*Timothy Shanahan | Editor*

# Why You Need to Be Careful About *Visible Learning*

In my last column, I advocated the notion that instructional decisions in reading should be guided by the results of meta-analysis. The basic idea was that meta-analysis summarizes the results of studies that evaluate the effectiveness of instruction. Meta-analyses are also the best place to go to settle instructional arguments because they are based on evidence drawn from multiple studies. It seems more likely that we could make something work in our schools if it has worked over and over again for others.

Not surprisingly, scientists cite meta-analyses more often than single studies (e.g., Carlson & Ji, 2011), and policymakers heavily depend on them, too.

That may sound like I'm writing a blank check for meta-analysis: If a meta-analysis says it's so, then it must be true. I can't go that far. If I did, I'd lose my plaque in the Curmudgeons Hall of Fame.

I use meta-analysis a lot when determining what works, and I think you should, too, but there are some things you need to know if you are going to be an educated consumer of this kind of research.

Recently, a colleague told me that she had found a shortcut to applying research to literacy instruction. I was intrigued. Unfortunately, what she described was as likely to mislead her instructional practices as to inform them. It simply isn't enough to read the summaries in Hattie's (2009) *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*, although that's not a bad place to start.

## Meta-Analysis Explained

*Meta-analysis* refers to a kind of research study that quantitatively combines the results of multiple studies. In the same way that research studies focus their attention on students or teachers as the subjects of their analyses, meta-analysis focuses on collections of independent research studies.

Before the development of meta-analysis, reviews of research were subjective, easily influenced by the biases and prejudices of the reviewers. Over the years, various schemes were proposed for conducting more objective reviews of research, but they all suffered from some basic problems. One idea was to count the numbers of studies that supported—or failed to support—a particular approach to teaching, kind of like tallying the runs in a baseball game (e.g., after-school programs are winning 7–2).

There are problems in summing up research in that way. For instance, studies may differ in the sizes of the learning impacts linked to the various versions of the intervention of interest. What if several of the positive results are minuscule—that is, there were only tiny learning gains—and the negative ones are big (perhaps the kids taught by the handy-dandy approach sometimes didn't even do as well as the control groups—yikes!)? Just counting votes in a case like that would lead you to think that teaching in a particular way was beneficial, although that would be doubtful overall.

Or another example: Evidence based on data drawn from thousands of kids shouldn't be treated as equal to the evidence drawn from dozens of them. Some kind of weighting is needed when the numbers of kids participating in the various studies differ by much.

Meta-analysis doesn't just count votes; it combines the data from the individual studies, considering the differing sizes of effects and sample sizes, so what results is a true average—in other words, a better idea of the likelihood that something will work for you and how well it might work.

Meta-analysts worry a lot about things like double counting that can mess up that "estimating the average" effect. If a study makes more than one comparison, such as looking at whether a program improves decoding, fluency, spelling, and reading comprehension, you wouldn't want to treat each of these comparisons as if they were drawn from independent studies. That might mislead you into thinking that the approach was more or less effective than was actually the case. Accordingly, meta-analysts are careful to use an average of such results

doi:10.1002/trtr.1570

within a study before trying to combine that result with the outcomes of other studies.

When I worked on the National Reading Panel, I was synthesizing the studies on oral reading fluency instruction. I found two studies that were strikingly similar in design, conducted by the same team of authors, but that differed in the number of students in each study. The similarities were suspicious. I called the authors and found out that they had implemented a study in several schools but hadn't been able to get all the data at the end, so they proceeded to publication with what they had. Later, when the missing data became available, they combined what they had originally published with the newly available results, which explained the difference in the number of subjects. If I had included both studies in my analysis, I would have been double counting those data, and the strengths of the outcome would have been overstated.

Another thing to know about meta-analysis is that it usually provides two types of results. The first outcomes, the main effects, are a summary or an average effect of the instructional intervention across all the studies and all the within-study comparisons that were made. For example, Graham and Hebert (2010) conducted a meta-analysis of studies that evaluated the effects on reading comprehension of having students write about the texts that they read. The results were consistently positive and sizable. They concluded that writing about text improves reading comprehension.

That main effect is truly an effect. What I mean is that the instructional innovation—in this case, writing about what one reads—really did improve student reading comprehension. We know that such writing actually affected reading, or caused the improvement in reading, because the researchers in all the original studies had some students writing about text while others did not. The meta-analysis reported the average effect that was obtained across all of these independent comparisons.

Yet, meta-analyses also usually report other results, moderator effects, that are correlational. Each study finds an effect, positive or negative, due to the treatment in question. These effects will vary, some being bigger and others smaller, and these variations in effect sizes can then be correlated with other study variables. These correlational results provide clues to questions that teachers may have about the instructional approaches, such as, With whom does this work best? or Which versions of the technique are most effective?

In Graham and Hebert's (2010) meta-analysis, for instance, the effects of summarizing were compared with the effects obtained when other kinds of writing about text (e.g., analyses or syntheses) were used. For elementary students, the summarization effects tended to be stronger, but this pattern reversed as students moved up the grades.

Remember, the original research studies didn't manipulate this aspect of instruction. What I mean is that none of the original studies compared summarization with other kinds of writing. That was just a pattern that emerged across these approximately 100 studies. Some of the studies compared summarization with reading alone or with reading and rereading. Other studies made similar comparisons using other kinds of writing. Looking at all those studies together revealed the pattern of summary writing being relatively more beneficial for younger readers, with greater learning payoffs from analytical writing later on. We'd be more certain of these results if someone actually tested that comparison directly, but until such studies are done, it seems reasonable to engage kids in more summary writing in elementary school, replacing this with more extensive writing about text as they progress through the grades.

A benefit of those correlations is that they can be useful for adjusting our estimates of how powerful instruction might be. For example, studies that randomly assign students to instructional conditions are likely to provide more accurate estimates of effects than studies that depend upon comparing already established groups. Meta-analysis allows us to compare the results of the randomized studies with the others to see if it makes any difference, and if so, how much. Being able to "correct" effect size estimates based on quality factors or other potentially important variables is valuable.

The main effects of a meta-analysis might indicate that an experimental approach leads to a full-semester learning difference for students. Sounds great, right? But then the meta-analyst compares the relative rigor of the studies, including random assignment, careful observation, length of treatments, type of assessment used, and so on. Those comparisons might lead to important adjustments in the effect size estimates. For example, effects tend to be substantially smaller with reliable standardized tests than with specially tailored experimenter-made tests. If the reason you want to adopt a particular teaching approach is that you expect it to improve standardized test performance, then that adjustment is of particular importance.

## What About *Visible Learning*?

In 2009, Hattie published *Visible Learning,* a summary of more than 800 meta-analyses, and since then he has published various offshoots on this scheme (e.g., Fisher, Frey, & Hattie, 2016). I keep a copy of the original text close at hand. I know of no other compendium of educational meta-analyses as complete (although, with a dozen or more relevant meta-analyses published each year on literacy alone, it is quickly going out of date). It is a very useful tool for me, as I get asked a lot of practical questions about topics that I don't necessarily research myself. Having a quick resource of this kind is valuable to me, as it is and would be for many other educators.

However, my concern is how that educator colleague said she used this worthwhile book. She believed that all she needed to do was refer to the appendix that showed the relative effects of more than 100 instructional approaches with a summary of the multiple meta-analyses that have been completed on each. That list, for her, was the agenda for what she needed to apply in her school to improve reading.

For me, Hattie's book is a quick source, a shortcut to finding meta-analyses on far-ranging topics of importance. For her, it was the end of the line, the proof that she needed to conclude that particular kinds of teaching approaches worked and to determine which would be relatively most important for improving her school's reading achievement.

It's just not that easy, unfortunately. In determining what works, Hattie's summaries point us in a useful direction, but they also can mislead in important ways. Using Hattie's results well requires critical reading on the part of educators—if they want to get it right.

There are several problems with Hattie's treatment of meta-analysis. The first is a lack of transparency. It simply is not always clear what he has done. He didn't provide a lot of explanation of how he combined the original meta-analyses, so one is forced to reverse-engineer his summaries just to figure out how he got there. This is especially an issue when I think I see an error. Sometimes I can figure it out, and other times I walk away scratching my head.

And errors there are. The problem that most concerns me—and one apparently not discussed by Hattie's critics—has to do with how one goes about combining results from multiple meta-analyses. As I've already explained, meta-analysis itself is about combining data in a valid way, but there is no agreed-upon methodology for then recombining the resulting meta-analyses. In medicine, meta-analyses are ongoing projects; when a new study is done on a topic, it simply gets added to the preceding collection of studies and the meta-analysis is recalculated with these new data (e.g., the Cochrane Collaboration).

What Hattie seems to have done is just take an average of the original effects reported in the various meta-analyses. That sometimes is all right, but it can create a lot of double counting and weighting problems that play havoc with the results.

For example, Hattie combined two meta-analyses of studies on repeated reading. He indicated that these meta-analyses together included 36 studies. I took a close look myself, and it appears that there were only 35 studies, not 36, but more importantly, four of these studies were double counted. Thus, we have two analyses of 31 studies, not 36, and the effects reported for repeated reading are based on counting four of the studies twice each!

Students who received this intervention outperform those who didn't by 25 percentiles, a sizable difference in learning. However, because of the double counting, I can't be sure whether this is an over- or underestimate of the actual effects of repeated reading that were found in the studies. Of course, the more meta-analyses that are combined, and the more studies that are double and triple and quadruple counted, the bigger the problem becomes.

Another example of this is evident with Hattie's combination of six vocabulary meta-analyses, each reporting positive learning outcomes from explicit vocabulary teaching. I couldn't find all of the original papers, so I couldn't thoroughly analyze the problems. However, my comparison of only two of the vocabulary meta-analyses revealed 18 studies that weren't there. Hattie claimed that one of the meta-analyses synthesized 33 studies, but it only included 15, and four of those 15 studies were also included in Stahl and Fairbanks's (1986) meta-analysis, whittling these 33 studies down to only 11. One wonders how many more double counts there were in the rest of the vocabulary meta-analyses.

This problem gets especially egregious when the meta-analyses themselves are counted twice! The National Reading Panel (National Institute of Child Health and Human Development, 2000) reviewed research on several topics, including phonics teaching and phonemic awareness training, finding that teaching phonics and phonemic awareness was beneficial to young readers and to older struggling readers who lacked these particular skills. Later, some of these National Reading Panel meta-analyses were republished, with minor updating, in refereed journals (e.g., Ehri et al., 2001; Ehri, Nunes, Stahl, & Willows, 2002). Hattie managed to count both the originals and the republications and lump them all together under the label Phonics Instruction—ignoring the important distinction between phonemic awareness (chldren's ability to hear and manipulate the sounds within words) and phonics (children's ability to use letter–sound relationships and spelling patterns to read words). That error both double counted 86 studies in the phonics section of *Visible Learning* and overestimated the amount of research on phonics instruction by more than 100 studies, because the phonemic awareness research is another kettle of fish. Those kinds of errors can only lead educators to believe that there is more evidence than there is and may result in misleading effect estimates.

Of course, a related problem arises when meta-analyses of very different scopes are combined. What if one of the meta-analyses being averaged has many more studies than the others? Simply averaging the results of a meta-analysis based on 1,077 studies with a meta-analysis based on six studies would be very misleading. Hattie combined data from 17 meta-analyses of studies that looked at the effects of students' prior knowledge or prior achievement levels on later learning. Two of these meta-analyses focused on more than a thousand studies each; others focused on fewer than 50 studies, and one as few as six. Hattie treated them all as equal. Again, potentially misleading.

Finally, Hattie's lists of effect sizes provided no information about the moderator analyses, although he selectively discussed some of these throughout the volume. As noted earlier, if flaws in the original research inflate an effect size, then it is important to note this. Similarly, if a particular approach to instruction works better at some grade levels than others (as is true of homework), has a more positive impact on math learning than reading (as is the case with frequent assessment), or has varied impacts on different kinds of reading outcomes (as is true of vocabulary, phonics, and fluency instruction), then it is important for teachers and principals to know this. Hattie addressed some of these concerns in the body of his books, but he ignored many of them, too.

My suggestion: Use *Visible Learning* as a resource, but read the original meta-analyses as well (and sometimes it will be worth even going back to some of the original studies to find out what was really being done). Obviously, the more kids who will be affected by your judgments and choices, the more careful you want to be in trying to apply research to teaching.

## REFERENCES

Carlson, K.D., & Ji, F.X. (2011). Citing and building on meta-analytic findings: A review of recommendations. *Organizational Research Methods*, 14(4), 696–717. doi:10.1177/1094428110384272

Ehri, L.C., Nunes, S.R., Stahl, S.A., & Willows, D.M. (2002). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Journal of Direct Instruction*, *2*(2), 121–166. (Reprinted from *Review of Educational Research*, *71*(3), 393–447)

Ehri, L.C., Nunes, S.R., Willows, D.M., Schuster, B.V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly*, *36*(3), 250–287. doi:10.1598/RRQ.36.3.2

Fisher, D., Frey, N., & Hattie, J. (2016). *Visible learning for literacy, grades K–12: Implementing the practices that work best to accelerate student learning*. Thousand Oaks, CA: Corwin.

Graham, S., & Hebert, M.A. (2010). *Writing to read: Evidence for how writing can improve reading*. Washington, DC: Alliance for Excellent Education.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Stahl, S.A., & Fairbanks, M.M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, 56(1), 72–110.

**The department editor welcomes reader comments.**

*Timothy Shanahan* is a Distinguished Professor Emeritus at the University of Illinois at Chicago, USA, and is a past president of the International Reading Association; e-mail shanahan@uic.edu. He also publishes a widely followed blog, Shanahan on Literacy: www.shanahanonliteracy.com.