on the trail and at the summit of a mountain. One solution is to take an average of the dependent effect sizes in each study, but this approach contradicts the previous literature showing that perceptions of crowding vary by location of the encounter[21]. Other solutions may include: (i) treating the effect sizes as independent, (ii) choosing only one effect size from each study, (iii) computing the degree of interdependence from intercorrelations and (iv) computing a weighted average of the correlations[24,25]. Ultimately, the solution of choice depends on the research question and the magnitude of dependence between the effect sizes[26,27]. Overall, researchers must balance these concerns (e.g. apples and oranges, garbage in–garbage out, file drawer) against the objectives of the meta-analysis and the research questions to be addressed. Meta-analysis is only as good as the individual studies from which it is composed. The tendency to overestimate the value of the results of a meta-analysis without considering the individual studies should be avoided. Careful documentation of the procedures that were followed can minimize these problems.

1. Philibert, A., Loyce, C. and Makowski, D., *Agric., Ecosyst. Environ*., 2012, **148**, 72–82.
2. Brandt, K., Srednicka-Tober, D., Baranski, M., Sanderson, R., Leifert, C. and Seal, C., *J. Agric. Food Chem.*, 2013, **61**, 7173–7180.
3. Glass, G. V., 2000; http://glass.ed.asu.edu/gene/papers/meta25.html (accessed on 10 June 2007).
4. Hedges, L. V., *Rev. Res. Educ.*, 1982, **13**, 353–398.
5. Campbell, D. T. and Stanley, J. C., *Experimental and Quasi-Experimental Designs for Research*, Rand McNally, Chicago, USA, 1966.
6. Cook, T. D., Campbell, D. T. and Day, A., *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Houghton Mifflin, Boston, 1979, vol. 351.
7. Cooper, H. M., *The Integrative Literature Review: A Systematic Approach*, Sage, CA, USA, 1984.
8. Glass, G. V., *J. Res. Sci. Teach.*, 1982, **19**(2), 93–112.
9. Glass, G. V., Mc Gaw, B. and Smith, M. L., *Meta-Analysis in Social Research*, Sage, CA, USA, 1981.
10. Cooper, H. M., *Rev. Educ. Res*., 1982, **52**(2), 291–302.
11. Hunter, J. E. and Schmidt, F. L., *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, Sage, CA, USA, 1990.
12. Hedges, L. V., *Stat. Sci*., 1992, **7**(2), 246–256.
13. Pham, B., Platt, R., McAuley, L., Klassen, T. P. and Moher, D., *Eval. Health Prof.*, 2001, **24**(2), 109–125.
14. Copas, J. and Shi, J. Q., *Biostatistics*, 2000, **1**(3), 247–262.
15. Light, R. J. and Pillemer, D. B., *Summing Up: The Science of Reviewing Research*, Harvard University Press, Cambridge, MA, USA, 1984.
16. Iyengar, S. and Greenhouse, J. B., *Stat. Sci.*, 1988, **3**(1), 109–135.
17. Rosenthal, R., *Psychol. Bull.*, 1979, **86**, 638–641.
18. Duval, S. and Tweedie, R., *J. Am. Stat. Assoc.*, 2000, **95**(449), 89–98.
19. Givens, G. H., Smith, D. D. and Tweedie, R. L., *Stat. Sci.*, 1997, **12**(4), 221–250.
20. Murtaugh, P. A., *Ecology*, 2002, **84**(4), 1162–1166.
21. Sutton, A. J., Song, F., Gilbody, S. M. and Abrams, K. R., *Stat. Methods Med. Res.*, 2000, **9**, 421–455.
22. Kraemer, H. C., Gardner, C., Brooks III, J. O. and Yesavage, J. A., *Psychol. Methods*, 1998, **3**, 23–31.
23. Rosenthal, R. and DiMatteo, M. R., *Annu. Rev. Psychol.*, 2001, **52**(1), 59–82.
24. Cheung, S. F. and Chan, D. K. S., *J. Appl. Psychol.*, 2004, **89**(5), 780–791.
25. Gleser, L. J. and Olkin, I., In *The Handbook of Research Synthesis* (eds Cooper, H. and Hedges, L. V.), Sage, New York, USA, 1994, pp. 339–355.
26. Rosenthal, R., Rosnow, R. L. and Rubin, D. B., *Contrasts and Effect Sizes in Behavioral Research*, Cambridge University Press, Cambridge, UK, 2000.
27. Shelby, B., Vaske, J. J. and Heberlein, T. A., *Leis. Sci.*, 1989, **11**, 269–291.

Ranjan Bhattacharyya and Avijit Ghosh are in the Division of Environment Science, ICAR-Indian Agricultural Research Institute, Pusa, New Delhi 110 012, India.*
**e-mail: ranjanvpkas@gmail.com*