The Promise of Meta-Analysis for Our Schools: A Q&A with NEPC Fellow Gene V Glass

Meta-analysis is a statistical technique that combines data from multiple studies on the same topic to explore trends. In today's education world, the approach is often associated with Australian education professor John Hattie, who is best known for his popular 2009 book, *Visible Learning*. Hattie synthesized the results of multiple meta-analyses to provide guidance on what influences student achievement. But long before Hattie came Gene Glass, whom the Oxford English Dictionary credits with coining the term in 1976 in connection with his work on psychotherapy. In the Q&A below, Glass explains and reflects on meta-analysis and its uses and abuses inside and outside of education. Glass is a Research Professor at the University of Colorado Boulder and a Fellow and senior researcher at the National Education Policy Center. A Regents Professor Emeritus at Arizona State University, he is also a Lecturer at San José State University. Glass has won multiple honors for his work, including the Palmer O. Johnson award of the American Educational Research Association, as well as the AERA's career award for Distinguished Contributions to Research. Trained as a statistician, he is an expert in psychotherapy research, evaluation methodology, and policy analysis.

Q: What is a meta-analysis? How, if at all, are meta-analyses useful for teachers, administrators, policymakers, journalists, and others outside of academia?

A: *Meta-analysis is a statistical technique to deal with the problem of extracting meaning from multiple studies of the same question. What is the correlation of SAT scores with freshman GPA? Can tutoring increase SAT-V and SAT-Q scores? How can we determine what these 25 studies of the question have to say?*

Q: When and why did researchers start conducting meta-analyses?

A: *Research in the soft sciences (viz., certain areas of psychology, sociology, and all of the minor disciplines like education, social work, business, nursing, and the like) exploded in the 1960s. Whereas before, one really well-done study on the effectiveness of Rogerian psychotherapy seemed to settle the matter, by the 1970s a few dozen outcome experiments competed for attention. Their findings were inconsistent to a greater or lesser extent. Their message was unclear.*

Q: How do meta-analyses differ from other types of research summary, research synthesis, or literature review?

A: *Narrative reviews, like those that populated journals such as Psychological Bulletin or the Review of Educational Research – both of which I edited in the 1970s, incidentally – were attempts to coalesce findings of multiple studies. They relied heavily on notions of "statistical significance," and they largely failed to reach a conclusion. Classic statistical significance in the soft sciences is attained by taking large samples; it's really that simple. Studies with large Ns – numbers of subjects, or observations – achieve statistical significance; studies with small Ns do not. Statistically significant results may not be of any practical significance. Paul Meehl called collections of significance tests "empirical power curves," i.e., worthless displays of which studies had large Ns and which did not.*

As research areas grew and studies on a single question could number in the dozens or even hundreds, attempts to discern trends in large masses of findings usually resulted in confusion. Prior to the introduction of meta-analysis, and regrettably too often afterwards, the typical research review ended with a call for more and better research, the vain search for the perfect study. But the perfect study never comes, and the mass of undigested study findings just lay there waiting to be analyzed.

Meta-analysis represents a simple change in perspective. Statistical methods aim to derive meaning from collections of data that in their individuality are uninterpretable or confusing. All 365 high-temperature readings for 2015 in Bismarck, North Dakota, reveal little; but 12 monthly averages displayed on graph paper paint a clear picture. The calculation of the correlation of SAT and GPA for 500 freshmen is what we call "primary analysis." The calculation of the average of 20 correlation coefficients from different studies of SAT & GPA correlation is called "meta-analysis." The findings of multiple studies are data for a meta-analysis, just as the data points in a single study are data for a primary analysis.

Q: What are the pros and cons of using a meta-analysis to summarize a body of research?

A: *People who disliked the findings of some early meta-analyses – particularly those in the field of psychotherapy outcome research – thought they spotted a fatal flaw in the approach. "You can't compare the results of studies unless the studies are the same." The Apples & Oranges problem, they called it.*

The meta-analysis critics had fervor and indignation on their side. One labeled it "meta-silliness." Unfortunately, they had no ally in logic. The consideration of these critics' objection led me eventually to a central problem in modern philosophy: the identity problem.

The first 100 or so pages of Robert Nozick's magnum opus Philosophical Explanations deal with the question, what does it mean to say that two things are identical? The very assertion that A and B are identical is self-contradictory, since two things that are identical are the same thing, hence there are not two things. If experiment A and experiment B are "the same," then there is no need to coalesce their findings because their findings will have to be the same.

I won't get into the details of how Nozick resolves the identity problem, but I will say that meta-analysis resolves the Apples & Oranges problem as he would have. The findings of studies A, B, C, D, etc. are arrayed and their variation as a function of characteristics X, Y, and Z are analyzed. For example, the 50 correlations of SAT and GPA for Males and the 65 correlations for Females are recorded and it is seen if they differ. If they do not, or if the correlations appear to differ very little, they might be averaged. Whether the SAT-GPA correlation question shows different answers across the mediating variable Sex, or Type of University, is not an a priori question; it is an empirical question answered by statistical analysis, or meta-analysis in this case.

Q: In recent years, John Hattie's syntheses of meta-analyses have grown wildly popular with K-12 practitioners and others. Do you have advice on how Hattie's work can best be understood and used?

A: *Hattie's work has been unfairly criticized, most inappropriately by Robert Slavin. Slavin claimed that, "The essential problem with Hattie's meta-meta-analyses is that they accept the results of the underlying meta-analyses without question. Yet many, perhaps most meta-analyses accept all sorts of individual studies of widely varying standards of quality." Well, Slavin is just flat wrong about this. Many meta-analyses have shown that distinctions between "good" and "bad" studies have proven to be irrelevant in accounting for differences in the results. As heretical as that may sound, it is nonetheless true. I suspect it arises from the fact that most collections of studies are not composed of "good" and "bad" studies, but of studies that can be classified as "good," "better," and "best."*

Hattie's contribution to discussions about education policy is that his work suggests where teachers and other might look to try to improve teaching and learning. All of education research fails to give directives for individual action. Rather, it illustrates perspectives one can take for making sense of individual experience. It's not that Duckworth's research on Grit tells any teacher what to do. It's that "persistence" and "resilience" might be one useful way for teachers to look at their students.

The reality of meta-analyses in education is that the findings of studies on a single topic like Feedback or Peer Tutoring can vary greatly. I have repeatedly observed that the effects of an intervention in teaching and learning vary substantially around their average. One study of Peer Tutoring might show a large benefit to student achievement and the next study might show a very small benefit or none at all. In this case, the take-away message is not that Peer Tutoring has an average benefit of .60 sigma; it's that Peer Tutoring, a promising intervention, can be done well or poorly. Good luck seeking the way to do it well.

Hattie's results, like the results of meta-analyses themselves, result in part from the under-

lying choice of outcome measures, which are often test scores with their attendant strengths and weaknesses. So, for example, an intervention that involves practicing taking tests may show big effects on outcomes that are little more than tests like those they have practiced. Some seemingly impressive interventions in Hattie's lists likely fall into that category.

Q: What are some common uses and abuses of the meta-analysis approach?

A: *Given that even single education interventions show extremely variable benefits, the premium in any meta-analysis is to discover the conditions under which the benefits are big or small. Peer tutoring might work well with tutors older than 13, but not as well or at all with younger tutors. Ignoring this fact or failing even to investigate it is to fall short of one's objective to produce practical knowledge.*

This is why meta-analysis has had a checkered history in education while having enjoyed great success in medicine. For every educator who says, "Meta-analysis is garbage-in-garbage-out," there are 10 MDs who say, "Yes, I learned about meta-analyses in med school, and I rely on their findings in my specialty." The difference is that meta-analyses in medicine have often shown consistent results across studies (e.g., clinical trials of a new drug) while meta-analyses in education do not. Surely this arises from the fact that interventions in medicine (e.g., intravenous injection of 10 mg of Nortriptyline) are uniform and well defined whereas even interventions that carry the same label in education are subject to substantial variation from place to place, or time to time. Giving students Feedback can take many different forms, some of which are effective and some of which are not.

Q: How would you like to see the approach used in the future?

A: *Some of the shortcomings of meta-analysis applied to the soft sciences could be overcome if those synthesizing the results of multiple studies had access to the raw data from those studies. Primary statistical analyses often obscure mediating relationships that might someday prove to be crucial. It is now well known, for example, that certain stimulants like caffeine will calm prepubescent children while they hype up postpubescent children. Studies that ignored the mediating variable Puberty and averaged across a wide range of ages have lost valuable knowledge that might be recovered if the original data were available to meta-analysts. So much could be learned by secondary analyses of original study data. Fortunately, the situation in medical research is far ahead of that in the soft sciences. "Numerous organizations now recommend or require raw data to be made available, including the International Committee of Medical Journal Editors, which recently proposed that clinical trial data sharing be a 'condition of ... publication.'"*

NEPC Resources on Research Issues