Can Meta-Analysis Make Policy?

This statistical technique is being used with increasing frequency to resolve conflicting studies in social science—and in the process overturning much conventional wisdom

In 1970, Richard Light, now a professor at Harvard University's Kennedy School of Government, was asked by the Department of Health, Education, and Welfare to answer a simple question: Does Head Start work? Answering the question wasn't simple, but not for lack of information. On the contrary, Light recalls, no fewer than 13 studies had evaluated the program. The first 12 he examined showed modest positive effects. But the thirteenth study, by far the largest, showed no effect at all. This sequence of events, Light says, reminded him of Mark Twain's remark about the 13th stroke of a clock: It cast doubt not only on itself, but on all that went before.

"I had no idea what to do," he says. "This big collection of studies seemed to come to a morass of conflicting conclusions." Light was dissatisfied with the usual procedure for writing scientific review articles—picking through the studies to find the ones that seem most solid—because it struck him as too subjective. Frustrated at his inability to come up with a solid answer, Light and Paul Smith, now at the Children's Defense Fund, wrote an article for the *Harvard Education Review* in 1971 that called on scientists to come up with more rigorous methods for reviewing research.

Light's dilemma exemplifies the difficulties policy-makers face in trying to translate research into social policy. Often there are scores—even hundreds—of studies, but few seem definitive, most conflict, and many approach the subject from differing angles. What Light did not discover for another few years was that a possible way out of this bind already existed. Now generally called "meta-analysis," it has been employed in various forms since 1904, when the English statistician Karl Pearson grouped data from British military tests to conclude that the then-current practice of vaccination against intestinal fever was ineffective.

Despite its long pedigree, the technique did not become common until the 1980s (*Science*, 3 August 1990, p. 476). Today, meta-analysis is especially prominent in medical research. But the flourishing of meta-analysis in medicine has obscured a similar growth in the social sciences, where the technique was endorsed by the National Research Council in 1992. As was shown at a conference earlier this year sponsored by the Russell Sage Foundation,* social scientists of every stripe have performed hundreds of meta-analyses, with more on the way. Studies of psychotherapy, hospital staffing, juvenile delinquency, the efficacy of funding increases in education—all have been put under the lens of meta-analysis. And in many cases the results, if they are to be believed, have profound implications for social policy.

META-ANALYSIS CONFOUNDS THE EXPERTS

Subject	Conclusion of Expert Review	Meta-Analysis
Psychotherapy	Worthless (Eysenck,1965).	Positive results, but little difference between varying approaches (Smith and Glass, 1977).
Delinquency prevention	Programs have no consistent positive effects (National Academy of Sciences Panel on Rehabilitative Techniques, 1981).	Many programs have modest good effects; skill-oriented, nonpsychologically oriented ones may have more than modest effects. Punitive schemes are counterproductive (Lipsey, in press).
School funding	Surprisingly little direct impact on educational outcome (Hanushek, 1989).	Important to educational outcome (Hedges *et al.*, 1994).
Job training	Effectiveness subject to bitter dispute.	Women show modest positive effects from programs that help find work, men from basic education; current systems do not match people and programs well (Cordray and Fischer, 1994).
Reducing anxiety in surgical patients	Inconclusive, but thought to have little potential for reducing length of stay and costs (Schwartz and Mendelson, 1991).	Inexpensive 30–90 minute preparation sessions can reduce length of stay with sharp impact on costs (Devine, 1994).

A solution ... or more problems?

But can the results of meta-analysis be believed? Meta-analysis remains a controversial method that has provoked dispute in every field to which it has been applied—including the social sciences. To begin with, few researchers have the necessary statistical expertise to conduct or interpret a full-scale meta-analysis. In addition, the technique implicitly rejects the traditional preference for a single, beautifully conducted, absolutely decisive study—a preference, Light says, that has led researchers in many fields to spend their lives reinventing the wheel. "They want their study to kill off the question," he says. "But because a single study can almost never do that, they constantly repeat the same research. You find 58 studies of the same question, and somebody out there is asking for funding to do the 59th."

Meta-analysis also challenges customary views of expertise. Ordinarily, review articles are supposed to be summaries in which great researchers survey their fields, figuratively settling back in their chairs and dispensing the wisdom. Meta-analysis isn't like that. "Instead," says Iain Chalmers, "meta-analysis seems to make some outside statisticians into the big names—and the big names inside the field don't like it. And they especially don't like it when the statisticians tell them their so-called expert judgment is all wrong."

More important, meta-analysis attracts dissent because pooling data from different studies seems to violate a cardinal scientific prohibition: against adding apples and oranges. Indeed, meta-analysts cannot directly combine data from different studies. Instead, they usually look at a statistical measure called the "effect size"—the difference between the result observed by, say, an experimental treatment and what would be expected if the treatment had no effect. In a typical individual study, the results are subjected to standard statistical tests of significance, which reject effect sizes near zero unless the sample size is very large. Meta-analysis, by contrast, looks at the distribution of *all* effect sizes, significant or not. If they are randomly clustered around zero, meta-analysis suggests that the treatment has

*National Conference on Research Synthesis: Social Science Informing Public Policy, 21 June.

no effect. But if they cluster off to one side, meta-analysis shows that something is going on, even if the individual results are not statistically significant in themselves. In this way, meta-analysis can amplify experimental phenomena that are too small to see in single experiments, and it can find a consistent pattern in apparently contradictory results—as demonstrated by a recent meta-analysis of juvenile delinquency prevention.

Juvenile delinquency: Prevention works

Conventional wisdom holds that social programs don't prevent juvenile delinquency, a viewpoint exemplified by sarcastic media treatment of midnight basketball programs in the recent crime bill. Nor is this attitude limited to the lay public; after a series of major research reviews in the 1980s, Ira M. Schwartz of the University of Michigan's Center for the Study of Youth Policy summed up expert opinion in the title of a 1991 article: "Delinquency Prevention: Where's the Beef?" Given the general despair, public opinion has swung toward punitive measures such as boot camps, mandatory sentences, and automatic treatment of juvenile offenders as adults in court.

Mark Lipsey of Vanderbilt University has been revisiting the question, poring through almost 500 controlled studies of delinquency prevention in a meta-analysis. (A preliminary version appeared in *Meta-analysis for Explanation: A Casebook*, published by the Russell Sage Foundation in 1992.) Lipsey has found that the reviews were wrong: Most delinquency programs work, albeit modestly. Typically, the studies compared rates of re-arrest among delinquent youths who entered a program with the rates among those who didn't. On average, Lipsey found a 10% reduction, from about 50% to about 45%. "It doesn't knock your socks off," says Lipsey. "But it's not trivial. There's a lot of communities that would like to see a 10% reduction in delinquency." In Lipsey's view, re-arrest is such a poor measure of outcome—"It measures police behavior, more than anything"—that the true benefits are probably understated. "But let's get it settled," he says. "Some things work! The question is, what works and why?"

Lipsey tried to answer that question in his meta-analysis by dividing programs into three categories: structured and behavioral (job training and behavior modification, for example); insight-oriented (family therapy, rap groups, and the like); and deterrent (such as shock incarceration and "scared straight" programs). Behavioral programs seem the most successful; many lead to impressive 20% to 30% reductions in re-arrest. "They tend to be structured environments that rely on teaching things, not psychodynamic insights," Lipsey says. On the other hand, he says, "when you do insight counseling for juvenile delinquents, you get more insightful juvenile delinquents. That's not bad by itself, but it's not where you want to put your tax dollar."

The worst performers were deterrent programs—"get tough, straighten them out, scare them away from a life of crime," Lipsey explains. "You average those out and you get a negative effect." Lipsey surmises that this is because "many of these programs take moderately impressionable hypermacho teenage kids and expose them to flamboyant models of abusive behavior, whether it's in the lifer mode or the drill-sergeant mode. In any case, it's a riveting image of some colorful personalities. I'm not sure that modeling verbally and physically abusive behavior for them is the best idea." Lipsey "guesses" that the current vogue for boot camps falls into this category, though the programs are too new for there to be much data either way.

Meta-analysis vs. vote-counting

By assembling a large number of studies, Lipsey hoped to avoid the data-quality problems that are an inevitable issue in the social sciences. But this cannot always be avoided, as shown by the debate over a meta-analysis of school funding. "People want solutions to the problems of the schools," says Eric A. Hanushek, an economist at the University of Rochester whose book *Making Schools Work* was published this month by the Brookings Institute. Typical reforms, he says, include lowering class size and raising teacher salaries, both of which require large funding increases. "Naturally," he says, "you want to know whether those ways of increasing resources matter"—that is, whether they raise student performance. In 1989, Hanushek reviewed 38 studies and found the "startlingly consistent" result that "there is no strong or systematic relationship between school expenditures and student performance." Somewhat to Hanushek's dismay, conservative critics widely publicized his work with the slogan "money doesn't matter."

Hanushek's review used a technique called "vote-counting"—he performed statistical regressions and tallied the studies that were positive and statistically significant. Because these were fewer in number than negative or nonsignificant studies, Hanushek concluded that school spending is not clearly related to educational performance. But vote-counting has come under fire. Despite its intuitive appeal, Larry V.

Is the Vote Count Biased?

Though researchers seldom admit it, says Stanford University statistician Ingram Olkin, the basis for traditional reviews of conflicting studies is often "vote-counting": Each study with statistically significant findings gets one vote for or against the hypothesis. But, as Olkin and Larry Hedges of the University of Chicago demonstrated in 1980, vote-counting is plagued by a fundamental problem: As the number of studies increases, the chance of a specific statistical error soars.

Statisticians divide errors into Type I and Type II. A Type I error lies in concluding that research has found an association or effect when one does not exist; Type II implies concluding there is no association or effect when one exists. When scientists say that their results have less than a 5% chance of being due to random fluctuations, they are almost always speaking of the possibility of Type I error. Type II error is frequently ignored.

It shouldn't be, Hedges and Olkin argue. Suppose researchers are trying to learn the effects of a drug to reduce blood cholesterol. They have conducted many studies, administering the drug to half the people in each study and using the other half as a control group. Suppose further that the actual effect of the drug is to reduce the cholesterol level by half a standard deviation—a statistician's way of saying that almost 70% of the people who take the drug have lower cholesterol levels than the mean of the people who do not (assuming that both groups are distributed along a bell curve).

To avoid Type I error, typical statistical tests look for lower than expected cholesterol levels. Much lower, in fact—two-thirds or three-quarters of a standard deviation, depending on the test. But if the real effect is smaller than that, most experiments will find smaller reductions. Studies that find the true effect—half a standard deviation—will be interpreted as finding no effect; only studies that found much more than the correct value will be interpreted as showing an effect. As the number of studies increases, it becomes less and less likely that a large proportion of them will find these unrealistically large reductions, thus diminishing the chance that the real effect will be observed. Given the preponderance of "no effect" experiments, in this case a traditional vote-counting review would report that the drug did little good—a Type II error.

–C.M.

The Power of Positive Thinking

Are the social sciences suited for meta-analysis? The question is raised by a startling review of meta-analyses in the social sciences by political scientists Mark Lipsey and David B. Wilson of Vanderbilt University in an article in last December's *American Psychologist*. Their "meta-meta-analysis" collected 302 meta-analyses on the efficacy of psychological, educational, and behavioral treatments. Astonishingly, only six found negative effects; fewer than 10% found negligible effects. Lipsey and Wilson were unable to explain this strong, unexpectedly positive showing as due to bias, oversampling, or poor data quality.

The figures surprised everybody, including Lipsey and the editors of *American Psychologist*, which will publish an array of comments this winter. To Eric Hanushek of the University of Rochester, the study demonstrates the inadequacy of meta-analysis for social science. "When statistics tell us something that stands in the face of common sense," he says, "it's time to abandon statistics." Harvard University education researcher Richard Light draws a different conclusion. "Putting on my optimist's hat," he says, they may be evidence "that we really have learned something in these fields in the last 20 years."

"At first I said, 'Come on, everything can't work,' " Lipsey recalls. "But then I realized that the kind of interventions that would end up in a meta-analysis would have to be fairly mature and widespread to attract the quantity of research necessary for meta-analysis. On the whole continuum of psychologically-based interventions, this is a very established, very mature set. ... I would by no means generalize to say that any psychological intervention has a chance of working. But the subset that survives the weeding out and editing is your tried and tested ones."

Not so, says Gene V. Glass, the education researcher at Arizona State University who coined the term "meta-analysis" in 1976. The efforts of behavioral scientists "are part of the common stock of understanding and people trying to help," he says. "Even if they are based on completely ineffective theories, they probably still do some good." Lipsey's analysis, in Glass's view, should not be taken as "supporting the state of social science theory."

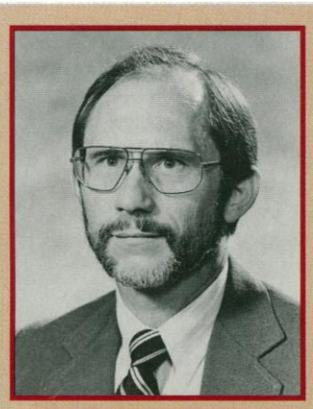
Indeed, Glass suggests, the Lipsey-Wilson study may be taken as evidence of the debility of social science. "You have mildly positive effects almost no matter what you do," he says. "That tells me you are measuring the effects of intervention alone, almost regardless of the context. There's no field of social science that seems to add anything to that, which is discouraging."

–C.M.

Hedges, now at the University of Chicago, and Ingram Olkin of Stanford University demonstrated in 1980 that vote-counting is increasingly unlikely to detect real positive effects as the number of studies increases (see box on p. 961). "It's paradoxical," Hedges admits. "The more information you have, the worse you do."

Last April, Hedges and two colleagues sought to demonstrate the point in *Educational Researcher* when they reviewed the same studies reviewed by Hanushek. Rather than vote-counting, though, they used meta-analysis. Just as Hedges and Olkin had argued, they found systematic positive effects that Hanushek's vote-counting had missed. Indeed, decreased class size, increased teacher experience, increased teacher salaries, and increased per-pupil spending were all positively related to academic performance. "Money does matter after all," they concluded.

Hanushek says his work has been caricatured as an argument that money makes no difference in any circumstances. Furthermore, he says, the Chicago meta-analysis did not demonstrate an error in his analysis, because the technique is not well suited to social science. As evidence, he points out that the Chicago researchers had to throw out 30% to 40% of the data because those data reported no significant effect—without reporting whether the insignificant effect was positive or negative, as needed in a meta-analysis. "It's not that it wasn't information," Hanushek says; "it just didn't fit their form of meta-analysis. The way they did that led to specific biases in their results, because they ended up with very selective sampling."



Meta-analysis shows some forms of juvenile delinquency prevention are effective: "Some things work!"

—Mark Lipsey

Statisticians ascendant

Many social scientists other than Hanushek also criticize the use of meta-analysis in the social sciences. Yet proponents of the method argue that whatever problems there are in the technique will have to be dealt with, because there is no other way to handle the explosion of data. According to Hedges, the world's researchers have produced 100,000 studies of depression. Yet new proposals continue to appear. "Is this a sensible situation?" he asks. "Do we really need more data?"

According to Frank L. Schmidt of the University of Iowa, the inability of a single study to definitively answer research questions demands the use of meta-analysis, and the result of its increased use, he predicts, could transform research in the behavioral sciences. As more discoveries are made by people who do not conduct primary research, research may split into two tiers, one group specializing in the conduct of individual studies, the other applying complex meta-analytic techniques to make scientific discoveries. "Such a structure raises troubling questions," Schmidt wrote in 1992. "How would these two groups be rewarded? What would be their relative status in the overall research enterprise? ... Is it the wave of the future?" As meta-analysis is exploited to analyze an increasing number of issues in the social sciences, the answer to that question should soon become clear.

–Charles C. Mann

Charles Mann is a science writer living in western Massachusetts.

Additional Reading

T. D. Cook *et al.*, *Meta-analysis for Explanation: A Casebook* (Russell Sage, New York, 1992).

H. M. Cooper and L. V. Hedges, *The Handbook of Research Synthesis* (Russell Sage, New York, 1992.)

National Research Council, *Combining Information: Statistical Issues and Opportunities for Research* (National Academy Press, Washington, D.C., 1992).

F. L. Schmidt, "What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology," *American Psychologist* **47**, 1173 (1992).