

Meta-analysis without the mantra: a reply to Wilson and “weighted analysis”

ROBERT L. BANGERT-DROWNS¹ &
ELIZABETH WELLS-PARKER²

¹State University of New York at Albany & ²Mississippi State University, USA

In 1976, Gene Glass first articulated the notion of meta-analysis as the application of statistical methods to literature review (Glass, 1976). In its most essential form, meta-analysis translates study outcomes into a common metric, a measure of effect size, and applies statistical methods to describe and analyze the distribution of these outcomes (Bangert-Drowns, 1995). Within its first decade, several distinctive approaches to meta-analysis already had emerged, distinguishable in terms of purpose, unit of analysis, treatment of study variation and typical analytical products (Bangert-Drowns, 1986). A quarter of a century after the articulation of meta-analysis, Wilson's (2000) survey of meta-analysis in research on substance abuse treatment echoes and extends familiar themes about which meta-analysts are reaching reasonable consensus. These include concerns about non-independence of effect sizes drawn from the same study, strategies for differentiating constructs in independent and dependent variables, dissatisfaction with techniques that only combine probabilities without analyzing variation among studies, questions of appropriate numbers of studies to constitute a meta-analysis and ways to represent treatment effects for various audiences and purposes.

In contrast, Wilson (2000) may impose premature closure on a complex topic by advo-

cating that effect sizes always be weighted in accordance with their sampling errors. Indeed, Wilson (2000) espouses what he calls “the meta-mantra: all analysis in meta-analysis is weighted analysis” (p. 427). In this, Wilson (2000) adopts the well-known procedure of weighting effect sizes by the inverse of their squared standard errors (Hedges & Olkin, 1985) and advocates its universal application in meta-analysis.

His exclusive advocacy of weighting led Wilson to disparage specific meta-analyses that used alternative analytical methods. For example, he judged all the significance tests of our meta-analysis on remedial interventions for drink/drive offenders (Wells-Parker *et al.*, 1995a) to be invalid, implying that the conclusions of our research are jeopardized; but this exclusive advocacy of weighted analyses ignores problems associated with weighting and with parametric testing in meta-analysis more generally. Here we highlight some of those neglected problems and our rationale for using unweighted analysis.

Because the sampling error of an effect size is a function of the sample size of the study from which it is drawn, weighting by the inverse of an effect size's variance gives studies with larger sample sizes greater influence in any subsequent analysis. On first blush, giving greater credence to larger studies might seem reasonable. Bear in

Correspondence to: Robert Bangert-Drowns, ED 110, State University of New York at Albany, 1400 Washington Avenue, Albany, New York 12222, USA.

Submitted 14th February 2001; initial review completed 19th February 2001; final version accepted 19th February 2001.

mind, however, that larger samples do not guarantee a more accurate measure of a "true" treatment effect. By chance, a large-sample study may make a poorer measure of a population parameter than a study with fewer subjects. All things being equal, larger samples are preferred because they make precise estimation more likely, but precision is not a necessary consequence of larger samples in any given instance.

More importantly, all things are rarely equal among studies, even replications, on a given topic. There is no reason to believe that a study of remedial treatment for driving-under-the-influence (DUI) offenders with 5000 subjects will be necessarily "better" than a study with 1000, or 500 or 100 subjects. Studies vary on numerous dimensions in addition to sample size, and a thoughtful reviewer may value the conclusions of even a small study if it meets other important criteria. For example, it is not difficult to imagine that larger studies might be more poorly controlled than smaller studies because data might be collected inconsistently over different sites, or extensive treatment implementations might be of dubious fidelity with planned intervention, or researchers might be more influenced by funding pressures and political interests. Similarly, it is possible that one study might implement a more intensive or theoretically consistent intervention than another; should the former necessarily be given less weight if its sample size is smaller than the latter?

Even though treatment characteristics and methodological qualities may be just as important as sample size, if not more so, in determining the value of a particular study finding, important study characteristics are often invisible to the meta-analyst because they go unreported in the original documents. In our meta-analysis of DUI intervention, we paid extraordinary attention to the investigation of study quality (Wells-Parker *et al.*, 1995a; Bangert-Drowns, Wells-Parker & Chevillard, 1997). Our efforts to measure research quality were hampered by the absence of important information in original documents and restricted range of several important methodological characteristics when they were reported.

The potentially negative consequence of weighting effect sizes by within-study sampling error is more likely to occur in cases where the range of sample sizes is quite large. For example, in our analysis of recidivism effects for DUI

remedial interventions (Wells-Parker *et al.*, 1995a), study sample sizes range from 60 to over 150 000 subjects. Obviously, weighting study effect sizes by their inverse variances could give considerable advantage to very large studies with thousands of subjects. In such a group of studies, information from the larger studies could overwhelm information drawn from the smaller ones in subsequent averages and analyses, diminishing the potentially important contributions of findings from smaller studies.

We are not the only meta-analysts to recognize the potential distortion that can be produced by weighting effect sizes from very large studies. Lipsey (1994), in his meta-analysis of juvenile delinquency treatment effects, Windsorized his sample sizes at 300 per treatment group "to prevent a few very large studies from dominating the results" (p. 96). He was concerned with study samples that ranged from 10 to 1000 subjects (median in the range of 101–150 subjects), much smaller than the samples in our studies.

In his discussion of strategies for the analysis of variation among study effect sizes, Wilson (2000) again follows the pioneering work of Hedges & Olkin (1985) and advocates the homogeneity statistic, Q . In essence, Q is distributed as a chi-square and represents a comparison of variation among effect sizes to what might be expected from their sampling error. What happens to Q when the sample sizes in the studies are large? Hedges & Olkin warn, "When the sample sizes are very large, ...rather small differences [among effect sizes] may lead to large values of the test statistic" (p. 123). The authors do not define "small" and "large" but do refer to samples of 10 or more subjects per group as "moderate-sized" (p. 124). In Wells-Parker *et al.* (1995a), the median sample size for studies that reported recidivism effects was 1145, with correspondingly low sampling errors. Among the studies judged to be of sufficient quality for detailed analysis, only five studies of recidivism effects had samples less than 200, and in our final regressions we included only studies judged to be of sufficient quality and possessing more than 100 subjects. Needless to say, in such a case, the Q statistic will be large simply as a consequence of the typically large samples; in fact, we calculated a Q of 10 836. When we Windsorized sample sizes to 300 per comparison group (although the consequences of Windsoriz-

ing in Q values is unexplored, especially Windsorizing to a value lower than the median), the Q statistic was still 883. When we removed studies of poorer quality and Windsorized the sample sizes, the Q statistic was 60. All of these values would have been statistically significant and suggested enormous heterogeneity, but it is unclear from the Q statistic alone if the heterogeneity reflected much more of practical importance than variation in sample size.

Wilson (2000) expresses surprise (p. 423) that not all meta-analyses employ "advanced" meta-analytical method, such as weighted analyses with tests of homogeneity. He contends that all of the significance levels reported in Wells-Parker *et al.* (1995a) are invalid because they are based on unweighted least-squares regression, but the above discussion should show that weighted regression is not a panacea, and there is considerable literature that has suggested caveats with and alternatives to weighted analyses. For example, Hedges & Stock (1983), in a reanalysis of data originally examined using conventional statistics, found that weighted analyses largely reproduced the findings of the conventional analysis. Hedges (1986) noted that weighting by sampling error requires special adjustments for different study designs, such as matched samples, pre-/post-testing and analysis of covariance, adjustments that are rarely employed. Kulik & Kulik (1989) showed that conventional, unweighted analysis of variance of effect sizes was a better analog to analysis of variance of raw data from six studies than was weighted analysis of effect sizes using homogeneity testing. They argue that "within-study variance [employed in homogeneity testing] is not the appropriate variance to use to test the significance of a group factor when studies are a random factor nested within groups" (p. 250).

Hunter & Schmidt (1990) have been pointed in their critique of homogeneity testing. "No other method of meta-analysis is built so squarely on the quicksand foundation of significance testing" (p. 483). They complain that homogeneity testing only attends to sampling error, not the host of other artifactual sources of variance that can add error to an effect size. They fear that its emphasis on statistical testing will distract the meta-analysis from attending to the actual distributions of effect sizes and the theoretical constructs that might describe their variation. They reiterate their con-

cerns in Hunter & Schmidt (1994) and complain that "As long as there are uncorrected artifacts, ... uncorrected artifactual variance will cause the chi-square test [homogeneity tests using the Q statistic] to reject the null hypothesis even when the unattenuated [effect sizes] are identical" (p. 335).

Of course, a more general concern can be raised about the validity of any inferential statistics, weighted or unweighted, assuming fixed or random effects, in meta-analysis. Inferential statistics generally assume random sampling and assignment to conditions. Sampling and assignment are never random in meta-analysis at either the level of subjects or studies. It is unclear to what populations of persons or studies valid generalizations are to be made. In part because of this lack of specification in population, Rubin (1990, 1992, 1993) complained that most meta-analysis is "literature synthesis", whose inferences about the characteristics of a poorly defined population is distorted by generalization from a limited set of treatment and methodological factors. He suggested that meta-analysis turn instead to extrapolating response surfaces, attempting to best estimate the relationship of effect size and an explicit range of scientifically relevant treatment factors under hypothesized "ideal" methodological conditions. Glass (1991) praised this altered conception of meta-analysis: "Meta-analysis should be conceptualized as building and extrapolating response surfaces, not as surveying the literature. The literature, published or not, on any topic is a huge unbalanced survey; coding it, measuring it, and describing it statistically can result in little more than a description of research customs and habits" (p. 1142). (Glass (1999) recently reiterated his praise of Rubin's suggestion and argued that meta-analysis should be replaced with Internet-accessible archives of raw data to allow ideal estimation of "complex data landscapes".)

In our opinion, meta-analysis as literature review is an enormously useful, but primarily exploratory tool. Its expanded perspective, explicit method and synthetic analysis can address questions that any individual study cannot. However, there are numerous factors that operate against the use of meta-analysis to derive precise measures of treatment effects. Meta-analysis is fundamentally dependent on authors of primary research and publication and archiving processes to allow a transparent view of the research enter-

prise. However, even if research reports were transparent, the research enterprise is biased by all sorts of political, financial and scientific factors. Publication and archiving processes can further cull available data, and authors themselves introduce, through error or omission, inaccuracies into the scientific record. Even in the impossible case of comprehensive, errorless and unbiased primary research and dissemination, the meta-analytical process is fraught with human judgement. What data resources to search, what studies to include, what characteristics to code for analysis, how effect sizes are calculated and how effect sizes are analyzed and interpreted are uncertain decisions about which any two reviewers may disagree (Bangert-Drowns, 1997). Current forms of meta-analysis are most useful for explicitly exploring patterns in an extant corpus and for suggesting areas of relative consensus and debate to policy makers and researchers (Wells-Parker *et al.*, 1995b).

We consciously decided to use conventional (“unweighted”) statistics in Wells-Parker *et al.* (1995a) for several reasons. First, there are the various difficulties in applying weighted analyses and homogeneity testing to studies of varied and typically large sample sizes, as outlined above. Secondly, unweighted analyses give every effect size equal value, regardless of sample size. Given unknown differences in characteristics of various studies with no clear evidence that characteristics were related to sample size, equal weight to each study is not an unreasonable analytic strategy. Thirdly, unweighted analyses have the advantage of being more transparent for readers; findings are clearly related to variation in the magnitude in effect size and not clouded by differences in sample size. Fourthly, compared to analyses weighted by the sampling errors of each effect size, unweighted analyses are likely to be more conservative, indicating fewer significant findings, not more. If there is error, the error is likely to be to claim too little.

Although Wilson (2000) may find our significance tests invalid, we think the question of weighted or unweighted analyses (or the use of inferential statistics in general) in meta-analysis is something about which reasonable people can disagree, and unweighted analyses have advantages to recommend them. We also feel confident that our essential findings in Wells-Parker *et al.* (1995a) are robust: that DUI remediation effects are typically small but probably

underestimated in the literature and that combined treatments are likely to be most helpful in reducing DUI recidivism. In light of these issues, we suggest that mantras are unnecessary in the conduct of meta-analysis, which is better served by thoughtful examination of the nature of the data in hand.

References

- BANGERT-DROWNS, R. L. (1986) A review of developments in meta-analytic method, *Psychological Bulletin*, 99, 388–399.
- BANGERT-DROWNS, R. L. (1995) Misunderstanding meta-analysis, *Evaluation in the Health Professions*, 18, 304–314.
- BANGERT-DROWNS, R. L. (1997) Some limiting factors in meta-analysis, in: BUKOSKI, W. J. (Ed.) *Meta-analysis of Drug Abuse Prevention Programs*, monograph 170, pp. 234–252 (Washington, DC, National Institute on Drug Abuse Research).
- BANGERT-DROWNS, R. L., WELLS-PARKER, E. & CHEVILLARD, I. (1997) Assessing methodological quality in narrative reviews and meta-analyses, in: BRYANT, K., WINDLE, M. & WEST, S. G. (Eds) *The Science of Prevention: methodological advances from alcohol and substance abuse research*, pp. 405–429 (Washington, DC, American Psychological Association).
- GLASS, G. V. (1976) Primary, secondary, and meta-analysis research, *Educational Researcher*, 5, 3–8.
- GLASS, G. V. (1991) Review of the book *The Future of Meta-analysis*, *Journal of the American Statistical Association*, 86, 1141–1142.
- GLASS, G. V. (1999) Meta-analysis at 25. Address presented to the Office of Special Education Programs Research Project Directors’ Conference, July, Washington DC, USA. <http://glass.ed.asu.edu/genepapers/meta25.html>
- HEDGES, L. V. (1986) Issues in meta-analysis, in: ROTHKOPF, E. Z. (Ed.) *Review of Research in Education*, vol. 13, pp. 353–398 (Washington, DC, American Educational Research Association).
- HEDGES, L. V. & OLKIN, I. (1985) *Statistical Methods for Meta-analysis* (San Diego, Academic Press).
- HEDGES, L. V. & STOCK, W. (1983) The effects of class size: an examination of rival hypotheses, *American Educational Research Journal*, 20, 63–85.
- HUNTER, J. E. & SCHMIDT, F. L. (1990) *Methods of Meta-analysis: correcting error and bias in research findings* (Newbury Park, Sage Publications).
- HUNTER, J. E. & SCHMIDT, F. L. (1994) Correcting for sources of artifactual variation across studies, in: COOPER, H. & HEDGES, L. V. (Eds) *The Handbook of Research Synthesis*, pp. 233–336 (New York, The Russell Sage Foundation).
- KULIK, J. A. & KULIK, C.-L. C. (1989) Meta-analysis in education, *International Journal of Educational Research*, 13, 221–340.
- LIPSEY, M. W. (1994) Juvenile delinquency treatment: a meta-analytic inquiry into the variability of effects,

- in: T. COOK, T., COOPER, H., CORDRAY, D. S., HARTMANN, H., HEDGES, L. V., LIGHT, R. J., LOUIS, T. A. & MOSTELLER, F. (Eds) *Meta-analysis for Explanation: a casebook*, pp. 83–128 (New York, The Russell Sage Foundation).
- RUBIN, D. B. (1990) A new perspective on meta-analysis, in: WACHTER, K. W. & STRAFF, M. L. (Eds) *The Future of Meta-analysis*, pp. 155–165 (New York, The Russell Sage Foundation).
- RUBIN, D. B. (1992) Meta-analysis: literature-synthesis or effect-size surface estimation? *Journal of Educational Statistics*, 17, 363–374.
- RUBIN, D. B. (1993) Statistical tools for meta-analysis: from straightforward to esoteric, in: BLANCK, P. D. (Ed.) *Interpersonal Expectations: theory, research, and applications*, pp. 400–417 (New York, Cambridge University Press).
- WELLS-PARKER, E., BANGERT-DROWNS, R. L., MCMILLEN, R. & WILLIAMS, M. (1995a) Final results from a meta-analysis of remedial interventions with drink/drive offenders, *Addiction*, 90, 907–926.
- WELLS-PARKER, E., BANGERT-DROWNS, R. L. & WILLIAMS, M. (1995b) The past is prologue: determining directions for research on DUI remediation from meta-analysis, *Addiction*, 90, 1595–1601.
- WILSON, D. B. (2000) Meta-analyses in alcohol and other drug abuse treatment research, *Addiction*, 95 (suppl. 3), 419–438.