HOW CAN YOU TELL WHEN THE FINDINGS OF A META-ANALYSIS ARE LIKELY TO BE VALID?

ROBERT SLAVIN, DIRECTOR OF RESEARCH AT JOHN HOPKINS UNIVERSITY

AUTHOR & CO-AUTHOR OF OVER 300 ARTICLES ON EDUCATIONAL PSYCHOLOGY FROM 1975 – 2020; AUTHOR & CO-AUTHOR OF 24 BOOKS

In Baltimore, Faidley's, founded in 1886, is a much-loved seafood market inside Lexington Market. Faidley's used to be a real old-fashioned market, with sawdust on the floor and an oyster bar in the center. People lined up behind their favorite oyster shucker. In a longstanding tradition, the oyster shuckers picked oysters out of crushed ice and tapped them with their oyster knives. If they sounded full, they opened them. But if they did not, the shuckers discarded them.

I always noticed that the line was longer behind the shucker who was discarding the most oysters. Why? Because everyone knew that the shucker who was pickier was more likely to come up with a dozen fat, delicious oysters, instead of say, nine great ones and three…not so great.

I bring this up today to tell you how to pick full, fair meta-analyses on educational programs. No, you can't tap them with an oyster knife, but otherwise, the process is similar. You want meta-analysts who are picky about what goes into their meta-analyses. Your goal is to make sure that a meta-analysis produces results that truly represent what teachers and schools are likely to see in practice when they thoughtfully implement an innovative program. If instead you pick the meta-analysis with the biggest effect sizes, you will always be disappointed.

As a special service to my readers, I'm going to let you in on a few trade secrets about how to quickly evaluate a meta-analysis in education.

One very easy way to evaluate a meta-analysis is to look at the overall effect siz One very easy way to evaluate a meta-analysis is to look at the overall effect size, probably shown in the abstract. If the overall mean effect size is more than about +0.40, you probably don't have to read any further. Unless the treatment is tutoring or some other treatment that you would expect to make a massive difference in student achievement, it is rare to find a single legitimate study with an effect size that large, much less an average that large. A very large effect size is almost a guarantee that a meta-analysis is full of studies with design features that greatly inflate effect sizes, not studies with outstandingly effective treatments.

Next, go to the Methods section, which will have within it a section on inclusion (or selection) criteria. It should list the types of studies that were or were not accepted into the study. Some of the criteria will have to do with the focus of the meta-analysis, specifying, for example, "studies of science programs for students in grades 6 to 12." But your focus is on the criteria that specify how picky the meta-analysis is. As one example of a picky set of critera, here are the main ones we use in Evidence for ESSA and in every analysis we write:

1. Studies had to use random assignment or matching to assign students to experimental or control groups, with schools and students in each specified in advance.
2. Students assigned to the experimental group had to be compared to very similar students in a control group, which uses business-as-usual. The experimental and control students must be well matched, within a quarter standard deviation at pretest (ES=+0.25), and attrition (loss of subjects) must be no more than 15% higher in one group than the other at the end of the study. Why? It is essential that experimental and control groups start and remain the same in all ways other than the treatment. Controls for initial differences do not work well when the differences are large.
3. There must be at least 30 experimental and 30 control students. Analyses of combined effect sizes must control for sample sizes. Why? Evidence finds substantial inflation of effect sizes in very small studies.

4. The treatments must be provided for at least 12 weeks. Why? Evidence finds major inflation of effect sizes in very brief studies, and brief studies do not represent the reality of the classroom.
5. Outcome measures must be measures independent of the program developers and researchers. Usually, this means using national tests of achievement, though not necessarily standardized tests. Why? Research has found that tests made by researchers can inflate effect sizes by double, or more, and research-made measures do not represent the reality of classroom assessment.

There may be other details, but these are the most important. Note that there is a double focus of these standards. Each is intended both to minimize bias, but also to maximize similarity to the conditions faced by schools. What principal or teacher who cares about evidence would be interested in adopting a program evaluated in comparison to a very different control group? Or in a study with few subjects, or a very brief duration? Or in a study that used measures made by the developers or researchers? This set is very similar to what the What Works Clearinghouse (WWC) requires

If these criteria are all there in the "Inclusion Standards," chances are you are looking at a top-quality meta-analysis. As a rule, it will have average effect sizes lower than those you'll see in reviews without some or all of these standards, but the effect sizes you see will probably be close to what you will actually get in student achievement gains if your school implements a given program with fidelity and thoughtfulness.

What I find astonishing is how many meta-analyses do not have standards this high. Among experts, these criteria are not controversial, except for the last one, which shouldn't be. Yet meta-analyses are often written, and accepted by journals, with much lower standards, thereby producing greatly inflated, unrealistic effect sizes.

As one example, there was a meta-analysis of Direct Instruction programs in reading, mathematics, and language, published in the *Review of Educational Research* (Stockard et al., 2016). I have great respect for Direct Instruction, which has been doing good work for many years. But this meta-analysis was very disturbing.

The inclusion and exclusion criteria in this meta-analysis did not require experimental-control comparisons, did not require well-matched samples, and did not require any minimum sample size or duration. It was not clear how many of the outcomes measures were made by program developers or researchers, rather than independent of the program.

With these minimal inclusion standards, and a very long time span (back to 1966), it is not surprising that the review found a great many qualifying studies. 528, to be exact. The review also reported extraordinary effect sizes: +0.51 for reading, +0.55 for math, and +0.54 for language. If these effects were all true and meaningful, it would mean that DI is much more effective than one-to-one tutoring, for example.

But don't get your hopes up. The article included an online appendix that showed the sample sizes, study designs, and outcomes of every study.

First, the authors identified eight experimental designs (plus single-subject designs, which were treated separately). Only two of these would meet anyone's modern standards of meta-analysis: randomized and matched. The others included pre-post gains (no control group), comparisons to test norms, and other pre-scientific designs.

Sample sizes were often extremely small. Leaving aside single-case experiments, there were dozens of single-digit sample sizes (e.g., six students), often with very large effect sizes. Further, there was no indication of study duration.

What is truly astonishing is that *RER* accepted this study. *RER* is the top-rated journal in all of education, based on its citation count. Yet this review, and the Kulik & Fletcher (2016) review I cited in a recent blog, clearly did not meet minimal standards for meta-analyses.

My colleagues and I will be working in the coming months to better understand what has gone wrong with meta-analysis in education, and to propose solutions. Of course, our first step will be to spend a lot of time at oyster bars studying how they set such high standards. Oysters and beer will definitely be involved!