Real Gold vs. Fool's Gold

The VISIBLE LEARNING™ Methodology for Finding What Works Best in Education

John Hattie
Arran Hamilton

A JOINT PUBLICATION
CORWIN
Cognition Education Group

Real Gold vs. Fool's Gold

The VISIBLE LEARNING™ Methodology for Finding What Works Best in Education

John Hattie | Arran Hamilton

A JOINT PUBLICATION

CORWIN | Cognition Education Group

CORWIN
A SAGE Publishing Company

FOR INFORMATION:

Corwin
A SAGE Company
2455 Teller Road
Thousand Oaks, California 91320
(800) 233-9936
www.corwin.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
18 Cross Street #10-10/11/12
China Square Central
Singapore 048423

Acquisitions Editor: Ariel Curry
Editorial Assistant: Eliza Erickson
Production Editor: Laureen Gleason
Copy Editor: Christina West
Typesetter: C&M Digitals (P) Ltd.
Proofreader: Victoria Reed-Castro
Graphic Designer: Lysa Becker
Marketing Manager: Kerry Garagliano

ISBN 978-1-5443-9992-8 (web PDF)

DISCLAIMER: This paper may direct you to access third-party content via web links, QR codes, or other scannable technologies, which are provided for your reference by the author(s). Corwin makes no guarantee that such third-party content will be available for your use and encourages you to review the terms and conditions of such third-party content. Corwin takes no responsibility and assumes no liability for your use of any third-party content, nor does Corwin approve, sponsor, endorse, verify, or certify such third-party content.

Contents

About VISIBLE LEARNING™

In 2008, Professor John Hattie published *Visible Learning*, a synthesis of more than 800 meta-studies covering more than 80 million students. The book revealed what education variables have the biggest impact on learning and created a new mindset that has swept educators around the world. Applying the Visible Learning methodology means that students are taught to know what they need to learn, how to learn it, and how to evaluate their own progress. Using the Visible Learning approach, teachers become evaluators of their own impact on student learning. The combination causes students to drive their own learning. Since 2008, Professor Hattie has teamed with highly influential educators to expand the Visible Learning canon with books, including *Visible Learning Into Action*, *Visible Learning for Teachers*, *Visible Learning for Mathematics*, and *Visible Learning for Literacy*.

Visible Learning$^{plus®}$ is the model of professional learning that takes the theory of Hattie's research and puts it into a practical inquiry model for teachers and school leaders to ask questions of themselves about the impact they are having on student achievement. Visible Learning^{plus} is a result of the collaboration between Professor John Hattie and Corwin with the aim to help educators translate the Visible Learning research. Through a global network of partners, Visible Learning^{plus} professional learning is implemented in over 20 countries in North America, Europe, and the Pacific.

Learn more at www.visiblelearningplus.com

About Corwin

Corwin, a SAGE Publishing company, was established in 1990, first as a professional book publisher, now as a full-service professional learning company, offering professional development solutions on the topics that matter most and the delivery methods that work best to achieve a school or district's objectives. Its many resources range from a library of 4,000+ books to on-site consulting to online courses and events. At the heart of every professional learning experience is the book content and author expertise that have made Corwin the most trusted name in professional development.

Learn more at www.corwin.com

About the Authors

Professor John Hattie is Laureate Professor at the Melbourne Graduate School of Education at the University of Melbourne and Chair of the Australian Institute for Teaching and School Leadership. His areas of interest are measurement models and their applications to education's problems, and models of teaching and learning. He has published and presented over 1,000 papers, supervised 200 thesis students, and published 31 books, including 18 on understanding and applying the Visible Learning™ research.

Dr. Arran Hamilton is Group Director of Education at Cognition Education. His early career included teaching and research at Warwick University and a stint in adult and community education. Arran transitioned into educational consultancy more than 15 years ago and has held senior positions at Cambridge Assessment, Nord Anglia Education, Education Development Trust (formerly CfBT), and the British Council. Much of this work was international and has focused on supporting Ministries of Education and corporate funders to improve learner outcomes.

Acknowledgments

We would like to thank the eight anonymous reviewers who provided feedback on an earlier draft of this manuscript. Any remaining errors or omissions are entirely our own.

Introduction

Gold is one of the most valuable natural resources in the world. It is rare, shiny, extremely malleable, resistant to corrosion, and a good conductor of electricity. Gold's monetary and symbolic value make it a coveted material for all kinds of goods, from jewelry to furniture, and of course (historically) as a currency.

Its value has created an entire market for counterfeit gold. Unless you happen to work for the London Bullion Market Association or another reliable trade source, it can be easy for the layperson to be duped by a realistic-looking piece of metal. Even fake gold often contains traces of real gold or is blended with other metals, making it even harder to tell the difference between real and "replica." Short of taking gold to an approved trader for inspection, there are methods anyone can learn that provide a fairly reliable way to judge whether a piece is genuine, including looking for the hallmark, dropping acid on the surface, trying to lift the metal up with a magnet, and placing shavings in water.

As educators, we are all on a continual hunt for metaphorical gold. The motherlode that we seek? The most effective interventions to drive up student learning outcomes. The challenge that we face is that almost every pedagogical approach or classroom product seems to come with a "hallmark" or appears to pass some sort of empirical test of fitness. But it can't all be (as) fit. Some of it *has* to be more golden than the rest. And this means there is potentially a "long tail" of foolishness: interventions that look shiny but that are a complete waste of everyone's time and effort.

This was exactly the challenge that the Visible Learning research was designed to address. The idea was to harvest, bundle, and synthesize as much research as possible on *every* type of education intervention, in order to build the definitive database (and narrative) on what works best and in what contexts. The intention was to give educators and policymakers access to a better-quality gold detector.

When *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement* was published in 2009, it was surprisingly well received. Despite it looking and reading like a telephone directory, the *Times Educational Supplement* described the book as "teaching's holy grail" (Mansell, 2008). Visible Learning also generated strong interest in policymaking circles.

Even in the early days, there was criticism. One of John's colleagues commented that she was amazed he could understand classrooms to the second decimal point and that reducing classrooms to index numbers could be considered akin to reducing society to unemployment numbers, inflation rates, and crime statistics.

As the years have gone on, the Visible Learning dataset has almost doubled. It now synthesizes the findings of over 1,600 meta-analyses of 96,000 individual research studies, involving more than 300 million students, into over 270 influences on student achievement.

Yet the criticisms of Visible Learning have kept abreast and have also more than doubled. Hardly any of the criticism has been about the *interpretation* of the data or the recommendations on how teachers and school leaders should best approach the business of improving student outcomes. These Visible Learning messages seem to be generally well received.

Virtually all of the criticism has been about the research method. It's been about the way that the original studies were selected and combined, as well as the statistical techniques that were used to build the database and combine the numbers.

A good story about what works best in education is worth nothing if it's built on dodgy data. In this paper, we explicitly unpack and explore this criticism head on, so that teachers, school leaders, and policymakers understand the how and why of the Visible Learning research and the types of inferences that can (and can't) be made from it.

On our journey, we will traverse the following waypoints:

1. Building on the Shoulders of Giants

This section takes us on a brief tour from the birth of educational research in the 1800s to the present day. It explains how, by building on the shoulders of giants, we now have a thriving educational research community that has undertaken primary research involving more than 300 million students. It introduces meta-analysis as a mechanism for making sense of this huge and often conflicting dataset.

2. Meta-Analysis

Since the publication of *Visible Learning* (2009), which has been described as "teaching's holy grail" by some (Mansell, 2008) and as a work of "pseudo-science" by others (Bergeron, 2017), debate has raged on the value of meta-analysis and about whether Visible Learning tells us anything of value. This section explores the many criticisms of meta-analysis. It awards points to the critics but argues that, if interpreted properly, meta-analyses make important contributions to understanding what works best in improving student outcomes.

3. Fool's Gold?

This section specifically addresses the methodological criticisms that have been leveled at the Visible Learning approach to meta-analysis. Again, we sustain hits from the critics but argue that partial sight is better than no sight. As Erasmus (1508) put it, "*In regione caecorum rex est luscus*," which roughly translates as "in the land of the blind, the one-eyed man is king" (Bland & Erasmus, 1814).

4. What Works Best?

We then move from the research process to the analysis and interpretation of evidence. Waypoint 4 recaps the findings of the Visible Learning research—namely, teachers make the biggest difference to student achievement, and their attitudes or mindframes are key to unlocking this impact. When teachers believe that individually and collectively they can make a real difference in the lives of their learners and see themselves as evaluators of their own impact, wonderful things happen.

5. Conclusion

In waypoint 5, we summarize, tie loose ends, and bring proceedings to a close.

1. Building on the Shoulders of Giants

To know what works best in improving learning, we need explicit processes for cataloging, evaluating, and enhancing our implicit approaches to teaching. We also need to better understand the structures of the human brain and how they enable learning. If we build and implement effective processes based on these dual insights, learning becomes visible. Teachers see learning through the eyes of their students and the students develop the skills to become their own teachers.

In medicine, there are standardized tests, diagnosis protocols, and treatments that apprentice doctors must master before they are deemed fit to practice (see American Medical Association, 2017). The same goes for airline pilots, who have structured preflight checks and in-flight protocols (see Federal Aviation Administration, 2016) that have been iterated through trial and error, extensive research, and the learnings from major air disasters.

The education profession has taken a different path. The pedagogics have debated everything from what to teach, how to teach, when to teach, and who to teach. There are progressivists, traditionalists, constructivists, behaviorists, and so on—each with very different views about what works and how to implement it. Educationists have flip-flopped on phonics versus whole-word reading, intelligence testing, and the benefits of parents reading to children.

We speculate that the reason education did not traverse the path of other professions was because the stakes seemed lower and accountability was more dispersed. In medicine, if a surgeon undertakes a triple-bypass heart surgery and decides midway through to try something new, the risks of death for the patient are high and the finger of accountability will point squarely at the doctor who decided to improvise. The same goes for aviation. Commercial airline pilots don't spontaneously barrel roll or have a go at landing when they feel like it. In both industries, it's pretty black and white: what doesn't work often kills you—quickly.

Education, by contrast, suffers from perverse incentives. Most things that a teacher could do in a classroom "sorta" work (Hattie, 2009). Short of instructing children to lick their fingers and stick them in an electric socket, we can't think of many things teachers could do in a classroom that would result in an immediate risk of death to their learners. So, the short-term stakes are lower.

Of course, the stakes are just as high in the long term. Although bad teaching won't give you cancer (at least not according to the research we've surveyed), it can significantly reduce your economic, social, and emotional well-being. In this case, what doesn't work kills or depresses you slowly. And the finger of accountability is not sure where to point. Was it Ms. Trunchbull who most contributed to your undoing or Mr. Chips? Who knows? Maybe both and maybe neither.

Despite the diversity of approaches to pedagogy, classroom management, curriculum, and school administration that have emerged, there has also, thankfully, been a corresponding increase in the amount of research into what works best.

Education researchers around the world spend their lives conducting primary research into what best unlocks student achievement. They regularly conduct studies at and with schools. Unlike many related disciplines, there is no history of conducting lab studies about teaching; instead, nearly all of these studies are conducted in regular classrooms. These studies can range in size and scope from a few of days of action research with a single school to a longitudinal study involving several hundred schools.

The research methods used can also vary tremendously, ranging from comparisons of experimental (e.g., new teaching method) and control groups and from pre- to postlongitudinal studies. The research findings from these different studies, using different methodologies, are disseminated in a variety of ways, including through presentations at academic conferences and publication in books, chapters, government reports, working papers, and peer-reviewed academic journals.

As we explain in our earlier paper "As Good as Gold?," many of the quantitative studies, as a matter of course, have sufficient data to calculate an effect size (Hattie & Hamilton, 2020). So rather than telling you whether something works or not (i.e., statistically significantly different from a zero effect), it quantifies on a universal scale how powerful (or how weak) the intervention is. In other words, if something works, does it have the impact of an unarmed person, a skilled archer, or a battle tank?

Effect size is relatively easy to calculate. It requires quantitative outputs (e.g., means and standard deviations of test scores) and it requires two sets of numbers: either pre- or postintervention with a single group, or the means from an experimental and control group (for an excellent overview of effect size, see Coe, 2002).

In education research, the most common way to calculate effect size is through the use of the Cohen's *d*:

$$d = \frac{x_1 - x_2}{SD}$$

In plain English, *d* is derived by taking the mean average of a pre (x_1) and post (x_2) set of scores, calculating the difference between these two means, and dividing this by a standard deviation (*SD*) for the dataset.

The output of this calculation is a numerical value that shows the gain or decline in performance from the intervention as a proportion of a standard deviation. So, an effect size of 0.20 means that the second basket of scores was 20% of 1 standard deviation higher, on average, than the first basket of scores.[1]

Jacob Cohen (1988) also developed a scoring table to help researchers interpret their findings, which was later updated by Shlomo Sawilowsky. Our analysis of the Visible Learning database shows, very generally, that a small effect size is <0.20, a medium effect size is 0.40, and a high effect size is >0.60, but these adjectives need to be treated with so much care that they are close to useless. Context matters.

As we state in our first Gold Paper, "As Good as Gold?," what may be "small" may be life-saving (e.g., the effect of taking aspirin to reduce heart attacks is $d < 0.01$, but the side effects are low and it could be the difference between life and death). What may be small may be cheap and worth adding, but what may be "high" may be too expensive and we may need to look for alternatives. Smaller effects may have more impact on what may be very hard to change (whole systems) than what may be easier to change.

[1] Note that there are two methods for calculating effect size (pre-post and intervention comparison) and they can lead to different interpretations.

The beauty of the effect size statistic is that it is a form of universal translator. No matter what testing instrument the researcher uses, no matter how the scoring is done, and no matter the subject (math or music), the student age (4 or 20), or the country (mostly Western countries), we can make meaningful comparisons.

So long as there are at least two sets of scores (means and standard deviation), it's possible to calculate the effect size.

Since the early 1980s, many quantitative educational researchers have habitually included effect size scores in their research outputs. Indeed, most of the major education societies demand it. This means that there are currently effect size data from 96,000 studies involving more than 300 million students.

The use of effect sizes has allowed us to combine many different research studies to identify the good bets for learning and make relative statements about what works best in education. This is a massive step forward.

Meta-Analysis Is Born

The real challenge is about how we then synthesize the findings from hundreds of disparate research projects that address similar research questions into an overarching meta-analysis or study of studies. This problem of aggregation has vexed scientists for quite some time.

It was not until the 1970s that Gene Glass coined the actual term "meta-analysis" to refer to "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass, 1976, p. 3).

Unlike the earlier pioneers in the field of statistical synthesis, Glass was an educationalist. Many wrongly believe that meta-analysis was invented in medicine and applied to education, but here the case is the opposite. John was in the audience on the day of Glass's presidential speech to the American Educational Research Association when "meta-analysis" was born; the frisson in the room was palpable as many realized that the world of literature reviewing was about to change.

One of the major reasons for the invention of meta-analysis was the increased bias in traditional reviews, with authors choosing their "golden eggs" and omitting critiques, and particularly because reviewers saw large influences based on tiny effects (and often they claimed that because the effect was significantly different from zero, it was therefore a "golden egg"). Soon after, to better understand and explore how to do a meta-analysis, John and his colleague Brian Hansford (1982) completed one of the early meta-analyses, on the relation between self-concept and achievement.

The aim of meta-analysis is to use approaches, such as effect size, to develop a pooled estimate by aggregating the findings of similar studies into a wider whole. At its most basic level, the approach involves averaging the averages of the findings from each piece of research to come to an overall average. In addition to identifying underlying patterns across various studies, meta-analysis also helps us to identify major disagreements between studies and explore the potential causes of those divergent findings. Another critical advance of meta-analysis was that it was now possible to statistically explore various moderators to the overall average; for example, were there differences relating to age, gender, country, subject, and so on? This was an exciting advance.

The new approach can also be taken one step higher— a meta-meta-analysis, which is an average of the various meta-analyses that aim to explain advances in student achievement. This is the mechanism that has been used to collect the Visible Learning research.

Both the meta- and meta-meta-analysis approaches involve a structured (re)search process. Meta-analysis starts by identifying a research question, such as "Do small class sizes increase student achievement" or "Do cooperative approaches result in better outcomes than competitive ones?"

Visible Learning represented an early, not the first but possibly the largest, attempt at blending across these many meta-analyses (i.e., the collection and collation of all of the educational meta-analyses into a single meta-meta-analysis) (Hattie, 2009). We believe that both the meta- and meta-meta-analyses provide a useful detector to educators and policymakers in divining education gold. But as Nietzsche (1891) reminds us, there is no such thing as immaculate perception. We need to understand the limits of meta-meta-analysis to ensure that we do not misquote, misapply, or misuse the research. In the next section, we outline the limitations.

2. Meta-Analysis

When *Visible Learning* was published in 2009, it was never intended to be a book on research methods. Depressingly, most of the critique has been about the research methods rather than the conclusions. Most of the criticism has been about the following:

1. The validity and reliability of meta-analysis as an approach in general,
2. The rigor with which it was employed in Visible Learning, or
3. A bit of both.

This section addresses point 1, the general criticisms about meta-analyses, clearing the ground for us to explore the specific Visible Learning–related critiques in section 3 ("Fool's Gold?"). Before we begin, we offer a word or two on the nature of claims and evidence in research.

In scientific, social, and educational research, we can distinguish between four types of claims: predictive, descriptive, speculative, and controversial.

No.	Type of Claim	Description
1	**Predictive**	The ability to make accurate predictions about future outcomes by reviewing patterns in historical data to develop universal laws, which foretell outcomes with precision—time and time again.
2	**Descriptive**	The ability to derive general principles about the past that seem to apply to the present and might have some value in the future.
3	**Speculative**	The ability to tell a story about the past with some degree of plausibility, which may apply to the present and possibly the future.
4	**Controversial**	The ability to spin a yarn about the past that is loosely based on the data but where there are several other equally plausible explanations available. This means that the yarn, likely, has limited application to the present and even less to the future.

What we want to make clear is that, along with all other educational research, Visible Learning does not fit into category 1 (predictive). This is because the ecosystems that Visible Learning attempts to map are complex and have many moving parts with many interdependencies. There are no known universal laws of learning (yet). At best, the findings from Visible Learning can be considered "probability claims": *if you implement x under y conditions, there is a high probability you will have z effect.*

The best we can hope for is to describe past outcomes and to speculate on their replicability in the future. The more data we have in hand, the more confident we can be in these speculative claims and, inversely, the less data, the less confident.

In some parts of the Visible Learning dataset, the evidence collected is broad and deep. In these situations, we can be more confident in our conclusions. In other places, conclusions must be more speculative and more controversial.

The Validity and Reliability of Meta-Analysis as an Approach in General

Some of the critiques of Visible Learning have argued that the whole enterprise is fatally flawed because it relies on synthesizing equally flawed meta-analyses, which has been labeled as "mega-silliness" (Eysenck, 1978) and as "statistical alchemy for the 21st century" (Feinstein, 1995). In the subsections below, we outline the main arguments.

The Argument From *Reductio ad Absurdum*

A common argument against meta-analysis is that complex phenomena cannot meaningfully be reduced to a number (Bailar, 1997; Borenstein, Hedges, Higgins, & Rothstein, 2009). This is because too much fidelity is lost in the recoding of real-time interactions between students and educators (which are deep, rich, and value laden) from a range of education contexts into a numerical effect size.

However, what you see largely depends on where you sit; and meta-analysis gives us a powerful step ladder. Each upward step enables us to take in more of the view but we get further and further away from the individual blades of grass. With an oxygen tank and a good set of thermals, we can continue climbing the ladder right up to 27,000 feet. At that height, we can see the "bigger picture"—for example, the widespread forest fire of an ineffective approach to teaching and learning. But it's only when we climb back down the ladder to 50 feet that we can see the pockets of effective practice in the midst of the fire. We just have to accept that we can't see everything all of the time.

In fact, there are many contexts in which simplification of data helps much more than it hinders. For example, temperature in the form of Celsius or Fahrenheit is a form of meta-analysis. When we take the temperature in a room, it is the average of the velocity of hundreds of billions of molecules. Some of these molecules move at extremely high speeds and are "hot" and others move much more slowly and are "cold." The temperature measured by our thermometer is the average motion of all these molecules. Nothing more and nothing less.

The current meta-meta-analyses and their numerical findings add an important perspective on classrooms. Of course, this is only as good as the fidelity of the interpretations.

Our major claim is *not* that meta-analysis and Visible Learning lead directly to specific policies about what to do. At best, they offer probabilities about likely high-impact interventions. But key still is the fidelity of the implementation in the local context and the skills of teachers to adapt and modify in light of their evaluations of the implementation in terms of the maximum impact on the learning lives of their students. Hence, know thy impact.

The Argument From Apples and Oranges

Others have argued that the problem isn't so much about quantification itself. They say it's more that the meta-analysts are attempting to bundle together different types of data that shouldn't be combined (Borenstein et al., 2009). There are two dimensions to this argument.

1. **The Perils of Combining Data From Different Tests.** The first dimension centers on calibration challenges with similar measuring instruments. It's basically about the fact that different experimenters use different tests to measure the same aspects of student achievement. So, how can you meaningfully blend together the findings from these studies

(of, say, student literacy) if one used locally developed multiple choice tests, another used professionally developed formative assessment tools, and yet another used summative examination results? These are, after all, different tests with different levels of sensitivity and different reporting outputs.

If we go back to our temperature example, this would be the equivalent of taking heat measurements in different times and places with a glass thermometer, a thermistor, a resistance temperature detector, a pyrometer, and an infrared device. Each of these measures the same thing but does so in a different way and with the potential that readings are not quite the same.

But the challenge is not insurmountable. It either requires cross-calibrating the instruments in advance or checking the error margin between the instruments and adjusting for this after the fact. Granted, in education research, where investigations are conducted using standardized achievement tests, this is more complex. It requires the meta-analysts to carefully review the tests used in the primary research and to make a judgment call about the reliability and validity of each instrument. There is potential for error, but this can be mitigated through careful selection of studies for inclusion.

2. **The Perils of Combining Data From Completely Different Contexts.** The second dimension is what's often referred to as the comparing apples with oranges problem. Imagine that instead of combining different tests of student literacy, you instead combined completely different categories of data (e.g., IQ tests, lesson observation scores, psychometric surveys, and assessment data) into the same pot.

   Sticking, again, with our temperature example, this would mean that instead of taking temperature measurements with different types of thermometers, we did it in completely different ways. In some instances, we measured cloud cover; in other cases, we recorded survey data from a range of participants who we asked to estimate the temperature; and in yet other cases, we measured the volume of ice cream sales.

   Although this makes direct comparison difficult, it's not entirely impossible. If we have access to historical data correlating cloud cover with ambient temperature, the average margin of error in people's estimates of temperature with actual temperature, and ice cream sales with temperature, we can, with a margin of error, make a good estimate at figuring out the temperature without a thermometer.

   Arguably, the same principle applies to any form of literature synthesis including educational meta-analysis, where it is possible to combine the results from studies that draw on quantified assessment data, attitudinal surveys, observational data, IQ tests, and so on. Some argue that this is like comparing apples and oranges but, in reality, it is more like *combining* them into a higher-order category called fruit—in exactly the same way that temperature is a higher-order category of the velocity of molecules.

Cognitive Bias

Another group of critics argue that we cannot trust the primary research that meta-analyses are drawn from. This is because even the most rigorous randomized controlled trials (RCTs), in which participants are divided into separate groups that either receive or do not receive the treatment, cannot be trusted. Unlike medical trials that are generally double blind (i.e., where neither the experimenter nor the subjects know who is receiving the *real* treatment), even single-blind trials (where only the subjects are unsure of whether they are in the treatment group) are virtually impossible to conduct in education (Sullivan, 2011).

Let's unpack this. In any research study, there is a risk that if the participants know that they are receiving a treatment, it may lead them to subconsciously change their behavior in anticipation that the treatment will cure their malady. In the literature on cognitive bias, this is referred to as the placebo effect, Hawthorne effect, or observer expectancy bias. In medical trials, researchers counter for this by double blinding. This means that trial participants are unaware of (or are blind to) whether they are getting the experimental medication or a sugar tablet. The same goes for the experimenters; they have no idea which tablets they are dispensing to whom. And this means that neither party can subconsciously influence the outcome of a trial.

The challenge with education research is that double blinding is a virtual impossibility. If a study involves, say, using a new literacy program, the teachers delivering the intervention cannot be made blind to the fact that they are doing something different. And often when teachers try new things, they do so with renewed enthusiasm; so how can we tell whether it's the enthusiasm or the literacy program that's driving up student achievement? The same goes for the students. Unless we lock them in a dungeon between classes, we can't stop them from mixing with other groups of learners and comparing notes about their respective classroom experiences.

An added challenge is that in many research projects, the experimenters were the ones who devised the intervention, so they have a vested interest in finding positive results. They don't want to feel foolish for spending months or years of their respective lives on a folly that's no better than average—so they *will* their findings to be higher and deliver the intervention in an expert manner that no other teacher could replicate. And *abracadabra*: 200 tons of education pyrite.

This was, in fact, one of the reasons Glass developed meta-analysis. He believed it would actually reduce bias! By combining enough studies together, we have access to a more robust and nuanced picture than we would if we just referred to one or two small-scale studies. It means that we can draw on and interpret the collective wisdom of the crowd, rather than being overly swayed by one or two extremely loud voices.

We both agree that experimenter and participant bias are significant challenges in education research. In fact, two of the interventions with the highest effect sizes in the Visible Learning database are collective teacher efficacy (d = 1.57) and student self-efficacy (d = 0.92). Both are measures of beliefs about teaching and learning. If teachers really, really believe that individually and collectively they can make a difference to learning and if students believe the same about themselves, magical things happen. Belief is the key.

The File Drawer Problem

Building on the argument that everyone loves a winner, many critics suggest that researchers are disinclined to publish studies with unfavorable outcomes. This is the file drawer problem.

The argument is that if a research finding is promising, it gets published. If it's a howler, it stays filed (and hidden) in the drawer (see Rosenberg, 2005).

Meta-analysts can only combine the studies to which they can get access. Although they can also speculate about the other types of studies still filed in a bottom drawer, they can't collect and sample them.

Robert Rosenthal (1979) introduced a method for estimating a fail-safe number of unpublished articles with an average null effect that would be needed to counter the reported effect size. With his method, the reader can sense whether this number is convincing enough to threaten the overall conclusion drawn by the researcher.

Another way to reduce the number of studies sitting unpublished and contrary in file drawers is to raise the bar. This is one of the reasons that Visible Learning sets the effect size bar so high (i.e., $d > 0.40$) and compares those above and below this hinge point. This helps us to weed out false positives, which are more likely to have lower cumulative effect size values, and focuses everyone's attention on the interventions with the highest probability of impact.

Yet because meta-analysts are passive collectors and aggregators of research, they can still only analyze what others have chosen to research. This means that there are potentially major gaps: some areas have been overmined, others lightly mined, and in yet others no gold has been retrieved at all.

Effect Size Is Stupid

Other critics have argued that effect size is not a proper tool that's used in statistics or even found in statistical text books (see Coe, 2002). Formulae for the calculation of effect size are, admittedly, absent from many earlier introductory statistics textbooks. Effect size is also absent from some statistics software packages and is rarely taught in Research Methods 101.

The trouble with this argument is that it conflates lack of circulation with lack of value. It's a bit like saying that a Rembrandt painting is no good because hardly anyone has one and only art connoisseurs can tell you what one looks like.

On the contrary, effect size data have been calculated for more than 75 years and the American Psychological Association (2009) has been officially encouraging researchers to include effect size data in their research since the mid-1990s. There are over 1,600 meta-analyses in education, and probably five times more in medicine and many more in other domains, so it is hard to argue that effect sizes are not a proper statistic. We invite any doubter to read Hedges and Olkin's (1985) excellent treatise on the foundations of effect sizes and then say there is no statistical basis!

Driving the Car Backward

Yet other critics point to the fact that meta-analysis is a rear-view mirror activity. In other words, it involves looking at old fruit rather than the buds that are currently forming. There's a lot of truth to this. Indeed, this is the essence of most re-searching—searching again for new ideas and interpretations. Some of the students who participated in studies that appeared in the earliest educational meta-analysis are likely grandparents by now.

So, the question is whether we can rely on these "elderly" data—some of which were collected in the 1950s to 1990s—to make decisions today. In some areas, we must be especially cautious. In our forthcoming "Not All That Glitters Is Gold," we highlight the fact that one of the challenges with the research in education technology is that the tech moves so quickly that it is difficult to make direct comparisons between different studies. Although what is surprising in this tech research is that the average effect has barely changed over the last 35 years despite major changes in the technology.

We can also ask the question about the changes over time by dividing the average effects by decade, or by correlating the size of the effect size with the publication date of the article. The point is that it is an empirical question whether time of study makes a difference. And this is a major feature of meta-analysis.

It's Not What You Do, But the Way That You Do It

A final area of criticism that's linked to both the rear-view mirror problem and the argument from *Reductio ad Absurdum* is how the heck do we actually implement? The challenge for anyone who reads a meta-analysis is that the landscape is analyzed from the very top of the step ladder. While we can make sense of the scene and the recommendations about "what works best," there is rarely a step-by-step recipe alongside that tells you how to implement it with fidelity in your context. In order to access that recipe, we have to map back from the meta-analysis in question to the individual studies that it surveys. But even this may not be useful.

The quality and fidelity of implementation is a perennial problem that strikes at the very heart of the whole educational improvement endeavor. In our forthcoming publication "Getting to Gold," we tackle this issue head on. We argue that the process of implementation, or what Michael Barber calls *deliverology* (see Barber, Moffit, & Kihn, 2011), is crucial. The meta-analyses can help you to home in on good bets for the "what," and implementation protocols provide the "how." We argue that good implementation requires a systematic approach that does the following:

1. Enables the identification of "wicked issues" that enough people in the room feel passionate about resolving.
2. Involves the development of multiple theories of change about how to solve the problem, by building a causal loop back from problem to interventions and tactics that could potentially lead to a solution.
3. Uses the data from research, including meta-analysis, to weed out the theories of change and zero in on the better bets.
4. Implements one or more of those better bets.
5. Reviews and tinkers to improve outcomes in your context. This involves reversing back from adjustments that don't yield gold and iterating things that do, to see if the yield becomes even stronger.
6. Repeats steps 1 to 5 over and over and over.

This is an approach that we have iterated ourselves, through trial and error, on Visible Learning systems-level improvement projects in the Northern Territory of Australia and also in Scandinavia.

We can use the findings from the meta-analyses to help choose high-probability interventions, achieve excellent implementation, give attention to what is and what is not working in the local context, and foster a commitment to focus on the evidence of impact on the learning lives of students. Through this we inch ever closer to an implementation science for education.

For a comprehensive list of the common Visible Learning critiques and my responses, please read the accompaniment to this paper, "Common VISIBLE LEARNING™ Methodology Critiques and Responses," available at https://www.visiblelearningplus.com/content/gold-papers.

3. Fool's Gold?

In the previous section, we discussed the various criticisms that have been leveled at meta-analysis in general. In this section, we zoom in on the criticisms that are specifically directed at the Visible Learning research, which is probably the largest attempt to combine the findings from different meta-analyses into a meta-meta-analysis. We also provide an appendix with more criticisms and responses, available online at https://www.visiblelearningplus.com/content/gold-papers.

The Challenges of Ranking

The Programme for International Student Assessment (PISA) ranks schools, QS ranks universities, the English Premier League ranks football (soccer) teams, and the Dow Jones ranks companies. Many of us follow these and other rankings with great interest. However, one of the common criticisms of *Visible Learning* (2009) centers on the fact that it contained an appendix listing the various influences on student achievement in ranked order and that these rankings have been perpetuated in public presentations about the research (Terhart, 2011). The argument is that this ranking creates the perception that by simply doing more of the things at the top of this list and perhaps stopping things ranked at the bottom, great things will happen.

We agree that interpreting the Visible Learning research simply by looking at the rankings is not helpful. It supposes that the world of education improvement is a one-dimensional affair rather than a complex ecosystem with inter-related moving parts. And what may be "small" may be life-saving (as we previously mentioned in our example of the effect of taking aspirin to reduce heart attacks). It also implies that each influence is separate and merely adding them together is defensible. Indeed, it took almost 20 years to write *Visible Learning* because of the time required to understand the relation between the many influences and to develop a model of why some are above and others are below the hinge point.

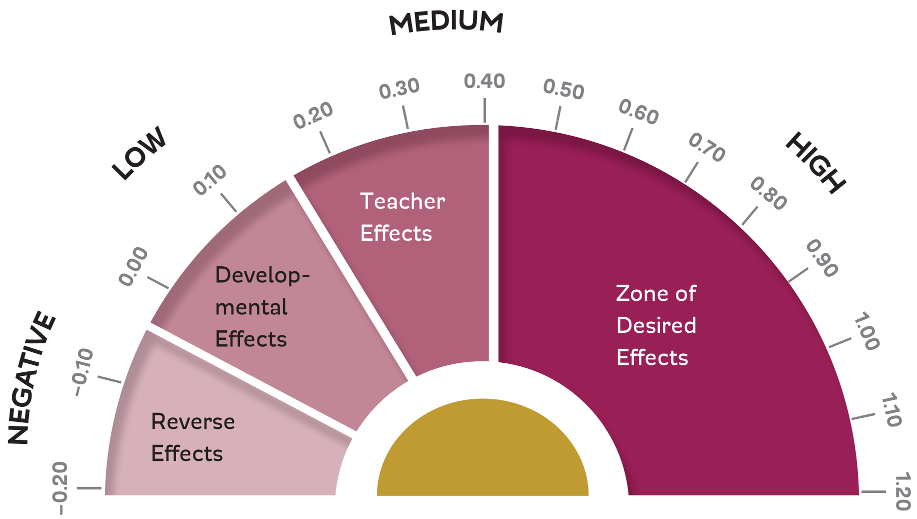
Consequently, in the latest publication of the research database, we have dispensed with rank ordering and instead organized the influences under seven domains and thirty-two subdomains. Within each subdomain, we listed each influence in alphabetical order, regardless of its effect size. We can't stop other people from ranking, but we hope that this simple shift in organization also shifts the way people interpret the research and apply it.

Use of the *d* = 0.40 Hinge Point

Hinges are great. They enable us to swing wardrobe doors and the arms of our eyeglasses from one position to another, with minimal use of force. Even our knees and elbows are hinges.

In *Visible Learning* (2009), both the analysis and the effect size barometer graphics presented $d = 0.40$ as a hinge point of sorts (see the following figure). The argument made was that in rule-of-thumb terms, any influence on student achievement that generated an effect size greater than 0.40 was, on balance, likely to be worth investing in.

At the time, the argument for the $d = 0.40$ hinge point was made on the basis that when the effect sizes from all 800 meta-analyses were averaged together, their mean average score was 0.40. Anything lower



than *d* = 0.40 was, by definition, "below average." Since then, the database of meta-analyses has grown to more than 1,600 and interestingly that mean average effect size across all influences has not really changed. Today it still stands at *d* = 0.40.

But we must not get too oversold on using *d* = 0.40 in all circumstances. The interpretation can differ in light of how narrow (e.g., vocabulary) or wide (e.g., comprehension) the outcome is, the cost of the intervention (Simpson, 2017), the challenge of learning how to implement intervention, and many other factors. When implementing the Visible Learning model, it is worth developing local knowledge about what works best in the context and not overly rely on 0.40. The 0.40 merely is the average of all 1,600 meta-analyses and serves as a worthwhile hinge compared with the usual zero (which allows nearly all to claim that their favorite strategy or influence can enhance achievement).

Common Language Effect Size

In the data tables at the back of the first edition of *Visible Learning* (2009), there was a column that reported common language effect size or CLE. This was included because one of the challenges with reporting effect size statistics is that very few people understand what they mean.

CLE was developed by McGraw and Wong (1992) as a way of communicating effect size in terms that lay readers would be more likely to understand. It is defined as "the probability that a randomly selected score from the one population will be greater than a randomly sampled score from the other population" (McGraw & Wong, 1992). CLE is presented as a percentage from 1% (i.e., much worse than chance), 50% (i.e., no better than chance), to 99% (i.e., near certainty) that a randomly selected score from one population will be greater than a randomly selected score from the other. The higher the CLE, the higher the probability that a randomly selected member of the treatment group scored higher than the control and that the treatment "works."

Embarrassingly, a coding error was made when the incorrect column was transposed into the final tables in the appendix of *Visible Learning* (2009). No one noticed this for almost 4 years, until it was picked up by a group of Norwegian students (see Topphol, 2012). That it took this long for anyone to spot the error gives food for thought about how many actually used the CLE. But it has been rectified in future editions and was only ever intended as a supplementary statistic buried in an appendix. In fact, in the most recent version of the data, the CLE has been removed completely. It seems that people were ready for effect size after all, which is perhaps why the CLE miscalculation was missed. The error changed the story and the major messages not one iota.

Garbage In, Garbage Out

We all know that the quality of a meal is strongly determined by its ingredients and that even Chef Gordon Ramsay would struggle to conjure up something worthy of a Michelin star if all he had on hand was tinned spaghetti and a slice of processed cheese. Another type of criticism leveled at the Visible Learning research has parallels with this. The argument is that the quality of the research included in the Visible Learning dataset is more akin to junk food than organic produce grown with love (Snook, O'Neill, Clark, O'Neill, & Openshaw, 2009). If the ingredients are junk, so the criticism goes, then so must be the results.

The argument is that the studies in the various meta-analyses included single or no blind studies (correct), that they used RCTs and pre-post and group comparison methods (correct), and that they used a variety of testing instruments ranging from standardized student achievement tests, IQ tests, self-perception survey data, quantified observations, teacher-made tests, and correlational data (also correct). Finally, the argument is that many of the studies are quasi-experimental and have no real control group and that others are much weaker correlational studies (correct again).

It is suggested that by including all of these types of data, rather than just the "gold standard" RCTs, the quality of the research has been significantly compromised. We agree (although we use "beyond reasonable doubt" and not RCTs as the gold standard), but it was a different kind of compromise that was made. If we only included the perfect studies or meta-analyses, there would be insufficient data from which to draw conclusions. Indeed, in the What Works Clearinghouse, which only allows RCTs and similarly high-quality designs, the median number of studies in each of the 500 reviews is two! It is hard to draw conclusions based on two studies.

So, we have a choice to make. We either limit ourselves to collecting the perfect studies or we mine the lot but take great care over how we interpret the data and the conclusions that we draw from them. In the case of Visible Learning, the latter approach was taken. We can also ask whether the quality of the study of meta-analysis makes a difference to the overall conclusions (and this is common practice with rare cases where quality makes a difference; indeed, it is more likely that quality matters when the average effect size is close to zero). We would rather be able to say something (nuanced) about everything than a lot about very little.

However, we think that more could be done to signal to readers about which research findings are more reliable and those that are speculative or even controversial. In England, the Education Endowment Foundation (EEF) took this step in their database of influences on student achievement. They included a padlock icon and the more padlocks that were displayed against an influence, the more secure the research findings.

We have taken the learnings from EEF and are implementing a similar confidence rating system to rate the quality of research, which is available on Visible Learning MetaX. Now we score each influence depending on the number of meta-analyses, studies, students, and effect sizes under each influence. We are also experimenting with additional weightings that take into account the predominant experimental design used in the research studies in each meta-analysis, although this is more difficult to do with fidelity (see Hattie & Zierer, 2018). However, this step is important because some of the "controversial" influences like Piagetian programs (*d* = 1.28), teacher credibility (*d* = 0.91), and one-on-one laptops (*d* = 0.16) are either based on a single meta-analysis and/or a small pool of research studies. Therefore, how these findings are interpreted and juxtaposed against the influences where we have higher confidence in the findings is key.

We are also working back through the 1,600 meta-analyses to specifically tag countries where the original research was conducted and overtly display this. We are taking this step because the majority of the research comes from developed English-speaking nations that are part of the G20. We can be relatively confident that research conducted in the United States will have some relevance or transferability to the United Kingdom, Canada, Australia, and New Zealand and vice versa. But we need to be careful about transposing this to developing country contexts like Sub-Saharan Africa or South Asia. The two ecosystems are quite different. In the developed world, education policy is focused largely on quality of provision and equity of outcomes. In the developing world, the challenges are all too often around access and infrastructure. And the data from one "world" tell us very little about the other.

Mr. Market

Benjamin Graham, the father of value investing and mentor to Warren Buffett, described the rollercoaster ride of the stock exchange through the allegory of Mr. Market. This Mr. Market is a manic-depressive type who swings wildly in mood. In the morning he is bullish, but in the evening he is bearish. He changes his sentiment about the value of companies as quickly as you might change your socks. To Mr. Market, the winners become the losers and the losers the winners and then back again.

One of the criticisms of meta-analysis and of the Visible Learning research is that it, too, suffers from a dose of Mr. Market. The argument is that as new primary research and meta-analyses are generated, these wildly inflate or dilute the messages of Visible Learning, which can never stay constant in light of the changing ticker tape of data.

To be fair, there is some truth in this—but not a lot. In the following table, we unpack the influences on student achievement that have suffered most at the hands of Mr. Market.

	Effect Size				
Influence	**2009**	**2012**	**2017**	**2019**	**Comment**
Teacher credibility	N/A	0.74	0.90	1.09	This influence relates student perceptions of teacher credibility and the impact of this on student achievement. It was not included as an influence until *Visible Learning for Teachers* (2012). Although there has been an increase in the effect size (ES) from 0.74 to 0.90 for this influence, the messaging or story around this has not changed: it has just been reinforced. That is, student perception of teacher credibility is very impactful.

	Effect Size				
Influence	2009	2012	2017	2019	Comment
Teacher-student relationships	0.72	0.72	0.52	0.48	Teacher-student relationships are about the quality of the relationship between teacher and student and the impact this has on student achievement. The research suggests that positive relationships enhance student outcomes. Although the ES has decreased, this is because the more recent meta-analyses were actually measuring subcomponents of this influence: • Nurmi (2012): This meta-analysis has an ES of 0.20 and focuses on students' sense of "closeness" and lack of conflict with the teacher. • Moallem (2013): This meta-analysis has an ES of 0.45 and focuses on students' sense of "belonging" in the class/school. Both of these effects are a bit lower than the other three, which are similar to each other. The original meta-analysis (Cornelius-White, 2007) has the highest ES (0.72) of all five studies. So, overall teacher-student relationships are still very important, but some aspects (e.g., closeness/belonging) appear to be not as important as the overall sense of whether or not the relationship is positive.
Providing formative evaluation	0.90	0.90	0.48	0.34	In the 2009/2012 dataset, there was only one meta-analysis on formative evaluation. Since then, there has been an additional meta-analysis and the reclassification of another meta-analysis that was previously counted under this influence. Both have contributed to the quite dramatic lowering of the ES for formative evaluation. As with feedback, the effects measured vary a great deal and the explanation as to why formative feedback is impactful or not could well be due to the type of feedback that is sought/received and how well it is acted on. The Visible Learning narrative is that providing formative assessment has the *potential* to be very powerful but how targeted, how specific, and how well the feedback is actually received, understood, and acted upon have a big impact on its efficacy for improvement.
Study skills	0.59	0.63	0.46	0.45	There have been two new meta-analyses added since 2012, which have brought the average ES down. The story still stays the same. The impact of study skills depends on the study skills being taught and when and where in the learning they are being used; that is, different skills are more or less effective than others

	Effect Size				
Influence	2009	2012	2017	2019	Comment
					and depend on which phase of learning they are being employed. In addition, teaching students "how" to study needs to be done in the context and/or alongside the learning area and not as an independent program in "how to study" to really get better effects. We recently completed a new synthesis of many learning strategies (Hattie & Donoghue, 2016) with an overall ES of 0.53.
Worked examples	0.57	0.57	0.37	0.37	In the 2012 list, there was just one meta-analysis on worked examples. Since then, one more has been added (Wittwer & Renkl, 2010), with an ES of 0.16. Because this is a less reviewed area of research, there is more likelihood of the effects changing as new research comes in. Worked examples are still regarded as a strategy that has the potential for moderate positive impact on student achievement.
Student-centered teaching	N/A	0.54	0.36	0.36	Two new meta-analyses have been added, in addition to the single meta on this topic from the 2012 dataset. The newer meta-analyses from Thomas et al. (2012) with an ES of 0.16 and from Bernard et al. (2013) with an ES of 0.37 have both changed the picture. The meta-analysis by Bernard et al. included over 290 studies and effects, which is much bigger than the other two studies, and so its findings have the dominant impact. The focus of much of the research included in these studies links to computer/technology-assisted environments and across a range of subjects where students are using these independently vs. more traditional teaching methods. This contrasts with the earlier meta-analysis by Preston (2007), which was small and focused on use of student-centered approaches in mathematics only.
Classroom management	0.52	0.52	0.35	0.35	In 2012, there was only one meta-analysis included under this category (Marzano, 2003, with an ES of 0.52). Since then, one new meta-analysis has been added (Korpershoek et al., 2016), which found a considerably lower ES (0.17). Both of these studies are looking at a wide range of factors contributing to an overall "classroom management." The overall finding that classroom management interventions are generally effective in enhancing student outcomes is in line with the findings of prior meta-analyses, so the story has not changed. But we need to unpack with greater care the specific element of classroom management that works best.

	Effect Size				
Influence	2009	2012	2017	2019	Comment
Pre-school programs	0.47	0.45	0.26	0.28	There have been five new meta-analyses of pre-school programs conducted over the last 5 years. These have found quite consistent small-to-moderate effects of the impact participation in a pre-school program has on school achievement in the first few years. In fact, of the 12 meta-analyses included in the Visible Learning database, the one outlier is that of La Paro and Pianta (2000), with an ES of 1.02. The other 11 are much more modest. So, the story is that involvement in pre-school programs is likely to have a small-to-moderate positive impact on students' early years learning but that by about year 4/5 of school, most students who were not involved in pre-school programs will have caught up.
Collective teacher efficacy	N/A	1.57	1.32	1.39	Collective teacher efficacy (CTE) is a relatively recent school-level construct subjected to meta-analysis (and of course, there have been many studies since Bandura promoted the notion in 1987) and it was therefore not included in the 2009 dataset. It is defined as the collective belief of the staff of the school/faculty in their ability to positively affect students. CTE was found to be strongly, positively correlated with student achievement. A school staff that believes that they can collectively accomplish great things is vital for the health of a school. The inference from the strength of this correlation is that interventions focusing on developing strong CTE may be a good starting point for positive change within the school system. The findings were consistent no matter what the subject area. But a key point of caution is that there is currently only one meta-analysis of 26 research studies. The evidence base is still too small to form anything more than speculative conclusions.

As we gather and review more data, it is to be expected that the average effect for some influences may change over time. If we didn't expect this, why would we bother to continue panning for this precious gold? But despite this, the addition of new data has resulted in evolution rather than revolution. As we shall go on to outline in section 4, the core messages of Visible Learning have been remarkably consistent during the last decade. We must, regardless, continue to see evidence that we may be wrong. This is consistent with how many philosophers argue how science progresses. All of us should also search for evidence to falsify a model. John continues to seek and add meta-analyses, as he wants to be first to decree if and when the Visible Learning model is wrong; so far, there has only been (wonderful) confirmation of the Visible Learning story, but the search must continue.

Overlapping Data

Where there is doubt in the Visible Learning research about whether a newly discovered meta-analysis fits into an existing category, often a new influence is added to the database. Back in 2009, there were 150 influences in the original database. In 2019, this now stands at 273 influences and counting. The challenge with this is that some are like nesting Russian dolls and are overlapping. They are "suitcase influences" and when we open up the suitcase and look inside, we see any array of socks, shirts, and toiletries that are subinfluences or moderators within the same category but which, in Visible Learning, are recorded as influences in their own right. We recognize that this can be confusing, and we hope that the new organization of the influences into domains and subdomains helps to clarify the overlapping of data.

Beyond the School Gates

A final barrage of criticism comes from those who argue that the Visible Learning research does not place sufficient emphasis on out-of-school variables (see Snook et al., 2009). We think this criticism is a little unfair because the research database reviews the Home domain, including family dynamics, family structure, and family resources. But it is true that the analysis and the interpretation of the data looks far more closely at the within-school influences. There are two reasons for this. The first is that the intended audience for the research is teachers and education policymakers, so it makes sense to focus much more on things that they can do something about. Neither teachers nor Ministers of Education can quickly or easily change the home environment of students but if they work together, they can significantly improve the school experience for learners from all home environments. And improving that school experience alone is enough to make a difference in the learning outcomes of *all* students.

The second reason for reducing emphasis on out-of-school influences is that schools can offer all students the opportunity to gain at least a year's growth for a year's input, regardless of home background. If out-of-school influences like socioeconomic status and the composition of the family are given primacy, teachers have an excuse for why they have not achieved a year's growth for a year's input. We have met teachers who explain away shortcomings in learning and growth as a deficiency of the students. Of course, students do not leave their home backgrounds at the school gate, so awareness of what students bring from the home is important and needs to be considered by schools in order to make differences in the learning lives of all students. There should be nowhere to hide, because anything less than a year's growth for a year's input is utterly unacceptable.

Toward a TripAdvisor for Education

Before we try new restaurants or book untested hotels, millions of us visit sites like TripAdvisor to review the experiences of other customers. These types of sites contain hundreds of millions of user-generated reviews of peoples' direct experiences. By reviewing this historical dataset, we collectively hope that we will be able to vector in on the best eateries and hostelries and sort the gold from the pyrite.

Of course, experienced users of such sites know that the data must be interpreted with great care. When we review the overall star rating for a venue (which is the mean average of all reviewer ratings), we have to be especially careful. We know that some establishments can get to the top of the rankings with a very small total number of reviews. And other venues with a thousand reviews can rank much lower, even if several hundred of those thousand reviewers have given them the top rating (*The Challenges of Ranking*).

When we read the actual reviews, we also quickly recognize that the different reviewers do not share an objective or even intersubjective set of standards about what the perfect hotel or restaurant looks like (*Comparing Apples and Oranges/Garbage In, Garbage Out*). One person's delight is, to another, the very definition of misery. So, we have to understand that when we review scores on sites like TripAdvisor, we are often comparing reviews derived from different senses of satisfaction and style.

It is also likely that, like one of us, some reviewers are only motivated to write about either utterly amazing or diabolical experiences (i.e., *The File Drawer Problem*). And we have all heard rumors about venues that hire PR companies to write flattering reviews or to contact folks who leave negative feedback to encourage them to moderate their opinions.

Once we have filtered the information, we then visit, say, a restaurant in utter excitement after reading the string of glowing reviews. But this is no guarantee that the actual soufflé we are served that day will be "perfect" like all the others (i.e., *Driving the Car Backward/Mr. Market*). Chefs have bad days, ovens play up, and some eggs are better than others.

At other times, perhaps we give a venue a little too much benefit of the doubt. When our experiences are out of kilter with the reviews, maybe we question ourselves and moderate our opinion to fit the data (i.e., *Cognitive Bias*).

Knowing all this does not discourage us, or many hundreds of millions of others, from using sites like TripAdvisor. We all understand the limitations of the methodology and most of us are careful in how we interpret the data. We would rather have access to an abundant supply of imperfect data that we can interpret than rely on one or two comprehensive but outdated reviews by professional restaurant critics. In fact, if you are anything like us, you might actually *enjoy* interpreting the data for the prize of an outstanding meal.

The same methodological challenges apply to meta-analyses. But we hope that, like us, you conclude that having access to a dataset on historical learning outcomes for 300 million students in a range of different contexts is like gold dust. It's just that we all need to get better at applying the analytic skills we use on TripAdvisor to *interpret* these educational data effectively.

4. What Works Best?

The Visible Learning research is based on a synthesis of (now) over 1,600 meta-analyses in education. This synthesis was compiled to address a vexing question in education: Why is it that we can we find so many studies that seem to provide "evidence" that each pet influence, method, or policy works?

From the 300 million students across the 96,000 studies in the Visible Learning database, it turns out that over 95% of influences on students have an average effect greater than zero. So in this sense, teachers and policymakers are correct to argue that there is evidence for most of what they do.

But the average effect of all influences is much higher, $d = 0.40$. The Visible Learning story is about the common themes underlying those influences greater than this average compared to those below this average (noting that nearly all can increase achievement). Ergo, we need to stop asking "What works?" and replace it with "What works *best*?" because almost everything "sorta" works.

As we have outlined in the previous two sections, most of the criticism about Visible Learning centers on the research methodology rather than the interpretation or prescriptions from the data. In these preceding sections, we have spent (a lot of) time unpacking, defending against, and occasionally sustaining blows from the slings and arrows of that criticism. But overall, we believe that we have justified the value of an educational TripAdvisor and highlighted the limitations of meta-analysis so that teachers, school leaders, and system leaders know how to use it better. Now that we have done this, we want to get back to the core messages of Visible Learning. Although we recognize that since these are not anywhere near as controversial as the perception of the research methods, we may well be preaching to the converted.

The Visible Learning Story

The major message of Visible Learning is "know thy impact" (Hattie, 2012). That is, teachers, school leaders, and system leaders need to ask about the merit, worth, and significance of their selected interventions—in the classroom, in the staffroom, and in the policy sector.

From the research, we know that the following things matter the most:

1. **Achieving teacher collective efficacy.** To achieve collective efficacy, teachers work collaboratively to plan and work together to critique their expectations, evaluate their impact on students, and decide where best to go next in light of their impact. Teachers are one of the most powerful influences on student learning. Collective efficacy is about teachers believing that by working together, they can have a significant positive impact on student achievement. The research tells us that where teachers have this shared belief, great things happen. Where teachers *don't* share this belief, things happen. This is self-fulfilling prophecy land.

   When school leaders create a trusting environment and provide the time and direction for teachers to act in ways that develop their collective agency (by collaboratively planning, sharing their teaching, giving each other evaluative feedback on what is working well or not so well in their teaching, and when they actively seek and act on feedback from their students about how the learning is going), then they can achieve significant effects on student outcomes.

2. **Knowing student prior learning.** When teachers know and understand the prior knowledge of their students and what they bring to the class in terms of dispositions and motivations, and then adapt and respond to this background information, then they can also achieve significant gains in student learning. If students are taught to have a clear understanding of their own prior knowledge and where their strengths and gaps lie, this has an equally powerful effect on achievement. Teachers need to enable students to have the opportunity to reflect on their own learning needs and to develop skills in self- and peer assessment.
3. **Emphasizing success criteria.** When teachers and students use practices that emphasize the appropriately challenging learning intentions and success criteria that are being evaluated or sought in the learning activities, students will be more likely to invest in their learning and will be able to achieve more highly and consistently. Everyone then knows where to focus their energy, time, and thinking.
4. **Using feedback, welcoming errors, and building trust.** Learning outcomes improve when the following take place:
   a. Teachers and students actively seek, receive, and act upon feedback and feed-forward about their impact (on whom, about what, to what magnitude).

   b. There is a trusting learning relationship existing between the teacher and students and between the students in classrooms.

   c. Errors are welcomed and the learning climate is safe and ordered.
5. **Structuring for deep on top of surface learning.** When teachers have developed the learning sequence in such a way that students first develop relevant and needed surface knowledge (the content or ideas) and then go on to develop deeper understandings (relate and extend these ideas), this combination of surface to deep learning can have a significant impact on improving student learning outcomes. For example, problem-based learning has been shown to be ineffective when students don't have enough relevant and necessary prior knowledge to be able to make their own connections and solve the problems. However, when students do have good surface knowledge, then problem-based learning has been shown to be very effective in consolidating and developing deeper understandings.
6. **Holding high expectations and the right levels of challenge.** We know that teachers need to get the level of challenge "just right." This is the Goldilocks principle of engaging students in goals and learning that is a challenge for them but where that the level of challenge is neither "too hard" nor "too boring." When this optimal zone of challenge is achieved, then students are often more engaged and motivated to learn and able to maximize their learning.

None of the "big six" messages painted above is terribly controversial. But the research also points to a "small four" list of messages, outlined below, about variables that have much less influence than we might intuitively think.

1. **Teacher demographics.** The research suggests that it's not who teachers are or how they were trained that matters. It is how they think.

2. **Student attributes.** By and large (except for prior knowledge, student self-efficacy, and their range of learning strategies), many of the attributes of individual students that have been the focus of research, such as gender and personality, account for very little of the variation in student achievement.

3. **Technology.** Technology is a large part of what happens in modern schooling and it is something that we need to understand so that we can maximize the benefits that may exist in technology. But it is important to acknowledge that existing research shows that for most areas of computer-assisted/based instruction, the impact on learning has been positive but small. Our forthcoming paper "Not All That Glitters Is Gold" will explore this dilemma more fully.

4. **Structure of schools and classrooms.** The structure of schools and classes (i.e., class size, ability grouping, multiage grade classes, etc.) makes very little difference to the amount of progress students make in their learning. Similarly, the effects of the types of schools that students attend make little difference: single-sex schools, religious schools, charter schools, length of school day, size of school, and so on all have small effects on student achievement.

5. Conclusion

The Visible Learning research represents (probably) the largest meta-meta-analysis of what works best in improving student learning outcomes. But the reaction to the research and its interpretation has been deeply polarizing. Much of the criticism has centered on the methodology rather than on the interpretation of the story underlying the research findings, with the argument that the enterprise has become lost in a world of tainted data and abstractions that are too removed from the everyday realities of educators. The critics are saying that Visible Learning is *fool's gold*.

In this paper, we have argued that this is far from the case. We started by sketching the emergence of the education research paradigm, culminating in the meta-analysis approach. We then explored the criticisms of meta-analysis and its application in the Visible Learning research, concluding that, like a kind of educational TripAdvisor, meta-analyses help us to identify what (has) worked best, although not necessarily whether it will work again; that's where the interpretation comes in.

Finally, we recapped on the less controversial element of Visible Learning, the six big messages, and the need to focus on the major themes: educators that DIIE bring learning alive. This means that they:

- participate in *Diagnosing* the status of students as they begin lessons,
- have multiple *I*nterventions that they can apply if their current intervention is not having the desired impact,
- attend to the dosage and fidelity of their *I*mplementation on the learning lives of students, and
- *E*valuate the students' responses to their interventions.

At its core, the Visible Learning message is about teachers seeing learning through the eyes of their students, so that those students become their own teachers. To our eyes, that is truly golden.

References

American Medical Association. (2017). *CPT (Current Procedural Terminology)*. Chicago, IL: Author.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Bailar, J. C. (1997). The promise and problems of meta-analysis. *New England Journal of Medicine*, *337*, 559–561.

Barber, M., Moffit, A., & Kihn, P. (2011). *Deliverology 101: A field guide for educational leaders*. Thousand Oaks, CA: Corwin.

Bergeron, P.-J. (2017). How to engage in pseudoscience with real data: A criticism of John Hattie's arguments in Visible Learning from the perspective of a statistician. *McGill Journal of Education*, *52*(1), 237–246.

Bernard, R. M., Borokhovski, E., Tamim, R., & Abrami, P. C. (2013). *Teacher-centered and student-centered pedagogy: A meta-analysis of classroom practices and processes*. Poster presented at the meeting of the American Educational Research Association, San Francisco, CA.

Bland, R., & Erasmus, D. (1814). *Proverbs, chiefly taken from the Adagia of Erasmus, with explanations; and further illustrated by corresponding examples from the Spanish, Italian, French & English languages*. London, England: T. Egerton.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: John Wiley & Sons.

Coe, R. (2002). *It's the effect size, stupid. What effect size is and why it is important*. Paper presented at the British Educational Research Association Annual Conference, Exeter, England. Retrieved from http://www.cem.org/attachments/ebe/ESguide.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research*, *77*(1), 113–143.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, *33*(5), 517.

Federal Aviation Administration. (2016). *Airplane flying handbook*. Oklahoma City, OK: U.S. Department of Transportation, Federal Aviation Administration, Airman Testing Standards Branch.

Feinstein, A. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology*, *48*(1), 71–79.

Glass, G. V. (Ed.). (1976). *Evaluation studies review annual* (Vol. 1). Beverly Hills, CA: Sage.

Hansford, B. C., & Hattie, J. A. (1982). The relationship between self and achievement/performance measures. *Review of Educational Research*, *52*(1), 123–142.

Hattie, J. (2012). *Visible learning for teachers*. New York, NY: Routledge.

Hattie, J., & Hamilton, A. (2020). "As good as gold? Why we focus on the wrong drivers in education." Thousand Oaks, CA: Corwin.

Hattie, J., & Zierer, K. (2018). *10 mindframes for visible learning: Teaching for success.* New York, NY: Routledge.

Hattie, J. A., & Donoghue, G. M. (2016). Learning strategies: A synthesis and conceptual model. *npj Science of Learning*, *1*, 16013.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Oxford, England: Routledge.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research*, *86*(3), 643–680.

La Paro, K. M., & Pianta, R. C. (2000). Predicting children's competence in the early school years: A meta-analytic review. *Review of Educational Research*, *70*(4), 443–484.

Mansell, W. (2008, November 21). Research reveals teaching's holy grail. *Times Educational Supplement*. Retrieved from https://www.tes.com/news/research-reveals-teachings-holy-grail

Marzano, R. J. (2003). *What works in schools: Translating research into action*. Alexandria, VA: Association for Supervision and Curriculum Development.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361–365.

Moallem, I. (2013). *A meta-analysis of school belonging and academic success and persistence Loyola University Chicago Dissertations*, 726. Retrieved from http://ecommons.luc.edu/luc_diss/726

Nurmi, J. E. (2012). Students' characteristics and teacher–child relationships in instruction: A meta-analysis. *Educational Research Review*, *7*(3), 177–197.

Preston, J. A. (2007). *Student-centered versus teacher-centered mathematics instruction: A meta-analysis* (unpublished Ph.D. thesis). Indiana University of Pennsylvania, Indiana, PA.

Rosenberg, M. S. (2005). The file-drawer problem revisited. *Evolution*, *59*(2), 464–468.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641.

Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised effect sizes. *Journal of Education Policy*, *32*(4), 450–466.

Snook, I., O'Neill, J., Clark, J., O'Neill, A. M., & Openshaw, R. (2009). Invisible learnings? A commentary on John Hattie's book: *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement. New Zealand Journal of Educational Studies*, *44*(1), 93–106.

Sullivan, G. (2011). Getting off the "gold standard": Randomized controlled trials and education research. *Journal of Graduate Medical Education*, *3*(3), 285–289.

Terhart, E. (2011). Has John Hattie really found the holy grail of research on teaching? An extended review of Visible Learning. *Journal of Curriculum Studies*, 43(3), 425–438.

Thomas, T., Sparkes, C., Alexander, K., Jackson, R., Silva, C., Walker, T., Mandel, E., Abrami, P., & Bernard, R. (2012). *Blurring boundaries between teacher and learner: The case for a constructivist approach to teaching with technology in post-secondary classrooms*. Poster presented at the meeting of the Society for Teaching and Learning in Higher Education, Montreal, QC.

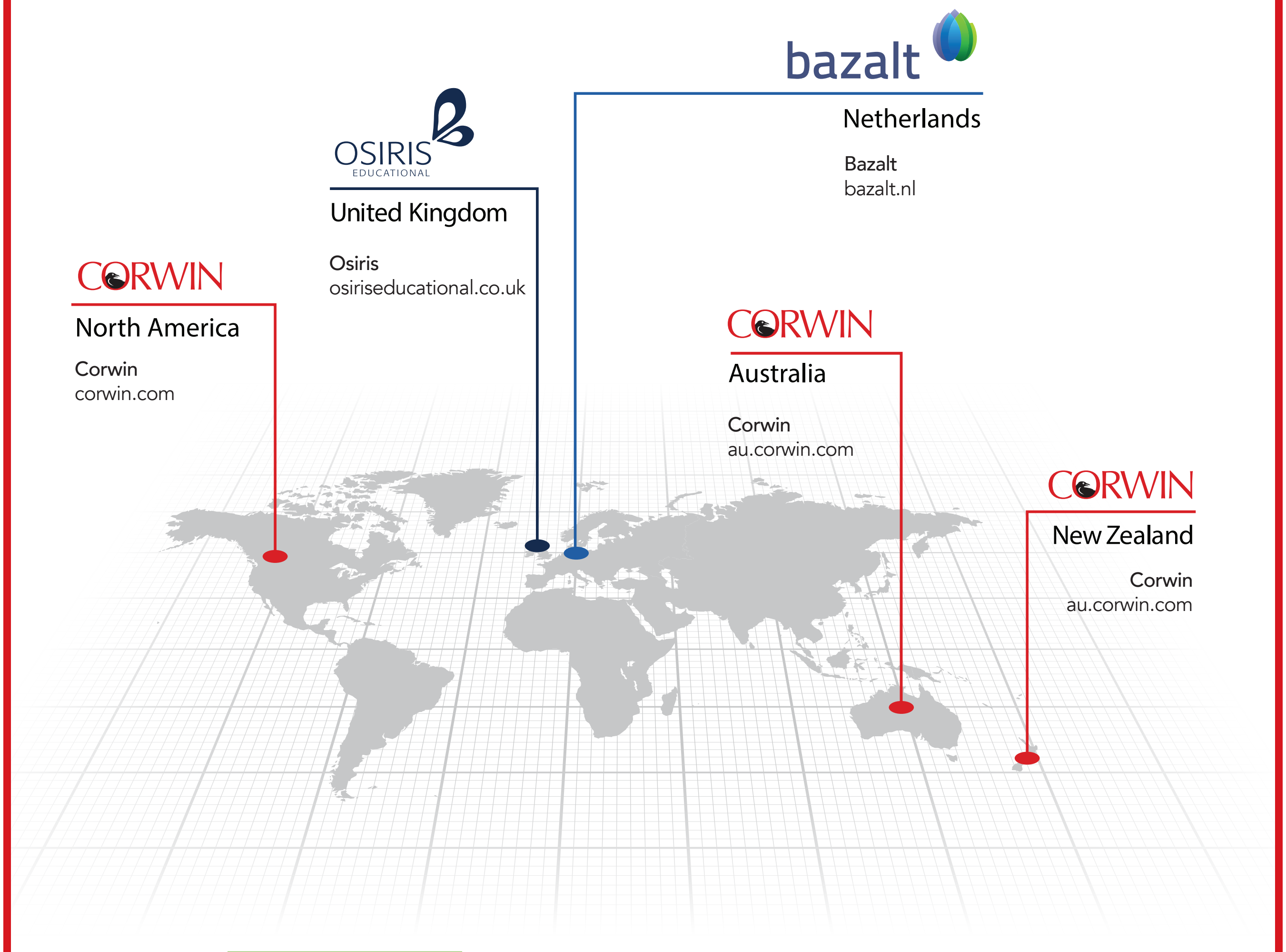
Topphol, A. K. (2012). Kan vi stole på statistikkbruken i utdanningsforskinga? [Can we rely on the use of statistics in education research?]. *Norsk Pedagogisk Tidsskrift*, *95*(6), 460–471.

Wittwer, J., & Renkl, A. (2010). How effective are instructional explanations in example-based learning? A meta-analytic review. *Educational Psychology Review*, *22*(4), 393–409.

CORWIN

To learn more

and get involved in the Visible Learning$^{plus®}$ Global Network



bazalt

Netherlands

Bazalt
bazalt.nl

OSIRIS EDUCATIONAL

United Kingdom

Osiris
osiriseducational.co.uk

CORWIN

North America

Corwin
corwin.com

CORWIN

Australia

Corwin
au.corwin.com

CORWIN

New Zealand

Corwin
au.corwin.com

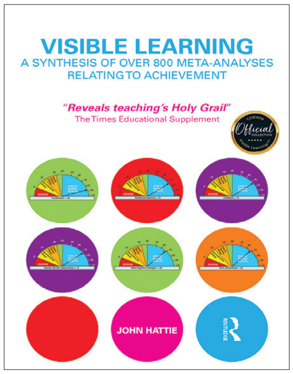
CORWIN Visible Learning$^{plus®}$

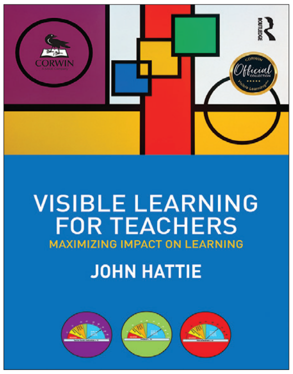
Contact Us: **www.visiblelearningplus.com** | **www.corwin.com**
Email: **visiblelearning@corwin.com** | Twitter: **@VisibleLearning**

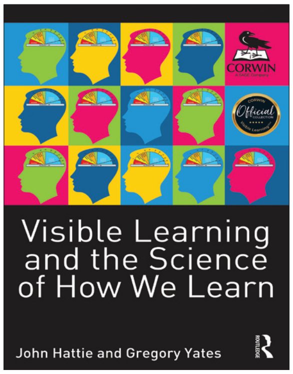
CORWIN Visible Learning$^{plus®}$

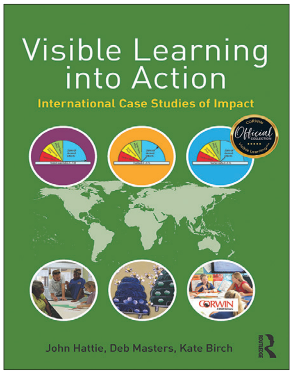
Build your Visible Learning™ library!

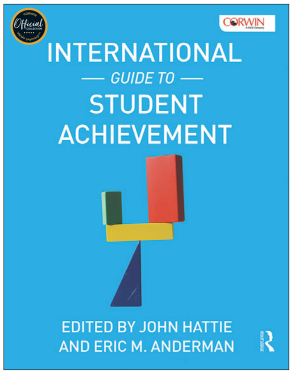
Foundation Series


VISIBLE LEARNING


VISIBLE LEARNING FOR TEACHERS

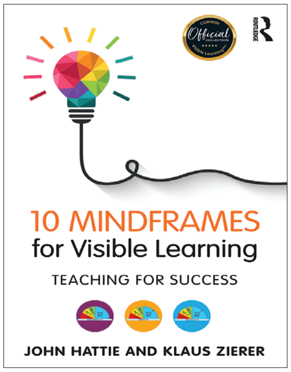

VISIBLE LEARNING AND THE SCIENCE OF HOW WE LEARN


VISIBLE LEARNING INTO ACTION

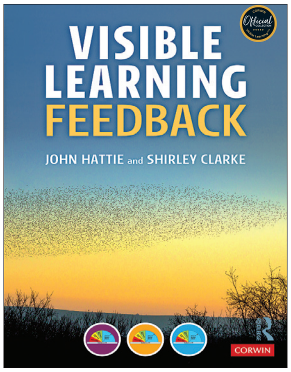

INTERNATIONAL GUIDE TO STUDENT ACHIEVEMENT

Visit corwin.com/vlbooks

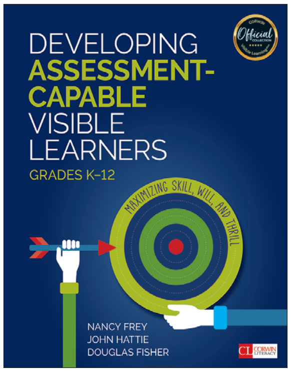
Impact Series



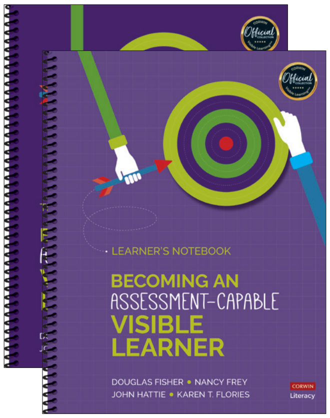
10 MINDFRAMES FOR VISIBLE LEARNING



VISIBLE LEARNING FEEDBACK

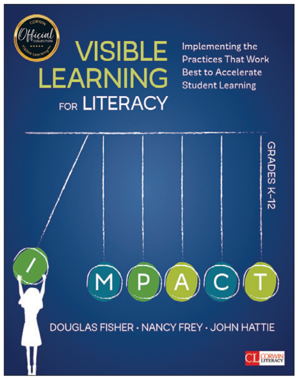

DEVELOPING ASSESSMENT-CAPABLE VISIBLE LEARNERS
GRADES K–12
MAXIMIZING SKILL, WILL, AND THRILL
NANCY FREY
JOHN HATTIE
DOUGLAS FISHER

DEVELOPING ASSESSMENT-CAPABLE VISIBLE LEARNERS, Grades K–12

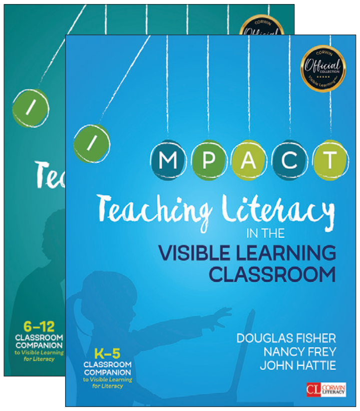

LEARNER'S NOTEBOOK
BECOMING AN ASSESSMENT-CAPABLE VISIBLE LEARNER
DOUGLAS FISHER • NANCY FREY
JOHN HATTIE • KAREN T. FLORIES
CORWIN Literacy

BECOMING AN ASSESSMENT-CAPABLE VISIBLE LEARNER Teacher's Guide & Learner's Notebooks Grades K–2, 3–5, & 6–12

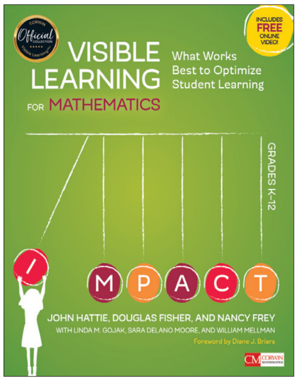
Practice Series


VISIBLE LEARNING FOR LITERACY
Implementing the Practices That Work Best to Accelerate Student Learning
GRADES K-12
MPACT
DOUGLAS FISHER • NANCY FREY • JOHN HATTIE

VISIBLE LEARNING FOR LITERACY, Grades K–12


MPACT
Teaching Literacy IN THE VISIBLE LEARNING CLASSROOM
6–12 CLASSROOM COMPANION
K–5 CLASSROOM COMPANION to Visible Learning for Literacy
DOUGLAS FISHER
NANCY FREY
JOHN HATTIE

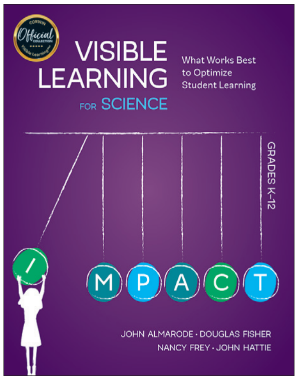
TEACHING LITERACY IN THE VISIBLE LEARNING CLASSROOM, Grades K–5 & 6–12


INCLUDES FREE ONLINE VIDEO!
VISIBLE LEARNING FOR MATHEMATICS
What Works Best to Optimize Student Learning
GRADES K-12
MPACT
JOHN HATTIE, DOUGLAS FISHER, AND NANCY FREY
WITH LINDA M. GOJAK, SARA DELANO MOORE, AND WILLIAM MELLMAN
Foreword by Diane J. Briars

VISIBLE LEARNING FOR MATHEMATICS, Grades K–12

Grades K–2 Classroom Companion to Visible Learning for Mathematics
MPACT
Teaching Mathematics IN THE VISIBLE LEARNING CLASSROOM
JOHN ALMARODE
DOUGLAS FISHER
KATERI THUNDER
JOHN HATTIE
NANCY FREY
CORWIN Mathematics

TEACHING MATHEMATICS IN THE VISIBLE LEARNING CLASSROOM, Grades K–2, 3–5, 6–8, & High School


VISIBLE LEARNING FOR SCIENCE
What Works Best to Optimize Student Learning
GRADES K-12
MPACT
JOHN ALMARODE • DOUGLAS FISHER
NANCY FREY • JOHN HATTIE

VISIBLE LEARNING FOR SCIENCE, Grades K–12

CORWIN
A SAGE Publishing Company

Helping educators make the greatest impact

CORWIN HAS ONE MISSION: to enhance education through intentional professional learning.

We build long-term relationships with our authors, educators, clients, and associations who partner with us to develop and continuously improve the best evidence-based practices that establish and support lifelong learning.

Cognition
Education Group

Ignite the global passion for learning

COGNITION EDUCATION GROUP is a leading provider of education consultancy, professional learning, teacher recruitment, early years and primary tutoring, e-learning and publishing services. Headquartered in New Zealand and operating around the world, our focus is to build the capability and expertise of educators and leaders to improve educational outcomes for all.

Common VISIBLE LEARNING™ Methodology Critiques and Responses

This document provides a succinct list of the common criticisms of the Visible Learning methodology, along with responses from John Hattie.

A. Issues of Meta-Analysis

1. *Critique:* **Weighting**. Some have argued that the effect sizes from each meta-analysis should be weighted by sample size.

   Response: This is reasonable; however, we have back-tested this for some domains and it made little difference to the aggregated effect sizes in these cases.

2. *Critique:* **Sample size**. Many of the controversial influences only have one to three meta-analyses.

   Response: The key is not necessarily the number of meta-analyses, but a combination of factors such as the number of studies in each meta-analysis, the number of effects, the sample size, and the quality of the meta-analysis. In Visible Learning MetaX, we include all of these features and researchers can make their own judgments and analyses and see the various confidence features.

3. *Critique:* **Sampling**. Visible Learning uses meta-analyses from atypical student populations (e.g., English language learners or individuals with attention-deficit/hyperactivity disorder, hyperactivity, or emotional/behavioral issues). Visible Learning also includes atypical subjects from nonstudent populations, such as doctors, tradesmen, nurses, athletes, sports teams, and military groups.

   Response: It is true that most meta-analyses include diverse samples and often within the meta-analysis they evaluate these factors. In Visible Learning MetaX, the nature of the sample is identified (preschool, elementary, high school, tertiary, across all K–12, special education or not, etc.). While there may be nonschool people in the meta-analysis, indeed there are relatively few because the major discriminator in the selection process was the presence of a school-based sample.

4. *Critique:* **Reductionism**. One number cannot summarize a research field; a common criticism of meta-analysis is that the analysis focuses on the summary effect and ignores the fact that the treatment effect may vary from study to study.

   Response: The goal of a meta-analysis should be to synthesize the effect sizes and not simply (or necessarily) report a summary effect. It is commonplace to investigate whether the overall mean is a sufficient statistic to explain the findings. If not, moderator analyses are the norm, and herein often lies the most interesting aspects of meta-analyses. In many ways, the analyses of their heterogeneity are among the most fascinating parts of synthesizing studies. In the same way, it took me 15+ years to work through this heterogeneity across the many meta-analyses.

5. *Critique:* **Quality and aggregation**. Visible Learning aggregates the findings of poor studies, thus setting low standards of judgment for the quality of outcome study.

Response: There is a robust discussion on how quality in meta-analysis should be included. In *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*, I said in 2009 that I would not attend to quality issues in that book *because* quality had been addressed in many places elsewhere (see references in the book), but some mischievously claimed he did not care about quality. First, most meta-analyses investigate the moderator effects of quality and exclude low ones if this is an issue. Second, the concern here is the quality of the meta-analyses and this is a less researched topic. In Visible Learning MetaX, we provide the journal and its impact factor and this then can be explored (although we accept that this is not the optimal quality measure and others are welcome). Third, it is an empirical question whether the quality of the study is a moderator. As stated in *Visible Learning*, Lipsey and Wilson (1993), for example, summarized 302 meta-analyses in psychology and education and found no differences between studies that only included random versus nonrandom design studies (*d* = 0.46 vs. *d* = 0.41), or between high-quality (*d* = 0.40) and low-quality (*d* = 0.37) studies. There was a bias upward from the published studies (*d* = 0.53) compared to nonpublished studies (*d* = 0.39), although sample size was unrelated to effect size (*d* = –0.03). Further, Sipe and Curlette (1996) found no relationship between the overall effect size of 97 meta-analyses (*d* = 0.34) and sample size, number of variables coded, and type of research design, and they found a slight increase for published (*d* = 0.46) versus unpublished (*d* = 0.36) meta-analyses. There is one exception that can be predicted from the principles of statistical power; and where the effect sizes are close to zero, then the probability of having high confidence in this effect is probably related to the sample size and quality of the study (see Cohen, 1988, 1990). The aim should be to summarize all possible studies regardless of their design and then ascertain if quality is a moderator to the final conclusions.

6. *Critique:* **The file drawer problem.** The file drawer problem invalidates meta-analysis. While the meta-analysis will yield a mathematically sound synthesis of the studies included in the analysis, if these studies are a biased sample of all possible studies, then the mean effect reported by the meta-analysis will reflect this bias. Several lines of evidence show that studies finding relatively high treatment effects are more likely to be published than studies finding lower treatment effects. The latter unpublished research lies dormant in the researchers' filing cabinet—hence, the *file drawer* problem.

   Response: Publications bias is perennially a problem for all research. Many ask how many unpublished studies sitting in someone's file drawer would it take to overturn the substantive findings of the meta-analysis? The number of studies aggregated in Visible Learning is large (more than 90,000) and the benchmark/bar for an above average affect is *d* = 0.40. We believe that the range of studies and high bar mitigate the file drawer problem.

7. *Critique:* **Fruit salad.** Visible Learning mixes "apples and oranges" in combining the findings of studies with varying methodological quality.

   Response: Any literature review involves making balanced judgments about diverse studies. A major reason for the development of meta-analysis was to find a more systematic way to join studies, in a similar way that apples and oranges can make fruit salad. Meta-analysis can be considered to ask about "fruit" and then assess the implications of combining apples and oranges, and the appropriate weighting of this combination. Unlike traditional reviews, meta-analyses provide systematic methods to evaluate the quality of combinations, allow for evaluation of various moderators, and provide excellent data for others to replicate or recombine the results. The key in all cases is the quality of the interpretation of the combined analyses. Further, as noted above, the individual studies can be evaluated for methodological quality.

8. *Critique:* **Important studies are ignored.** The studies included in Visible Learning are cherry-picked and leave out some of the most important studies.

   Response: It is hoped that the important studies are included! If not, they can be added from traditional reviews (as in many of the Visible Learning sections on the various influences). Other meta-analyses are continually being added to Visible Learning MetaX and many of the meta-analyses are explicit about their criteria for finding and selecting studies.

9. *Critique:* **It changes.** Why does the Visible Learning database keep adding more meta-analyses and more influences?

   Response: Of course, more need to be added, because this is the nature of research—we continually question, query, replicate, and validate previous studies. Furthermore, Visible Learning is built on Popperian principles. We seek falsifiability—maybe the next meta-analyses will question the underlying Visible Learning model and we want to be the first to acknowledge this. So far, however, every meta-analysis added provides confirmation not disconfirmation. It is exciting that researchers are still finding fascinating influences to investigate and to add to the database. Visible Learning MetaX will allow researchers to see when new meta-analyses are added so that they can explore the implications themselves.

10. *Critique:* **The effect sizes in Visible Learning change over time.**

    Response: The claim is that the Visible Learning rankings and effect sizes are not consistent, and we do find the same result when a new meta-analysis on the same topic is added. It would be even more remarkable if the average effect for any influence stayed exactly the same when more studies are added! And indeed, most are very similar. Yes, some have changed, most often because the first meta-analyses may not have had sufficient studies to provide stability of the mean influence.

11. *Critique:* **Rear view of the world.** What can studies based on previous studies tell us about the future?

    Response: Yes, the studies are "historical"; that is, they report past findings and cannot show that the future must be the same. This is what re-search means. The aim is to learn from what has occurred to better inform where to go, in the same way that checking the rear-view mirror when driving helps us move forward safely. To ignore the past permits opinions, fads, beliefs, and desires to dominate; our mission, educating students, demands more and at minimum to not repeat past errors, to learn how to scale up success, and to optimize the highest probability interventions.

12. *Critique:* **A narrative review is better.**

    Response: Yes, of course it is. It is the interpretation that matters most, whether the interpretation is based on primary, secondary, or meta-analyses. A key element of the Visible Learning research is about providing that interpretation or the story behind the data.

B. Effect Sizes

13. *Critique:* **Mixing different effect sizes is bad science.**

    Response: There are many forms of effect sizes, and there are many books that outline these. In Visible Learning, there are two major forms: (1) comparison group and (2) growth over time. Either is based on many forms of statistics, mainly means and correlations. Care is needed when considering them, and the aim of Visible Learning is to provide such care. Across the 1,600 effect sizes in the current Visible Learning database, there are no mean differences relating to whether the effects are based on a correlation (0.40) or based on mean differences (0.40). More interestingly, the effects from influences classified as "Causal – an intervention" are much higher (0.50) than influences that are more correlates (0.30), causal with no intervention (0.28), and causal intervention based on context (0.28). This does not mean we ignore these differences; in the discussion of any influence, there can be important nuances in interpretation and the Visible Learning books discuss these. Too often, critics merely look at the tables, ignore the story, and wrongly conclude that we do not discuss these important *moderators.*

14. *Critique:* **Common language effect size (CLE) is wrong.**

    Response: There was an unfortunate error in early editions of *Visible Learning*, where the wrong column in a spreadsheet was used to populate the CLE in the appendix. CLE was introduced as an alternative (not substitute, as some have claimed) way of interpreting the effect sizes. Clearly it was not successful, as it took a few years for some Norwegian students to detect the error, which was immediately acknowledged and updated in subsequent reprints of *Visible Learning*. The error did *not* mean the effect sizes or any interpretations based on the effect sizes in *Visible Learning* were incorrect.

15. *Critique:* **Half of the statistics in Visible Learning are wrong.**

    Response: A Twitter troll of a comment supposedly attributed to me—that half the statistics in Visible Learning are wrong—has perpetuated this myth. The source (a conference in London) was videoed, so I went back and checked—I never made this comment, and it is wrong. All of the data are provided in the appendices of *Visible Learning* (2009) and *Visible Learning for Teachers* (2012) (and now in Visible Learning MetaX) and anyone can check them. Yes, the early versions of CLE were incorrect but incidental to the interpretations in Visible Learning. This is a myth.

    For more about this critique, read "Effective Debate: In Defence of John Hattie" by Stuart Lord (2015).

16. *Critique:* **Effect sizes are not used by mathematicians.**

    Response: The claim is "Mathematicians don't use it. Mathematics textbooks don't teach it. Statistical packages don't calculate it." First, statisticians use it, developed it, teach it, and calculate it. Effect size is referenced in most basic statistics books, has been subjected to numerous studies, is used in all meta-analyses, is highly recommended by the American Psychological Association (2009), and is hotly debated in many sources. Effect sizes do exist. There are many forms of effect sizes, there are many statistical treatises debating effect sizes, and just because some mathematicians do not use them does not make effect sizes not real (see Hedges & Olkin, 1985, for an early and excellent analysis of effect sizes; and more recently, see Coe, 2002, for an overview).

17. *Critique:* **Effect sizes should not be used in education, as they came out of medicine.**

    Response: Interestingly, the reverse is true. Effect size started life in education research and was later adopted by medicine.

18. *Critique:* **The variability of effects is ignored.**

    Response: This is incorrect. For every influence in Visible Learning, there is an estimate of the variance of the mean of the effects (see each dial). Further, there is an established methodology about whether the variance of the effects is so heterogeneous that the average may not be a good estimator. Conducting such tests is basic practice in meta-analyses, and readers were encouraged to go to the original studies to see these analyses. An estimator of the variance was included within each influence (see the dial for each influence) and appropriately commented when these were large. Much time has been spent studying many of the influences with large variance (e.g., feedback) and the story is indeed more nuanced than reflected in the average effect. Feedback is among the highest but also most variable effects; although much feedback is positive, much is also negative. For example, it is critical to distinguish between giving and receiving feedback, between "how am I going" and "where to next" feedback, and how students and teachers receive and interpret feedback (Hattie & Clarke, 2019; Hattie, Gan, & Brooks, 2017; Hattie & Timperley, 2007).

C. Interpretation of Effect Sizes

19. *Critique:* **The use of the hinge point of *d* = 0.40 seems arbitrary.**

    Response: The hinge point is the average of all 1,600+ meta-analyses, and one interest in Visible Learning was those influences that surpassed the average, particularly in comparison with those below the average. It is fascinating that this overall average of *d* = 0.40 has not changed since John's first publication in 1989. The *d* = 0.40 is merely an overall summary of many influences, across many situations, ages, content, and so forth, and serves to organize influences in Visible Learning.

20. *Critique:* **The hinge point does not control for moderators and mediators.**

    Response: The claim is that U.S. Department of Education benchmark effect sizes per year level indicate another layer of complexity in interpreting effect sizes; studies need to control for the age of students as well as the time over which the study runs, and the claim is that Visible Learning does not do this. The reality is that it did, does, and always will. The hinge point is an average; we should never be seduced by the flaw of the average, and moderators and mediators to any average are key concerns in all educational studies. In many places in Visible Learning and elsewhere, the moderators and mediators are a continual source of fascination and debate.

21. *Critique:* **The average effect size can be moderated by age.**

    Response: Given that the hinge point is the overall average, it should always be evaluated relative to all moderators, including age. Care is needed, however, to then not make claims like "effect sizes are moderated by age" and provide tables of these age effects without attending to the nature of the assessment. So often these tables come from narrow measures of achievement in core subjects such as reading and numeracy. Such measures are often brief (40- to 120-minute tests), a "mile wide and an inch deep," and do not reflect the richness of reading and numeracy. In these cases, it is not surprising that there would be greater gains in the younger years of schooling and lower gains in the upper years of schooling. This should not be confused with assessments of reading and numeracy in the upper years based on what is being taught, where you can get various effect sizes (including, contrary to critic claims, >0.40).

22. *Critique:* **There are so few moderators discovered in Visible Learning that we can use the average.**

    Response: It is the case that there are few moderators to the average effect sizes. This is not inconsistent with an earlier critique (then termed "aptitude-treatment interactions") by Cronbach and Snow (1977), but this does not mean we should not eternally seek them. Moderators are the essence of the model of allowing for individual differences, for differentiation, and for centering on the child. When we implement interventions, we need to continually ask about who is affected, how, why, when, and by what magnitude. From such investigations, we may or may not find generalized moderators and this is an outcome finding; yet in education, we rarely start with the premise that one intervention fits all. Where there were moderators in *Visible Learning* (2009), these were noted. For example, the average effect for homework was 0.29, but the effect size was low for primary school (0.15) and much higher for high school (0.64). This is noted, an interpretation is provided, and this case shows the low interpretative power of the average (0.29).

23. *Critique:* **It is wrong to focus on influences with high effects sizes and leave out the low ones.**

    Response: Absolutely. Some of the low effects may be critical. At least, it is critical to ask why they are so low, and one of John's interests is exploring some of these (in particular, subject matter knowledge, modern learning or open environments, class size, and retention). I have asked why the effects are so low for class size, especially when it should be expected that reduced size should allow more opportunity for introducing some of the higher effects (Hattie, 2010). We are exploring the conditions in which deeper subject matter knowledge does matter, and we are involved in a major project about the optimal collaborative teaching to realize amazing influences on students in open environments. Just because an effect is not >0.40 does not mean it is not worthwhile; it means it may need deeper exploration to make the impact higher.

24. *Critique:* **Correlation does not imply causation.**

    Response: Yes, this is basic, although contested by some, and structural models as measures over time can move more to claims about causation. The role of a researcher is to interpret with care and to not slip and make or infer causality. Visible Learning aimed to build a model that involves causation, used evidence from the many meta-analyses to build and defend this model, and made strong statements that any such model is subject to falsification. This is the fine line for all interpretations and causal claims are legitimate if they are backed with evidence.

25. *Critique:* **The noninclusion on qualitative studies.** The following is an example of this criticism:

    > Let me state the basic shortcoming more bluntly. The non-meta-analytic and qualitative or mixed methods studies Professor Hattie has excluded are precisely the research investigations that do make visible not only (a) that class size matters to student achievement, but also (b) what the observed effects of different class sizes are on classroom teaching and learning practices as a whole, and furthermore (c) which sub-groups of students are most materially affected by larger or smaller class sizes and the attendant changes in classroom processes they require.

    Response: Yes, qualitative studies are not included in meta-analyses, and yes, they can add richness to the workings of classes. One of the most exciting developments since Visible Learning was published is the emergence and growth of meta-synthesis of qualitative studies (see Kennedy, 2008; Suri, 2013) and we look forward to reading a similar synthesis of these studies to the Visible Learning work. I have also used many of these nonempirical studies in trying to understand the effects of many of the influences—and class size (Hattie, 2005); in this case, these studies helped explain why the effects of class size are so low!

26. *Critique:* **Meta-analyses are not sensitive to instruction.**

    Response: The typical claim is that meta-analyses fail to consider the fact that different outcome measures are not equally sensitive to instruction (Popham, 2007). This is not the case in all meta-analyses and is certainly a major issue when interpreting the implications of meta-analyses. Controlling for sensitivity to instruction is more appropriate in meta-analyses and indeed, we would be massively advantaged if this moderator was included more often. A good five-level classification has been provided by Ruiz-Primo, Shavelson, Hamilton, and Klein (2002) for the distance of an assessment from the enactment of curriculum, with examples of each:

    1. Immediate, such as science journals, notebooks, and classroom tests;
    2. Close, or formal embedded assessments (for example, if an immediate assessment asked about number of pendulum swings in 15 seconds, a close assessment would ask about the time taken for 10 swings);
    3. Proximal, including a different assessment of the same concept, requiring some transfer (for example, if an immediate assessment asked students to construct boats out of paper cups, the proximal assessment would ask for an explanation of what makes bottles float or sink);
    4. Distal, for example a large-scale assessment from a state assessment framework, in which the assessment task was sampled from a different domain, such as physical science, and where the problem, procedures, materials, and measurement methods differed from those used in the original activities; and
    5. Remote, such as standardized national achievement tests.

27. *Critique:* **Meta-analyses do not control for costs.**

    Response: Yes, few do (but see Yeh, 2008), but the costs of implementation can be included as the Education Endowment Foundation has done. Of course, the costs need to be considered when making decisions about what interventions to use. For example, would you invest in the huge and recurrent costs of reducing class size compared to implementing the lower-cost and scalable solutions such as direct instruction, reciprocal teaching, or formative assessment? All are excellent empirical questions and are an area we are attempting to address in Visible Learning MetaX.

28. *Critique:* **Others are now recanting their own use of effect sizes.**

    Response: Yes, some have, although their interpretations seem to remain the same. The most famous case is Dylan Wiliam, who claimed:

    > In retrospect, therefore, it may well have been a mistake to use effect sizes in our booklet "Inside the black box" to indicate the sorts of impact that formative assessment might have. It would have been better to talk about extra months of learning which considers the fact that the annual growth in achievement, when measured in standard deviations, declines rapidly in primary school (one year's growth is over 1 standard deviation

> for five-year-olds, and only around 0.4 for 11-year-olds). That said, in answer to Michael Dorian's question, in arriving at our subjective estimate of 0.4 to 0.7 standard deviations for the impact of formative assessment, we did rely more on studies that were classroom-based, over extended periods of time, and which used standardized measures of achievement.
>
> I do still think that effect sizes are useful (and are far more useful than just reporting levels of statistical significance). If the effect sizes are based on experiments of similar duration, on similar populations, using outcome measures that are similar in their sensitivity to the effects of teaching, then I think comparisons are reasonable. Otherwise, I think effect sizes are extremely difficult to interpret. (Didau, 2014)

The reality is that the use of effect sizes has grown significantly over the past three decades. The perceived value of using effect sizes is now so strong that many professional bodies, journal editors, and statisticians across various disciplines have mandated their inclusion as necessary in order to clarify and substantiate differences in research findings (for example, American Psychological Association, 2001, 2009; Baugh & Thompson, 2001; Kline, 2004).

29. *Critique:* **Visible Learning ignores debates about what is worth learning.** The criticism is as follows:

> Only in one sentence is a look at the material side thrown in: "Educating is more than teaching people to think—it is also teaching people things that are worth learning" (p. 27). That might have been the starting point of a discourse about the substance of education and teaching, but Hattie does not follow this possible line of thought. So, one wonders: Where is the beef? In the chapter on "curriculum" one would expect more information about the substance, the content of school learning. However, again nothing can be found there. The chapter is divided into specialized areas: reading, mathematics, and other curricular elements. Under these headings, the reader again finds reports on certain specialized teaching methods and their effect sizes. The question of content, the question of the pedagogical significance of subjects, reflections about problems, and possibilities of legitimizing curricular decisions (Why include this—why exclude that?) are completely ignored. (Terhart, 2011)

Response: Visible Learning is not about the aims of education and is not a treatise of what is worth learning. I have written on these topics elsewhere, and they are indeed critical topics.

30. *Critique:* **Visible Learning is only about achievement and this is not all that school is about.**

Response: Visible Learning (2009) starts by saying, "Of course, there are many outcomes of schooling, such as attitudes, physical outcomes, belongingness, respect, citizenship, and the love of learning. This book focuses on student achievement, and that is a limitation of this review" (p. 6). Others are now synthesizing effects relating to motivation, interest, and affect; we have recently synthesized "how we learn" (Hattie & Donoghue, 2016). We wish others would synthesize health and physical outcomes. For example, Mitchell (2014) has focused on special needs students, and we are delighted when this more rounded view of the many outcomes of schooling is reviewed. Achievement remains central to the outcomes of schooling.

31. *Critique:* **Visible Learning ignores socioeconomic effects.**

    Response: Not at all. The Visible Learning dataset explicitly captures meta-analysis on socioeconomic status (SES) and other out-of-school influences. It also acknowledges that SES has an above average impact on student learning outcomes. However, a key message of Visible Learning is that teachers make a difference and high-impact strategies will improve learning outcomes, irrespective of a student's background or starting point.

32. *Critique:* **There are many risks involved when interpreting meta-analyses.**

    Response: Damn right there are. There is a whole compendium of risks to interpretation to statistics and these are not unique to meta-analyses. For example, Andrade and Cizek (2010) write,

    > Black and Wiliam [1998] noted [that] effect size is influenced by the range of achievement in the population. An increase of 5 points on a test where the population standard deviation is 10 points would result in an effect size of 0.5 standard deviations. However, the same intervention when administered only to the upper half of the same population, provided that it was equally effective for all students, would result in an effect size of over 0.8 standard deviations, due to the reduced variance of the subsample. An often-observed finding in the literature—that formative assessment interventions are more successful for students with special educational needs (for example in Fuchs & Fuchs, 1986)—is difficult to interpret without some attempt to control for the restriction of range, and may simply be a statistical artefact. (p. 20)

However, this problem with restriction of range can occur in primary, secondary, and meta-analyses. Campbell and Stanley (1963) highlight many other possible threats to the validity of interpretation of statistics, no matter whether primary, secondary, or meta-analysis is used. Care is always the by-line and we have tried to be careful when making interpretations.

D. The Visible Learning™ Model

33. *Critique:* **There are alternative interpretations based on the Visible Learning data.**

    Response: Absolutely there are, but so far no one has deduced an alternative explanation. We challenge you to do so, as that is how science progresses. Refute the Visible Learning theory and build a new one, please. The data are all available, Visible Learning MetaX is a gold mine, and we will be the first to acknowledge explanations that explore more, are bold and subject to refutation, and help schools have the desired impact on their students. So far, no one has provided an alternative theory based on these data!

34. *Critique:* **It is but a model.**

    Response: Yes, and like any good model, it not only aims to explain what we know (the evidence) but also to project what we may seek to now know. It is speculative.

35. *Critique:* **The main messages of Visible Learning *defy* widespread teacher experience.**

    Response: Sometimes yes, sometimes no; it depends on the mindframes of the teacher. Sometimes evidence, indeed, defies "common sense," sometimes it confirms and provides permission to continue, and sometimes it surprises and needs triangulation. Experience is also interpreted and it is this interpretation that always needs questioning, refuting, and evaluating.

36. *Critique:* **Visible Learning reports the opposite conclusion to that of the actual authors of the studies it reports on (e.g., "class size," "teacher training," "diet," and "reducing disruptive behavior").**

    Response: Visible Learning is not about repeating, summarizing, and copying—it is about interpreting. One of the major claims in Visible Learning is that almost everything works, and this has led many researchers to find positive evidence and then make claims about importance. But importance is a relative concept. A good example is class size; the preponderance of evidence does show that reducing class size has a positive impact on student achievement, but the size of this positive effect is relatively small (Hattie, 2005, 2016). Many authors on class size have made claims of importance without considering relative strength. Another example is a recent meta-analysis on teacher performance pay with an overall effect size of 0.04 (Pham, Nguyen, & Springer, 2017).

37. *Critique:* **The influences are not separate.**

    Response: Absolutely right, and this is repeatedly emphasized in *Visible Learning* (2009) and *Visible Learning for Teachers* (2012). Hence, the Visible Learning model helps explain the overlaps, the interactions, and the meaning underlying the various influences. It is not possible to merely add the effect size from two influences together—and this overlap of the many influences is why it took me 15 years to write the first book.

38. *Critique:* **Narrowness for breadth of outcome (i.e., Visible Learning just focuses on student assessment outcomes).**

    Response: Agreed. This does not, however, mean Visible Learning is worthless, but it does raise issues with the narrow excellence that many claim about the purpose of schooling. We agree (as noted above) that there are other important aims and outcomes of schooling, but achievement is still one major one. A related concern is that the narrower the outcome, then the probability of a higher effect size compared to a wider conception of outcome (e.g., it is easier to get a higher effect size for vocabulary than comprehension). This is important to consider when building theories and explanations and when interpreting both meta-analyses and effects in classrooms.

References

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Andrade, H., & Cizek, G. J. (2010). *Handbook of formative assessment*. New York, NY: Routledge.

Baugh, F., & Thompson, B. (2001). Using effect sizes in social science research: New APA and journal mandates for improved methodology practices. *Journal of Research in Education*, *11*(1), 120–129.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139–148.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally & Company.

Coe, R. (2002). *It's the effect size, stupid. What effect size is and why it is important*. Paper presented at the British Educational Research Association Annual Conference, Exeter, England. Retrieved from http://www.cem.org/attachments/ebe/ESguide.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York, NY: Irvington Publishers.

Didau, D. (2014). Old Hat(tie)? Some things you ought to know about effect sizes. *The Learning Spy*. Retrieved from https://learningspy.co.uk/myths/things-know-effect-sizes/

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, *53*(3), 199–208.

Hattie, J. (2005). The paradox of reducing class size and improving learning outcomes. *International Journal of Educational Research*, *43*(6), 387–425.

Hattie, J. (2012). *Visible Learning for teachers: Maximizing impact on learning*. New York, NY: Routledge.

Hattie, J., & Clarke, S. (2019). *Visible Learning: Feedback*. New York, NY: Routledge.

Hattie, J., Gan, M., & Brooks, C. (2017). Instruction based on feedback. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 290–324). New York, NY: Taylor & Francis.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Hattie, J. A., & Donoghue, G. M. (2016). Learning strategies: A synthesis and conceptual model. *npj Science of Learning*, *1*, 16013.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Oxford, England: Routledge.

Hattie, J. A. C. (2010). On being a 'critic and conscience of society': The role of the education academic in public debates. *New Zealand Journal of Educational Studies, 45(1),* 85–96.

Hattie, J. A. C. (2016). The right question in the debates about class size: Why is the (positive) effect so small? In P. Blatchford, K. W. Chan, M. Galton, K. C. Lai, & J. C. L. Lee (Eds.), *Class size: Eastern and Western perspectives* (pp. 105–118). London, England: Routledge.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Kennedy, M. M. (2008). Contributions of qualitative research to research on teacher qualifications. *Educational Evaluation and Policy Analysis*, *30*(4), 344–367.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*(12), 1181.

Lord, S. (2015, March 22). Effective debate: In defence of John Hattie. *The Learning Intention.* Retrieved from https://thelearningintention.net/effective-debate-in-defense-of-john-hattie/

Mitchell, D. (2014). *What really works in special and inclusive education: Using evidence-based teaching strategies* (2nd ed.). Abingdon, England: Routledge.

Pham, L. D., Nguyen, T. D., & Springer, M. G. (2017, April 3). *Teacher merit pay and student test scores: A meta-analysis*. Nashville, TN: Vanderbilt University.

Popham, W. J. (2007). *Classroom assessment: What teachers need to know* (5th ed.). Boston, MA: Allyn & Bacon.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*(5), 369–393.

Sipe, T. A., & Curlette, W. L. (1996). A meta-synthesis of factors related to educational achievement: A methodological approach to summarizing and synthesizing meta-analyses. *International Journal of Educational Research*, *25*(7), 583–698.

Suri, H. (2013). *Towards methodologically inclusive research syntheses: Expanding possibilities*. London, England: Routledge.

Terhart, E. (2011). Has John Hattie really found the holy grail of research on teaching? An extended review of Visible Learning. *Journal of Curriculum Studies*, *43*(3), 425–438.

Yeh, S. S. (2008). The cost-effectiveness of comprehensive school reform and rapid assessment. *Education Policy Analysis Archives*, *16*, 13.