23 Loew, E.R. (1994) *Vision Res.* 34, 1427–1431
24 Fleishman, L.J., Loew, E.R. and Leal, M. (1993) *Nature* 365, 397
25 Bennett, A.T.D. and Cuthill, I.C. (1994) *Vision Res.* 34, 1471–1478
26 Maier, E.J. and Bowmaker, J.K. (1993) *J. Comp. Physiol. A* 172, 295–301
27 Bowmaker, J.K. (1991) in *Vision and Visual Dysfunction: The Perception of Colour* (Gouras, P., ed.), pp. 108–127, Macmillan Press
28 Maier, E.J. (1994) *Vision Res.* 34, 1415–1418
29 Maier, E.J. (1992) *J. Comp. Physiol. A* 170, 709–714
30 Derim-Oglu, E.N. and Maximov, V.V. (1994) *Vision Res.* 34, 1535–1539
31 Jacobs, G.H. (1993) *Biol. Rev.* 68, 413–471
32 Jacobs, G.H. and Deegan, J.F., II (1994) *Vision Res.* 34, 1433–1441
33 Tovée, M.J. (1994) *Trends Neurosci.* 17, 30–37
34 Wehner, R. (1989) *J. Exp. Biol.* 146, 63–85
35 Hawryshyn, C.G. and McFarland, W.N. (1987) *J. Comp. Physiol. A* 160, 459–465
36 Hawryshyn, C.G. and Bolger, A.F. (1990) *J. Comp. Physiol. A* 167, 691–697
37 Hawryshyn, C.G., Arnold, M.G., Bowering, E. and Cole, R.I. (1990) *J. Comp. Physiol. A* 166, 565–574
38 Land, M.F. (1991) *Nature* 353, 118–119
39 Cameron, D.A. and Pugh, E.N. (1991) *Nature* 353, 161–164
40 Coemans, M.A.J.M., Vos Hzn, J.J. and Nuboer, J.F.W. (1994) *Vision Res.* 34, 1461–1470
41 Emmerton, J. (1983) *Percept. Psychophys.* 34, 555–559
42 Emmerton, J. and Remy, M. (1983) *Experientia* 39, 1161–1163
43 Loew, E.R., McFarland, W.N., Mills, E.C. and Hunter, D. (1993) *Can. J. Zool.* 71, 384–386
44 Kunz, Y.W., Wildenburg, G., Goodrich, L. and Callaghan, E. (1994) *Vision Res.* 34, 1375–1383
45 Loew, E.R. and Wahl, C.M. (1991) *Vision Res.* 31, 353–360
46 Browman, H.I. and Hawryshyn, C.W. (1992) *Vision Res.* 32, 2303–2312
47 Browman, H.I. and Hawryshyn, C.W. (1994) *Vision Res.* 34, 1397–1406
48 Thorpe, A. and Douglas, R.H. (1993) *Vision Res.* 33, 301–307
49 Meyer-Rochow, V.B. (1991) *J. R. Soc. N. Z.* 21, 169–177
50 Vanewright, R.I. and Boppre, M. (1993) *Philos. Trans. R. Soc. London Ser. B* 340, 197–205
51 Weiss, M.R. (1991) *Nature* 354, 227–229
52 Lehrer, M., Horridge, G.A., Zhang, S.W. and Gadagkar, R. (1995) *Philos. Trans. R. Soc. London Ser. B* 347, 123–137
53 Menzel, R. and Shmida, A. (1993) *Biol. Rev.* 68, 81–120
54 Chittka, L., Shmida, A., Troje, N. and Menzel, R. (1994) *Vision Res.* 34, 1489–1508
55 Goldsmith, T.H. (1980) *Science* 207, 786–788
56 Vlitala, J., Korpimaki, E., Palokangas, P. and Kolvula, M. (1995) *Nature* 373, 425–427
57 Finger, E. and Burkhardt, D. (1994) *Vision Res.* 34, 1509–1514
58 Brunton, C.F.A. and Majerus, M.E.N. *Proc. R. Soc. London Ser. B* (in press)
59 Bernard, G.D. and Remington, C.L. (1991) *Proc. Natl Acad. Sci. USA* 88, 2783–2787
60 Brainard, G.C. *et al.* (1994) *Vision Res.* 34, 1521–1533
61 Okano, T. *et al.* (1992) *Proc. Natl Acad. Sci. USA* 89, 5932–5936
62 Johnson, R.L. *et al.* (1993) *Biochemistry* 32, 208–214
63 Cserhati, P., Szel, A. and Rohlich, P. (1989) *Invest. Ophthalmol. Visual Sci.* 30, 74–81
64 Szél, A. and Rohlich, P. (1992) *Exp. Eye Res.* 55, 47–52
65 Robinson, J. *et al.* (1993) *Proc. Natl Acad. Sci. USA* 90, 6009–6012
66 Dittrich, W., Gilbert, F., Green, P., McGregor, P. and Grewcock, D. (1993) *Proc. R. Soc. London Ser. B* 251, 195–200
67 Windig, J.J., Brakefield, P.M., Reitsma, N. and Wilson, J.G.M. (1994) *Ecol. Entomol.* 19, 285–298

Statistical power of methods of meta-analysis

The review of meta-analysis (MA) by Arnqvist and Wooster[1] was informative and useful. It highlighted how MA can derive conclusions from a set of research studies, even if the individual studies have low statistical power (probability of correctly rejecting the null hypothesis[2]). Ironically, they appear to have overlooked a major point concerning the statistical power of the conclusions from their own example of MA. They applied MA to a hypothetical example of 15 different studies of the effect of a factor *x* on a response variable *y*. As well as stating other conclusions, they noted in Box 2 that:

> We then test for heterogeneity across studies by performing a diffuse test of homogeneity, but the null hypothesis that all studies share a common effect size [$\chi^2_{(14)} = 13.43$, $P > 0.5$] cannot be rejected...To conclude, this simple MA has...indicated that the outcome of the 15 studies are statistically indistinguishable and thus, in that sense, indeed 'consistent'.

However, the authors fail to state the statistical power of the MA method itself in their example! If power is low, then there is a high probability of committing a Type II error, that is, failing to reject the null hypothesis when the null is false. In the above quotes, the authors thus apparently committed the far-too-common error of failing to reject some null hypothesis and then concluding that the null must therefore be true, without asking what the statistical power was of the method of inference, given the data[3–6]. Because the authors failed to state anything about the power of the MA in this case, they are not justified in automatically concluding that the 15 studies are consistent. It is possible that they are indeed consistent, but it is also possible that they are not and that the MA method simply did not have a high probability of detecting inconsistency (i.e. rejecting the null hypothesis of homogeneity).

While the authors stated in Box 4 that the probability of making Type II errors is 'drastically reduced' with MA methods, the readers of *TREE* would benefit greatly if the authors could inform us of the statistical power of their particular example of MA, especially for different effect sizes (degrees of homogeneity among studies). As well, there are undoubtedly general factors that affect statistical power of MA methods. In classical statistical methods, for instance, power is affected by α, sample size, sample variance, and the true effect size[2]. What affects the power of MA methods? What guidelines can the authors provide for drawing conclusions from MA when a researcher fails to reject some null hypothesis?

Randall M. Peterman

School of Resource and Environmental Management, Simon Fraser University, Burnaby, B.C., Canada V5A 1S6

References

1 Arnqvist, G. and Wooster, D. (1995) *Trends Ecol. Evol.* 10, 236–240
2 Dixon, W.J. and Massey, F.J., Jr (1983) *Introduction to Statistical Analysis* (4th edn), McGraw Hill
3 Parkhurst, D.F. (1990) in *Acting Under Uncertainty: Multidisciplinary Conceptions* (von Furstenberg, G.M., ed.), pp. 181–201, Kluwer
4 Peterman, R.M. (1990) *Can. J. Fish. Aquat. Sci.* 47, 2–15
5 Peterman, R.M. (1990) *Ecology* 71, 2024–2027
6 Shrader-Frechette, K.S. and McCoy, E.D. (1995) *Trends Ecol. Evol.* 7, 96–99

Reply from G. Arnqvist and D. Wooster

Peterman raises the issue of statistical power in meta-analysis (MA). Indeed, the basic principles for formulating and rejecting null hypotheses are shared between MA and conventional inferential statistics. A failure to reject a null hypothesis cannot, in itself, validate acceptance of the null. With regard to the quote from Box 2 in our review[1], we concluded only that the results of the studies were statistically indistinguishable (as evaluated by a simple test of homogeneity). The phrase 'consistent' in Box 2 relates to the phrase 'inconsistent' in Box 1, and was used to contrast the conclusions from a narrative review with our meta-analytic synthesis. In other words, we wanted to illustrate and make it clear that there are no objective grounds for claiming that the studies in our example are 'inconsistent' (i.e. the null cannot be rejected), a subjective conclusion often found in narrative reviews of sets of studies such as this.

However, we wish to make two points. First, statements such as 'inconsistent' are, for the most part, meaningless when multiple studies are compared, unless accompanied by quantitative statistical evaluations. Second, diffuse tests, used for illustration in our simple example, are often relatively blunt instruments for hypothesis testing. A wide variety of more-complex and focused tests for comparing subgroups of studies are available, and should be employed in concert with diffuse tests of homogeneity in more extensive MAs[2].

Formal methods for estimating statistical power of meta-analytic inferences are under development. Unfortunately, none of these methods has yet been published, and thus none is available. It is, nevertheless, clear from simulations that different methods of combining probabilities (i.e. omnibus tests) have different properties, including different statistical power, and some general guidelines as to when different tests should be used have been given[2]. There is no doubt that the power of an MA widely exceeds that of the single studies in general, especially when simple hypotheses are being tested[2,3]. For any given significance level (α), three major factors determine the 'power' of an MA: the number of studies included, the quality of each of the included studies (e.g. their sample sizes), and the true effect size (or true difference in effect size between subgroups of studies). Unfortunately, the meta-analyst often faces a trade-off between the relevance and the power of the analysis[4]: as the number of potentially relevant covariates (or grouping variables) increases, the power decreases, since the number of studies in each unique subgroup decreases and the number of statistical tests increases. The most efficient way to avoid this problem is undoubtedly to base the MA on as large a number of studies as possible[4]. An alternative strategy may be to employ a bayesian approach to meta-analytic inference[5].

Thus, until formal methods of estimating the power of meta-analytical inference are available, a failure to reject a null hypothesis is not informative of whether or not the null is actually true.

Göran Arnqvist

Dept of Biology, University of New Mexico, Albuquerque, NM 87131-1091, USA

Dave Wooster

Center for Evolutionary Ecology, University of Kentucky, Lexington, KY 40506-0225, USA

References

1 Arnqvist, G. and Wooster, D. (1995) *Trends Ecol. Evol.* 10, 236–240
2 Cooper, H.M. and Hedges, L.V., eds (1994) *The Handbook of Research Synthesis*, Russel Sage Foundation
3 Arnqvist, G., Rowe, L., Krupa, J.J. and Sih, A. (1995) *Evol. Ecol.* 9, 1–20
4 Matt, G.E. and Cook, T.D. (1994) in *The Handbook of Research Synthesis* (Cooper, H.M. and Hedges, L.V., eds), pp. 503–520, Russel Sage Foundation
5 Louis, T.A. and Zelterman, D. (1994) in *The Handbook of Research Synthesis* (Cooper, H.M. and Hedges, L.V., eds), pp. 411–422, Russel Sage Foundation

Serious bias in journal impact factors

In their recent *TREE* article subtitled 'Don't worry', Statzner *et al.*[1] give two reasons for the relatively poor Impact Factors (IF) of even the best ecological journals. Here I highlight two further reasons for bias in IF, and suggest that we have every reason to be worried.

IFs are an index of the rate at which papers published in a journal are subsequently cited; they are supposedly a quantitative and objective measure of the quality of journals and, by implication, the scientists who publish in them (the higher the IF, the higher the quality). However, the top-ranked journals in molecular biology and biochemistry have IFs ten times higher than the top ecology, evolutionary biology or animal behaviour journals[1]. Statzner *et al.* suggest that this is partly because of the size of the discipline and the time taken to conduct subsequent research after a seminal paper appears.

I have recently shown[2] that two additional parameters (that are rather easier to quantify, but again are unrelated to the quality of the science) can explain a huge amount of the variation in IF between journals. These are the publication lag (mean time from submission to publication of papers) and the turnover rate (taken as the mean percentage of references in a journal that were published within the past two years – a measure of how much the discipline relies on recently published material).

The variation in these two variables is much greater between than within disciplines: in my survey of 'top' journals in the biological and biomedical sciences, I found that biochemical and molecular journals had much shorter mean publication lags than ecology journals (6.2 ± 0.5 months cf. 16.3 ± 1.2) and higher turnover rates ($28.2 \pm 2.5\%$ cf. $9.5 \pm 1.2\%$). The two variables are to some extent correlated: authors of ecological papers simply cannot cite many references from the past two years because they submitted their manuscript before most of these papers were published. Nonetheless, both variables were found to be significant in a multiple regression with IF as the dependent: they explained almost 72% of the tenfold variation in IF of top journals across the biological and biomedical sciences. This is due to the calculation of IFs being based on very recent citations.

We are therefore faced with the alarming fact that some disciplines (notably ecology, but also non-molecular medicine) are given poor rankings in the IF league tables for reasons totally unrelated to their quality. Given that the IFs of the journals in which we publish are increasingly being used as 'reliable' indicators of our own quality[1], we should be extremely worried, and should highlight these systematic biases before the use of unrefined IFs becomes even more prevalent.

Neil B. Metcalfe

Division of Environmental and Evolutionary Biology, Graham Kerr Building, IBLS, Glasgow University, Glasgow, UK G12 8QQ

References

1 Statzner, B., Resh, V.H. and Kobzina, N.G. (1995) *Trends Ecol. Evol.* 10, 220
2 Metcalfe, N.B. (1995) *Nature* 376, 720

Erratum

Purrington, C.B. and Bergelson, J. *Trends Ecol. Evol.* 10, 340–342 (August 1995)

In Fig. 1, incorrect shading was introduced into the artwork, completely obscuring the distinction between two categories in the pie charts. The correct version is printed below. We apologize for this error. **Ed.**

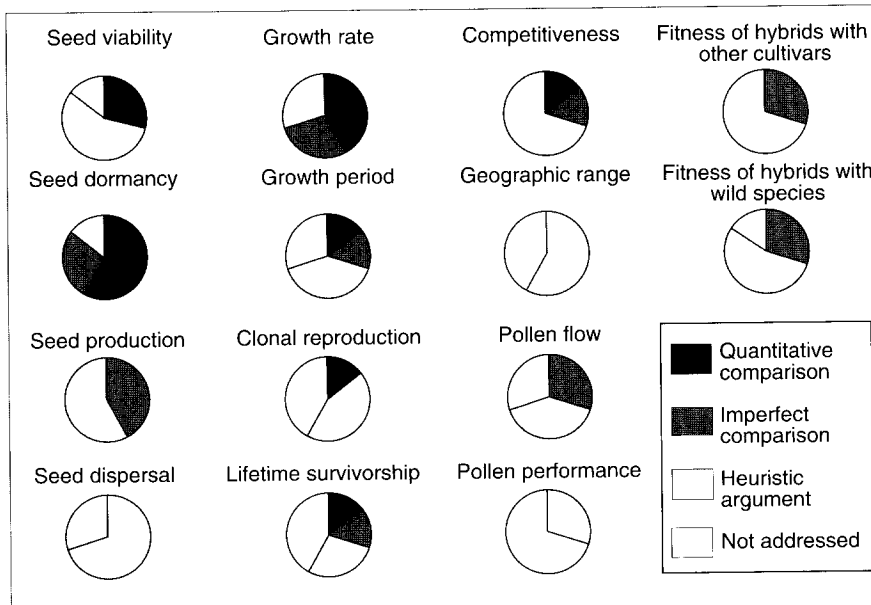

Fig. 1. Approved petitions (as of 15 March 1995) for nonregulated status for transgenic crops, categorized for their treatment of 14 issues relating to performance of plants containing transgenes. Possible types of evidence are shown in four categories, those containing quantitative comparisons (based on experimental comparison of transgenic and nontransformed genotypes), imperfect comparisons (based on observations or flawed experiments), heuristic arguments (unsubstantiated suppositions that experimentally derived data are not needed), and those lacking any discussion of the issue. When present at all, discussions of hybrid fitness in the petitions were typically restricted to one or two characters; the figures relating to fitness of hybrids therefore substantially overestimate the attention paid to this issue. *All petitions provided by USDA APHIS.*