Retrieval Practice
retrievalpractice.org

Meta-Analysis

EFFECT SIZES AND META-ANALYSES: HOW TO INTERPRET THE "EVIDENCE" IN EVIDENCE-BASED

Kripa Sundar, Ph.D.
Pooja K. Agarwal, Ph.D.

Based on a wealth of research by cognitive scientists, there are four powerful teaching strategies that transform students' long-term learning:

1. Retrieval practice boosts learning by pulling information out of students' heads, rather than cramming information into students' heads.
2. Spacing boosts learning by spreading lessons and retrieval opportunities out over time, so learning is not crammed all at once. In this way, forgetting is a good thing for learning.
3. Interleaving boosts learning by mixing up closely related topics, encouraging discrimination between similarities and differences.
4. Feedback-driven metacognition boosts learning by providing the opportunity for students to know what they know and know what they don't know.

In this guide, we want to empower you to **question the "evidence" in the evidence-based practices** you encounter. Specifically, we want to equip you with the tools to assess whether summative evidence that's presented based on effect sizes is trustworthy, such as meta-analyses (e.g., a recent meta-analyses on retrieval practice) and meta-***meta***-analyses (e.g., John Hattie's research).

We have separated this guide into three parts:

- **Section 1:** An overview of meta-analyses
- **Section 2:** Introduction to meta-meta-analyses
- **Section 3:** Effect size statistics, tables, and more

Note: As you read this guide, you may question: "Where will I find studies to read?" Go to scholar.google.com. Say you are looking for studies that focus on retrieval practice in elementary grades, type "retrieval practice elementary" in the search bar to find a host of studies. Visit RetrievalPractice.org for more tips on accessing research.

What you should look for to best interpret the "evidence" in evidence-based?

QUESTIONS TO ASK

When you're reading an individual study:

- What type of questions were used to measure learning?
- Who were the participants? How many students participated? Are they in a similar context to your class?
- What was being compared? Was there a comparison group that didn't receive the strategy?
- How was the study carried out? Can you implement the strategy similarly in your classroom?

When you're reading a meta-analysis:

- Does the meta-analysis compare apples to apples, or apples to oranges, in terms of learning outcomes? Is learning measured similarly across all the individual studies?
- Is there consistency across the learning strategies? Do they sound similar, consistent, and specific? (for example, "feedback" can be implemented in many different ways!)
- When was the meta-analysis published? Have there been a lot of studies since then?

When you're reading a meta-meta-analysis:

- Have the authors clearly defined the criteria for studies to be included in the meta-analysis?
- How many meta-analyses went into the meta-meta-analysis?
- How recently were the meta-analyses published?
- Since a meta-meta-analysis is "twice removed" from the original study, how relevant are the results for you and your students?

SECTION 1: AN OVERVIEW OF META-ANALYSES

What is a meta-analysis?

A meta-analysis is a systematic quantitative summary of relevant research on a specific topic. This statistical method is popularly used in the medical field.

For example, consider research we conducted on retrieval practice, where pulling information "out" of students' heads (via brain dumps, flashcards, etc.) improves long-term learning. In one study, middle school students participated in lessons with low-stakes quizzes compared to lessons without quizzes. Students' performance on an end-of-semester exam was greater for information learned during low-stakes quizzes compared to non-retrieved information. Simply put, this study provides evidence of the benefits of retrieval practice.

A meta-analysis combines a number of studies on a topic statistically, such as retrieval practice, to find an average effect, and then looks to see if there are any patterns across different contexts of implementation. For example, in a meta-analysis we conducted in 2017, we found that retrieval practice consistently benefitted learning across 118 studies. In other words, the meta-analysis assures us that over time and across studies, there is a trend of positive effects from retrieval practice for learning.

John Hattie's Visible Learning is *not* a meta-analysis. It is a meta-*meta*-analysis, or a summary of multiple meta-analyses.

Are meta-analyses better than individual studies?

Individual studies are valuable. They give you more context and detail. Plus, they are easier to follow than a meta-analysis, where the statistics can be overwhelming (see Section 3, page 11, of this guide for a how-to on the statistics).

Meta-analyses, on the other hand, give us the lay of the land in terms of research in the field: what has been studied, what hasn't, and how strongly the effect was replicated in other studies. Keep in mind that while individual studies can take 1-3 years to research and publish, meta-analyses can take a similar amount of time or longer, depending on the volume of individual studies included and the number of researchers involved.

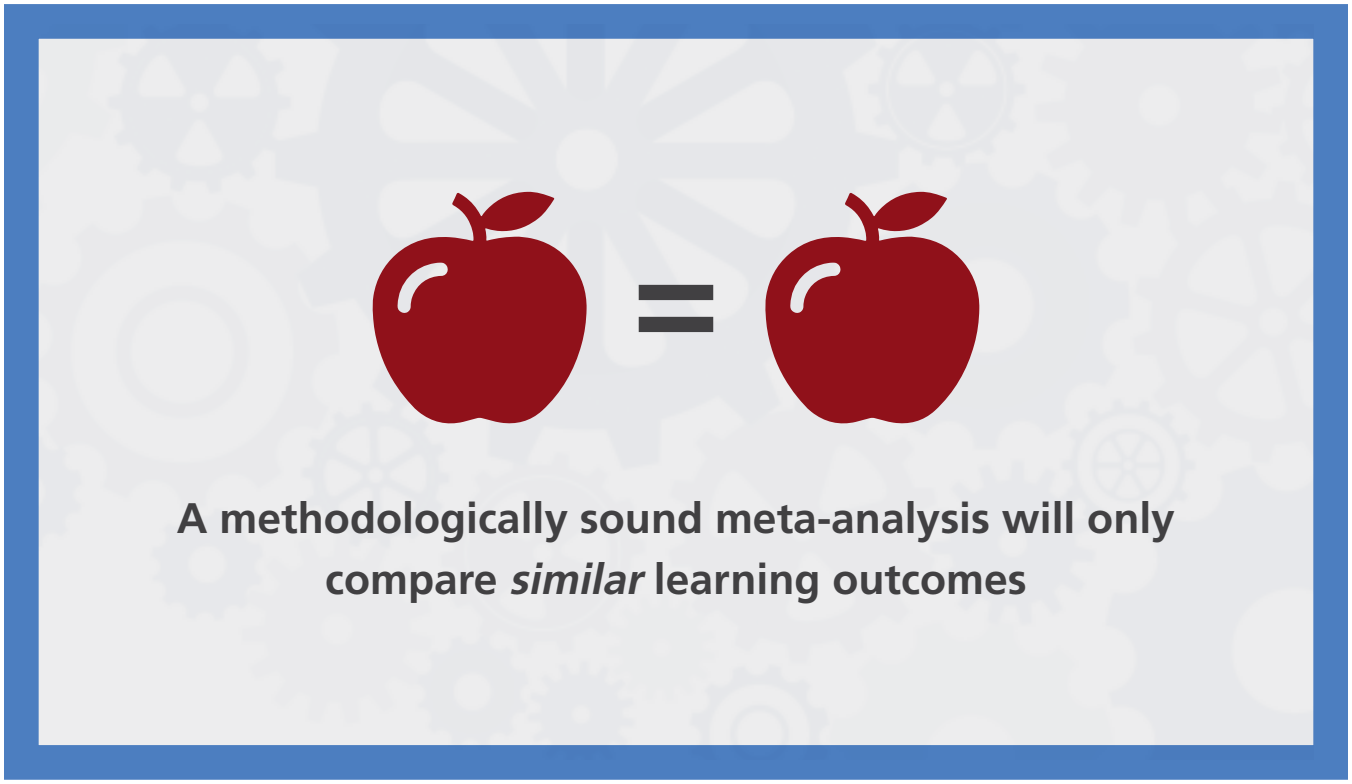
Are all meta-analyses trustworthy?

Not all meta-analyses are as rigorous as they may seem. Here are three red flags that a meta-analysis could be less than trustworthy:

Not necessarily. When it comes to learning outcomes measured in a meta-analysis, researchers may be comparing apples to oranges.

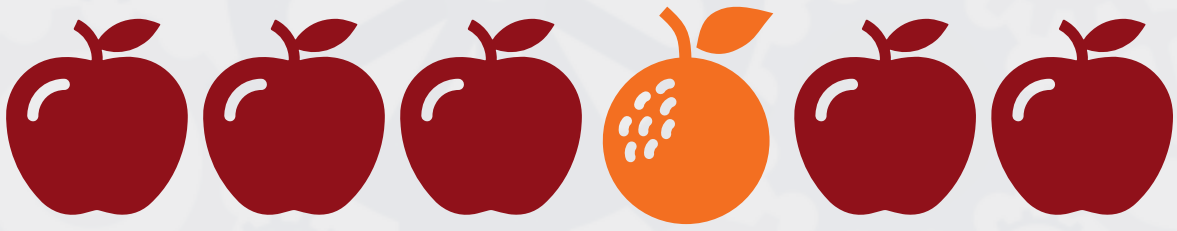
Researchers have to decide on criteria for inclusion based on a research question and then find specific studies that fit the criteria. A methodologically sound meta-analysis will only compare similar learning outcomes: retention of information, transfer of knowledge, higher order learning, etc. You want to compare apples to apples in a meta-analysis.

For example, when our team conducted a meta-analysis on the effect of retrieval practice, we wanted to include studies that compare learning outcomes between groups that did and did not practice retrieval. Importantly, we only included studies that measured student learning and transfer of knowledge in similar ways, i.e. using exams. One of the most important questions we asked: Did every study included in the meta-analysis have similar definitions for learning outcomes? We looked at what the researchers were measuring and excluded studies that say, focused on measuring anxiety only. When we had a set of studies that focused on learning, we looked to see how they measured it — was it a test of recall? Were researchers looking for application of content? Were they measuring both? And then grouped similar outcomes together.



The learning strategies being researched in a meta-analysis are not consistent.

Look for consistency in learning strategies and research methods. If learning strategies included in the meta-analysis are **not consistent and logical**, then beware! For example, if you find a meta-analysis that groups together "feedback" strategies including teacher praise, computer instruction, oral negative feedback, timing of feedback, and music as reinforcement, does that sound consistent to you? Drawing conclusions from such broadly defined learning strategies can be problematic, particularly when it comes to knowing which specific strategy to implement in the classroom.



If learning strategies included in the meta-analysis are not consistent and logical, then beware!

The meta-analysis does not include recent studies.

Results can change as more studies are conducted on the topic, and research will continue to be published. Keep an eye out for how many studies are included in the meta-analysis and when the individual studies were published. You want to look at results from more recent studies to observe potential changes in trends and effect strength or direction.

Take for example, retrieval practice. There have been several meta-analyses on the topic each adding on studies or approaching it from different angles. How frequent? Well, just in the last decade, there have been meta-analyses done in 2012, 2014, and two in 2017: ours, summarizing retrieval practice studies overall and another focused on studies in the classroom.

As we worked towards the publication of our 2017 meta-analysis on retrieval practice, we had stopped including articles published 2014 onwards to allow time for analysis and writing. Then, the publication process took about two years and wow! There were dozens of articles published between finishing our manuscript and it getting published.

In 2018, we summarized about 58 studies on the seductive details effect and a month after we submitted our paper for publication, 12 more studies were published on the same topic! We then had to redo our analysis and found that while the overall effect remained unchanged, the nuances of effect changed. Eventually, we were able to publish the updated findings.

Look at the year the meta-analysis was published and look for the publication dates of the studies that were included. How close to the current date is the meta-analysis and the included studies? If there is a significant number of new studies (say 20 or more) published since the meta-analysis, the nuances in findings may be outdated–even if the overall effect still holds.

SECTION 2: INTRODUCTION TO META-*META*-ANALYSES

Is John Hattie's Visible Learning a meta-analysis?

No. John Hattie's work is a meta-***meta***-analysis, where he combines findings from multiple meta-analyses that have already been conducted.

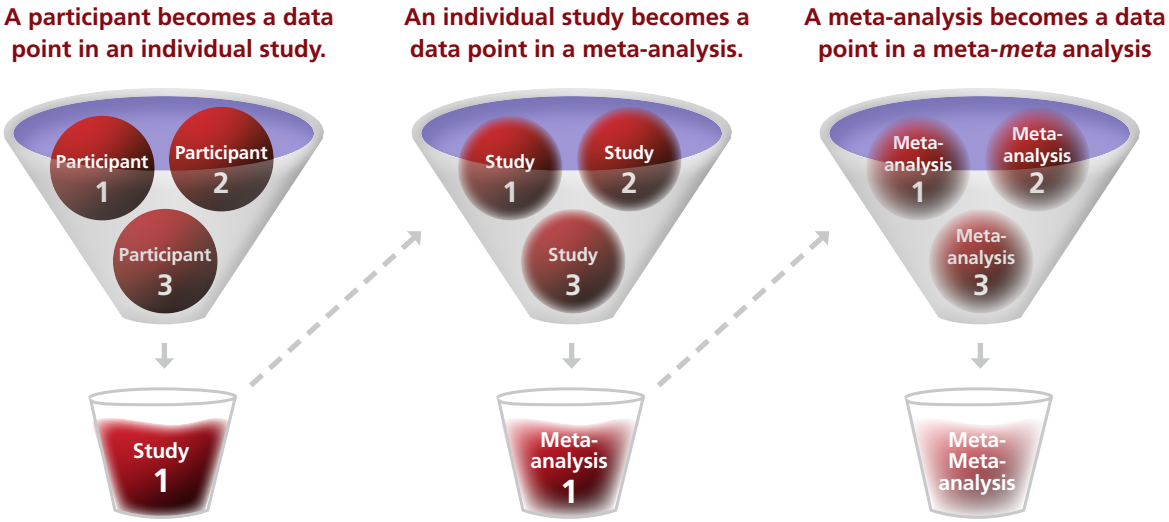
If you are familiar with Visible Learning, you may have asked yourself these questions:

1. If retrieval practice is so powerful, why is Hattie's score for retrieval practice only 0.54?
2. If spacing is so powerful, why is Hattie's score 0.60?
3. Why is the score for interleaving only 0.21?
4. Why are the scores for feedback and metacognition higher (0.70 and 0.60, respectively)?

Hattie's scores for these learning strategies are **twice removed** from the original individual studies that had specific comparisons using specific measures. Plus, Hattie's scores aren't really effect sizes (see Section 3, page 12 for an explanation). Even when conducted rigorously, scores from a meta-meta-analysis can be very different from the original effect sizes in an individual study or from a meta-analysis, which limits how informative it can be for classroom implementation.

What exactly is a meta-*meta*-analysis?

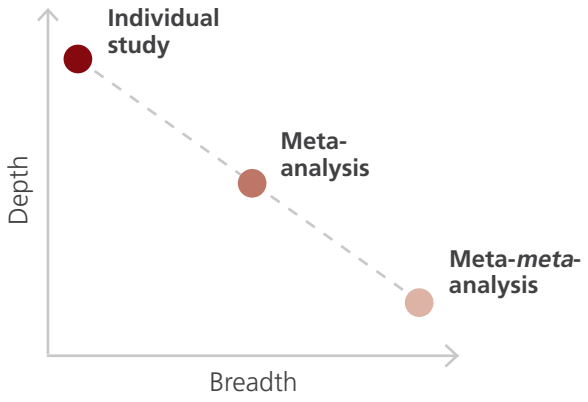
Think of a meta-*meta*-analysis as watered down coffee. Just like how someone who has watered down coffee can say they had "real" coffee, a meta-*meta*-analysis is also research–just not as impactful (or caffeinated).



With each step, the picture becomes fuzzier. At the very start, an individual study will tell you the exact learning strategy, procedure, and student performance. When that study is summarized in a meta-analysis, it is combined with other similar studies. It is useful and reassuring that more than one study found a similar effect for a learning strategy, but less useful than an individual study to learn how to implement it. In sum, a meta-analysis is once removed from the context of an original study.

Now, imagine if we took a meta-analysis and combined it with another meta-analysis. We've essentially **averaged two averages!** Take those two and add even more meta-analyses to the mix. What if all the meta-analyses had different criteria for including studies or focused on different learning outcomes? None of this is accounted for in a meta-meta-analysis beyond the researchers' judgement (partly why Hattie's work is criticized so much). A meta-meta-analysis is now twice removed from an original study. It can offer a broad picture of the state of research, but not much in terms of how to implement a specific strategy in your classroom.

For example, Hattie provides a score of 0.18 for "web-based learning." While the individual studies on web-based learning that were included in his score may have been rigorous and well-controlled, you can imagine that combining or collapsing over many studies waters down the actual intervention: What exactly is meant by "web-based learning?" This is an example where looking at an individual study may provide more insight in terms of implementation of a learning strategy than a meta-analysis or meta-meta-analysis of multiple studies with different definitions.



If Hattie's Visible Learning is not a meta-analysis, should I trust his work?

Don't throw the baby out with the bath water. Consider the underlying idea that Hattie presents: learning that is visible or demonstrated. This tangible proof of learning makes for easier assessment, which is often how we identify gaps and next steps. Even so, **ignore the specific rankings, comparisons, and scores assigned to the different activities**. Why? For two reasons:

1. There are questions regarding his methods (particularly the inclusion of low-quality studies and his calculation of his scores) that are unresolved within the academic research community.
2. Research is constantly being published. This means that the summary score you read nine years ago may have changed direction or strength based on newer research; like when I had to update my meta-analysis in just one year.

Consider the underlying idea that Hattie presents: *learning that is visible or demonstrated.*

SECTION 3: EFFECT SIZE STATISTICS, TABLES, AND MORE

What is an effect size?

An effect size is like a yardstick that measures and compares the effectiveness of a learning strategy within a study and also between studies. In other words, effect sizes serve a few purposes:

1. to look at the effect of a strategy in an individual study,
2. to compare strategies from two or more studies, and
3. to compare multiple effects with a meta-analysis.

As a rule of thumb, researchers interpret Cohen's *d* and Hedge's *g* in these categories:

- **very small effect < 0.2**
- **0.2 < small effect < 0.5**
- **0.5 < medium effect < 0.8**
- **0.8 < large effect**

In education research, for example, you may want to examine the effect of retrieval practice compared to highlighting. You would ask one group of students to practice retrieval and another group of students to highlight while reading a textbook. Then, you'd compare who learned more, on average. Now, what if you came across a different study that compared retrieval practice to concept mapping? How could you compare findings between the results of both studies? Researchers use a standardized effect size to draw comparisons across studies.

Commonly used effect sizes like Cohen's *d* or Hedge's *g* quantify how much of an effect learning strategy X has had on learning outcome Y on average. You can calculate it yourself if you have the means and standard deviations of the comparison groups. Meta-analyses typically use Hedge's *g* which adjusts for differences in sample sizes of the studies.

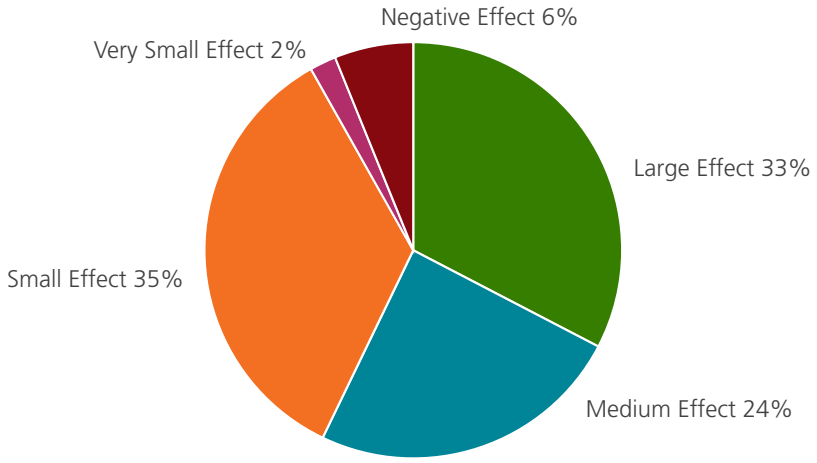
Are Hattie's scores the same as effect sizes?

No. Hattie uses the term "Cohen's *d*," but there are several statistical concerns with his calculation methods. We urge teachers to recognize that Hattie's scores can not be equated to what a majority of the research community calculates and interprets as effect sizes.

Consider this example: In Hattie's work, retrieval practice ("practice testing") had a score of 0.54. Our 2017 meta-analysis that summarized over 200 comparisons showed an effect of 0.70. A recent review of the literature demonstrated that, out of 49 effect sizes (Cohen's d), the majority (57%) revealed medium or large benefits from retrieval practice. In relation to the red flags on pages 5-6 of this guide, researchers compared apples to apples (classroom studies, not including laboratory studies), the learning strategies were consistent (low-stakes quizzes during class periods), and recent studies were included (through 2018). In other words, both the recent meta-analysis and the review of the literature demonstrated larger and more consistent benefits from retrieval practice than Hattie's score suggests.

Distribution of effect sizes from 50 classroom studies on retrieval practice

Review of the literature by Agarwal, Nunes, & Blunt (2021)



How would an effect size apply in my classroom?

Our meta-analysis on retrieval practice in 2017 found that nearly 100 years of research summarized to a 0.70 effect size (with a confidence interval of 0.63 to 0.78). If you took a rule of thumb interpretation, then this qualifies as a big effect. But what does that look like in your classroom? Based on our meta-analysis, if your students typically average an exam score of 85 (plus or minus 10 points) without retrieval practice, then using retrieval practice can move the average up to an exam score of 92. Pretty neat, huh?[1]

1 Here is a quick look at these calculations. With an average score of 85 and a standard deviation of 10, we can use the average effect size of retrieval practice (g = 0.70) as follows: 85 + (10 x 0.70) = 92 points. The range is calculated in the same way: 85 + (10 x 0.63) = 91.3 points and 85 + (10 x 0.78) = 92.8 points. You can use the same calculation for other individual studies, too!

What should I be mindful of when interpreting effect sizes?

As with all things in research, it depends. Here are two caveats to consider when interpreting effect sizes.

1. **Negative effect sizes are as important as positive effect sizes–it just depends on the question.** The effect size number denotes the strength, and the positive/negative of the effect size denotes the direction for interpretation. For example, does drinking coffee make you more productive? Yes, of course, and that'd be a positive effect size. As a second example, what is the impact of mindfulness training on distractedness? If we want to reduce distracted behavior, then a negative effect size would be a good thing.

2. **Effect sizes must be interpreted in context.** Consider this question: does drinking coffee make you more productive than drinking Redbull? If there is little difference between the two drinks, then you'd get a small effect size. Now, think about this second question: does drinking coffee make you more productive than drinking water? If there is a large benefit for coffee, you'd get a large effect size. But if you only look at effect sizes, you might think that Redbull has no effect on productivity–quite the opposite from the actual result. In terms of your classroom, retrieval practice will likely have a larger effect size when compared to let's say, restudying. But if you regularly practice concept mapping in your classroom, you may see a smaller effect size when compared to retrieval practice since both strategies improve student learning with similar average effects.

How do I read the tables in a meta-analysis?

You got your hands on a meta-analysis. As if the 35 pages of tiny text were not intimidating enough, there are an additional 15 pages of imposing tables. Doesn't feel like an afternoon of fun. It could be, though! Let's work through an example. See the table below from our 2017 meta-analysis on retrieval practice.

Rows 1 and 2: All (fixed-effects model) and All (random-effects model)

- This is the grand summary of all studies included in our meta-analysis. A fixed effect model "fixes" the interpretation of findings to the included studies only, but a random effects model includes a varying error model allowing us to generalize findings more broadly. In this example, let's go through the random-effects model (row 2), column by column.
- The meta-analysis summarized data from 15,427 participants (N) across 272 comparisons (k) of learning outcomes for students who learned with and without retrieval practice.
- On average, retrieval practice has an effect size of 0.70 (g+)
- This effect can fluctuate though, ranging from 0.63 (lower) to 0.78 (upper) standard deviations (known as the 95% confidence interval)

Rows 3-8: Comparison treatments

- Moderators, or conditions of effect, show how the effect can change. In this table, comparison treatments included restudying, fillers, a mix of activities, etc.
- The effect varied significantly between the different comparison conditions: The bottom row, Between-levels QB, has $p < .001$. When $p < .05$, it means that there is a less than 5% chance that there is no difference in variation of effects across conditions.
- When looking at the effect sizes for the different comparison treatments, we can see that retrieval practice is much more effective than rereading (g+ = 0.51), a filler activity (g+ = 0.93), and a mix of activities (g+ = 0.71).
- But keep an eye on the number of studies column (k)! If you have a strong effect (a high g+) and many studies (a high k), you're golden. On the other hand, notice that the large effect of retrieval practice vs. a mixture of activities (0.71) is drawn from only 18 studies. You may want to be more cautious about this finding, compared to the robust effect of retrieval practice vs. restudying from 195 studies.

N = Number of participants across studies

g+ = Standardized effect size

Test of heterogeneity = What is the probability that there is *no* difference in effects among included studies?

Between levels (Qb) = Is there a difference between the conditions of influence?

k = Number of comparisons of retrieval practice vs. no retrieval practice

95% confidence interval = The range for the effect size

Overall Weighted Mean Effect Size

				95% confidence interval		Test of Heterogeneity
	N	*k*	*g+*	Lower	Upper	*p*
All (fixed-effect model)	15,427	272	0.61*	0.58	0.65	<.001
All (random-effects model)	15,427	272	0.70*	0.63	0.78	

Comparison treatments

Restudying/ Re-reading	10,491	195	0.51*	0.47	0.55	<.001
Filler or no activity	3,452	58	0.93*	0.86	1.01	<.001
Mixture of restudying and filler activities	1,340	18	0.71*	0.59	0.82	.01
Not reported	144	1	0.68*	0.35	1.01	
Between-levels (QB)						<.001

* $p < .05$

+ This table has reduced columns than a table in a published meta-analysis for purposes of an example.

For research, resources, and tips, visit **retrievalpractice.org**

©2021 We'd like to thank Dr. Olusola Adesope for spearheading the 2017 meta-analysis on retrieval practice and for introducing Kripa to meta-analyses early in her academic career. We also thank Dr Dominic Trevisan for his collaboration in completing the meta-analyses.

Retrieval Practice
retrievalpractice.org
ask@retrievalpractice.org

@RetrieveLearn
/RetrievalPractice