Disturbed by Meta-Analysis?

KENNETH W. WACHTER

NOT LONG AGO AN ARTICLE IN *Mercury* CALLED "THE lunacy of it all" (*1*) examined published claims of correlations between manifestations of mental illness and phases of the moon. The authors J. Rotton and I. Kelly found different studies centering their statistically significant relationships on different lunar phases. "To deal with this problem," they wrote, "we resorted to *meta-analysis*, which is a statistical procedure for combining results from different studies. In our meta-analysis, we found no evidence for commonly held beliefs about the effects of a full Moon" (*1*, pp. 75 and 95).

More and more scientists from all fields are resorting to "meta-analysis" when they review a body of scientific literature. Glass (*2*) coined the name "meta-analysis" in 1976; it designates research synthesis that uses formal statistical procedures to retrieve, select, and combine results from previous separate studies. A boom in meta-analyses is under way, but the boom is not being universally welcomed. I have friends who would gladly class meta-analysis itself among the forms of lunacy whose relationship to lunar phases Rotton and Kelly were examining. Meta-analysis has waxed rapidly. Should we expect it to wane as rapidly, like the inconstant moon?

My friends' skepticism centers, I think, around four charges. First is the suspicion that "garbage in and garbage out" are here being camouflaged by fancy statistics. Second is the view that an elementary mistake is being made when a potpourri of study outcomes is treated with procedures invented for a statistical sample of experimental outcomes, without the benefit of the controlled conditions, homogeneous measurement scales, and statistical independence that make the procedures valid. Third is a fear that the study of previous studies is being reduced to a routinized task of coding relegated to a research assistant, upping output per author-month by suppressing any role for wisdom. Related to this fear is a fourth feeling, that meta-analysis accommodates itself to a world in which bad science drives out good by weight of numbers.

The first of these charges—the camouflage charge—is surely misdirected. There is very little fancy statistics at all in meta-analysis. The textbooks of the subject are exceedingly basic, down-to-earth, and opposed to all mystification. Examples, my favorites, are those by Hedges and Olkin, by Light and Pillemer, by Rosenthal, and by Wolf (*3*). Nor is there much tendency toward camouflage. Bad studies are exposed more baldly when one tries to extract or compute the size of an effect on a comparable basis study by study than when one relies, as informal research reviewers often do, on authors' own summaries of their conclusions. Authors of meta-analyses are forever deploring poor quality in studies rather than papering it over with calculations.

The second charge—flouting the conditions for rigor—runs counter to the first; it could be interpreted, constructively, as a call for fancier, more flexible and robust statistical procedures. The healthy preoccupation of the leaders of the field with basic scientific method and common sense may have left them slow to take up hard statistical problems like the quantification of design effects and the modeling of nonindependence among studies. But methods tailored to the messiness of the task may well be possible. A precedent is provided by the progress on the so-called "file-drawer problem." If studies tend to remain in the file drawers, unpublished, when they find effects to be statistically insignificant, then the consensus of published studies will tend to exaggerate statistical significance. An index of Rosenthal's, the "fail-safe sample size," is now widely used to gauge this danger, while Iyengar and Greenhouse (*4*) have pressed statistical methods of maximum-likelihood estimation into service for more refined allowances. An admitted problem is being brought under control.

Meta-analysis is exempt from none of the problems of rigor that beset traditional research synthesis. Combining results from experiments on the same phenomenon conducted under different conditions entails a model, formal or informal, and the scientific adequacy of the model is always a crucial question. Assessing the adequacy of a model relative to some universe of discourse is not a matter for "procedures." But the increasing emphasis in statistical practice on sources of systematic error may have ideas to contribute in the long run.

Other problems of this kind are on the agenda. Medical meta-analyses have turned up correlations between the absence of randomized controlled cases and the estimated strength of treatment effects, giving high visibility to the problem of modeling the quality of study designs. The issue of independence is more daunting. Not only do networks of loyalties and shared prior beliefs make studies by separate research teams less than independent, but so do the very cumulative properties of science that we all applaud. Some dependent stochastic process is here waiting to be modeled. Sooner or later an ingenious modeler is bound to take up the challenge, and sounder methods are going to be found.

Not all my friends who press the second charge, however, are going to be satisfied by sounder methods. When a collection of studies is too messy, they see no reason for systematic comparisons. My own feelings about this charge have been affected by an experience of mine reviewing one of the diciest of all exercises in meta-analysis. The question was the short-term effect of school desegregation on blacks' achievement test scores. The known studies numbered 157. The studies with some semblance of controlled design numbered 19. The National Institute of Education commissioned six scholars with different prior views to do separate meta-analyses on these 19 studies. The results are unpublished, but Harris Cooper (*5*) has published studies of the process of belief change among the writers as they carried out their analyses and among readers of their reports. Given the chance to read the six analyses myself, I watched my own beliefs change and compared Cooper's systematic, questionnaire-based findings. The problems of control and comparability are as messy in these analyses as they can ever be. Do the problems vitiate the point of doing meta-analysis?

About a third of the 19 studies had announced small negative effects, most of the rest small positive effects, and the remaining handful moderate positive effects. The noncomparable characteristics of design and context that could be identified from the study reports appeared, upon analysis, to account only tenuously, if at all, for the differences in outcome. Each of the studies had faults, but the faults differed, and I found the impression of a relatively random

The author is professor of demography and statistics, University of California, Berkeley, CA 94720.

scatter of outcomes with an identifiable central tendency hard to shake. Like the authors of the analyses and like most of Cooper's surveyed readers, I found myself led to believe that likely real effects were near the center of the scatter, smaller in magnitude than the prior guesses I and others had offered. Without erasing all differences of opinion, here meta-analysis did induce convergence of views. This is its usual function. Fifty estimated gains do not frequently turn into an overall estimated loss under statistical examination, even though in principle such extrapolation could be sound. Statistical examination usually highlights common ground. It did so with the desegregation studies, despite the palpable problems with each and all of them. My experience with this example indicates to me that, even with extremely problematic data, arraying a collection of studies with differing faults together in a systematic way does have persuasive power. Persuasive power or seductive power? I think persuasive power.

What about my friends' third charge, that routinization is crowding out wisdom? Despite the intentions of the founders, meta-analysis has burgeoned not least because it seems to sanction the use of research assistants in what is the most time-intensive and might otherwise be the most thought-intensive stage of preparing research reviews. Coding is a traditional task for research assistants, and when the study of previous papers is conceived to be a coding enterprise, conducted in accordance with rules that can be prespecified, it is hard to resist leaving research assistants the job. Codable kinds of judgments of study quality tend to become the basis for weighting studies in the analysis. Codable kinds of outcomes tend to be seized on to the exclusion of subtle chains of argument. Reading is devalued as a task. Can we devalue reading without devaluing writing? Can we devalue writing without devaluing thinking? It isn't hard to blame on meta-analysis a whole host of distressing trends.

The serious core of such fears finds expression in the fourth charge, the issue of weight of numbers. Procedures like those of meta-analysis that deal in bulk with scientific findings cannot go deeply into the rigor of individual arguments. But in science a single good study ought to be able to stand against any number of weak ones. Vote counting ought to be irrelevant. It is worse than sad to think this only true "once upon a time."

Meta-analysts worry about the many studies left in file drawers, implying an unrepresentative lot to analyze. Skeptics worry about too many studies not left in wastebaskets, and instead crowding journals, leaving the still small voice of signal to be drowned in noise. To those with a statistical turn of mind, the first is a problem of bias, the second a problem of variance. Both equally deserve attention. But the first is a problem for technical methodology, whereas the second is a problem for the values of the scientific community. As science is organized today, defending standards brings few rewards and many costs.

Those who feel anxious on these grounds, however, make a mistake if they see proponents of meta-analysis as enemies, not allies. Meta-analysts are among the few who actually pressure journal editors to impose standards of rigor and documentation as conditions for publication. Why are they not joined by more of us? The stricter the standards for publication, the lower are the costs forever after of sifting the papers that repay attention from the mass of papers that demand attention; the better are the chances of avoiding bulk-processing of scientific contributions. To those, once again, with a statistical turn of mind, the problem could be called a problem of "Type I" and "Type II" errors. As in so many aspects of society, avoiding Type I errors of rejecting the good has taken almost total precedence over avoiding Type II errors of accepting the bad. However deplorable, weight of numbers does now often tell over quality of argument. When volume grows beyond a point, formalized procedures like those of meta-analysis acquire almost irresistible appeal.

Viewed from this perspective, many of the trends that are making a place for meta-analysis in today's research world *are* disturbing. To this extent, I agree with the skeptics, my friends. But meta-analysis itself is not some lunatic fashion of the 1980s that will quickly wane. If, as I think, it is a symptom of disturbing trends, it is also a response.

REFERENCES

1. J. Rotton and I. Kelly, *Mercury* **15**, 73 (1986).
2. G. V. Glass, *Educ. Res.* **5**, 3 (1976).
3. L. Hedges and I. Olkin, *Statistical Methods for Meta-Analysis* (Academic Press, New York, 1985); F. M. Wolf, *Meta-Analysis: Quantitative Methods for Research Synthesis* (Sage, Beverly Hills, CA, 1986); R. Rosenthal, *Meta-Analytic Procedures for Social Research* (Sage, Beverly Hills, CA, 1984); R. Light and D. Pillemer, *Summing Up: The Science of Reviewing Research* (Harvard Univ. Press, Cambridge, MA, 1984).
4. S. Iyengar and J. Greenhouse, *Stat. Sci.*, in press.
5. H. Cooper, in *Social Psychology Applied to Education*, R. Feldman, Ed. (Cambridge University Press, Cambridge, England, 1986).