As Good as Gold?

Why We Focus on the Wrong Drivers in Education

John Hattie
Arran Hamilton

A JOINT PUBLICATION

CORWIN

Cognition Education Group

As Good as Gold?

Why We Focus on the Wrong Drivers 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 978-1-5443-9529-6 (web PDF)

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 John Almarode, Peter DeWitt, Shaun Hawthorne, James Nottingham, Ainsley Rose, Ariel Curry, Ray Smith, and Julie Smith for their critique of early drafts of this paper.

Introduction

In the late 1890s, Seattle was a modest logging town nestled in the Pacific Northwest region of the United States. Many of its inhabitants were engaged in the backbreaking work of harvesting and shipping timber south to San Francisco. Like the rest of the United States, Seattle was reeling from the Panic of 1893. This was the greatest financial crisis of the era and, with U.S. gold reserves depleted, there was severe economic depression throughout the country. Unexpectedly, this economic downfall set the stage for Seattle to change its destiny.

On July 17, 1897, the *SS Portland* chugged into the Seattle harbor with nearly two tons of gold from the Alaskan Yukon. In a small city haunted by news of the country's dwindling gold coffers, it's no wonder the scene sparked a frenzy. Erastus Brainerd, the mustachioed editor of the *Seattle Post-Intelligencer*, saw an opportunity to change the city's fortunes and took it, proclaiming Seattle the "gateway to the Yukon and the *only* such portal." His brilliant campaign worked, bringing over 70,000 prospective gold miners ("stampeders") from all over the country. Seemingly overnight, Seattle became the base camp for the Yukon Gold Rush.

Once the stampeders arrived in Seattle, businesses were ready to equip them with camping supplies, guidebooks, maps, food, and even sled dogs to help them survive their journey and, theoretically, find gold. While a few miners did strike it rich in the Yukon, the fortunes of most were less positive: many never left Seattle, died on the voyage to the Yukon, or simply came home unsuccessful in their efforts. The hunt for gold was all too often a fool's errand.

The real winners in the Yukon Gold Rush were the businesses in Seattle, leveraging desperate stampeders' visions of gold into hard cash in their own pockets.

In this paper, we argue that in education we're seeing a similar "gold rush" and that it's been going on for a long time. Like modern-day Erastus Brainerds, many education product and service providers promise golden (educational) richness—if only we adopt their specific curricula, app, or training program. And through these promises, they seek to unleash the stampeders to their foothills.

But in this case, the stampedes are being funded by the public purse; governments around the world collectively spend countless billions each year on teaching resources, education technology, curriculum materials, and teacher professional development. We witness with despair that much of this investment is having insufficient impact. Too much is being invested in shiny things that look like gold but deliver little: the wrong drivers in education.

In this paper, we examine the seductive factors that have lured us all to embrace false premises, and we describe the hallmarks of the education "gold" that are worth our time and investment.

We have organized our thinking into four sections, summarized below.

1. Where Is the 4% Going?

Global expenditure on education exceeds USD $3.5 trillion per annum—with approximately 4% of this (or USD $140 billion) being invested in education

products, resources, and in-service teacher professional learning. We argue that despite this investment, the returns are way too low.

2. A Glitch in the Matrix

We make the case that ingrained cognitive biases make us all naturally predisposed to invest in educational products and approaches that conform with our existing worldview and to only grudgingly alter our behavior in the face of significant conflicting evidence. We argue that educators and policymakers must fight hard to overcome their cognitive biases and to become true evaluators of their own impact.

3. How Do We Really Know?

We explore some of the key challenges with theory and evidence generation in education—including the limitations of using lesson observations, student achievement data, and meta-analysis to distinguish convincingly between education pyrite and education gold.

4. Conclusion

Finally, we conclude that we must keep our vision focused on what is true "education gold." Our hope is that you will understand that we must worship evidence of impact.

1. Where Is the 4% Going?

We can all think of times when organizations or individuals we know spend their resources on the wrong things, declare success, and congratulate themselves on a job well done—despite overwhelming evidence to the contrary.

Sometimes they even fail to collect evidence because the mirror of reality is too much to bear. More often, though, it's the case that they neglect to look for the contrary evidence that their selected intervention may not have worked—*and that alternative actions may have yielded far greater results*. We put that last bit in italics because, in a very real sense, it's the bigger crime.

In education, for every weak resource or intervention that didn't move the needle on student progress or was only tepidly "successful," it means that a far more effective initiative wasn't in place to improve teaching and learning. A year, two or three years, or more can be squandered in this manner, not to mention the money involved. Michael Fullan (1982) famously lamented:

> Nothing has promised so much and been so frustratingly wasteful as thousands of workshops and conferences which led to no significant change in practice when teachers returned to their classrooms. (p. 315)

We suspect the same adage would apply to much of the technology and paraphernalia that educators also deploy in their schools.

The 4% Well Spent Could Be the Golden Bullet

Globally, the resources expended on education products, resources, and teacher training are gargantuan. According to the World Bank (2017), the Gross World Product (GWP), which is the sum of the Gross Domestic Product (GDP) for every nation on earth, currently exceeds USD $75 trillion per annum.

Of this USD $75 trillion, approximately 4.7% is spent on education, which in hard currency is about **USD $3.5 trillion per annum** (United Nations Educational, Scientific and Cultural Organization Institute for Statistics, 2013). To put it in perspective, this expenditure is greater than the combined economic activity of Russia and India, with their conjoined population of 1.4 billion citizens. Globally, we spend a lot on education. And rightly so.

But when we dig a layer down and try to uncover how much of this USD $3.5 trillion is spent on teacher salaries, infrastructure (buildings and information and communications technology), administration, transportation, teacher professional learning, and resources, it starts to get a little murky and we need to make some careful assumptions.

Most of the funding is for fixed and reoccurring costs that cannot be adjusted without great care and without expending high levels of political capital (these items are shown in gray in Table 1). In short, for better or worse, we are stuck with these "gray costs" and must make sure that the buildings, transportation, and, most importantly, teachers are primed for the most effective use in their core task: educating young people.[1]

The area where there is the most flexibility is the estimated 4% of global education budgets in the bold zone in Table 1—those that are available for the procurement of education products/resources for use in the classroom and for in-service teacher professional learning. We estimate that, globally,

[1]It is also worth noting that when Purchasing Power Parity (PPP) adjusted education expenditure per country is cross-tabulated to performance in the Programme for International Student Assessment (PISA) assessments, there is no clear relationship between higher spending and higher PISA performance. Finance alone cannot guarantee improved education outcomes. What that funding is spent on makes all the difference.

Table 1: Breakdown of Education Expenditure

No.	Expenditure Area	Estimated Percentage*	Estimated Total Global Expenditure (Rounded)
1	Facilities Operation, Maintenance, and Construction	10%	USD $352 billion
2	District, Regional, and National Administrative Support and Oversight	10%	USD $352 billion
3	Transportation and Food Services	9%	USD $316 billion
4	Student Services (health, nutrition, special needs, speech therapy, etc.)	7%	USD $246 billion
5	Teacher Salaries and Benefits	56%	USD $197 billion
6	**Education Products and Resources (books, education technology, etc.)**	**3%**	**USD $105 billion**
7	**In-Service Teacher Learning**	**1%**	**USD $35 billion**

***Note:** These percentages were estimated by reviewing education expenditure categories published by public authorities in G20 countries. Note that different budgetary accounting/reporting principles are applied in different jurisdictions and across time, increasing the probability of error. In addition, estimated percentages are unlikely to be representative of the expenditure profile of developing and fragile states, where items 1, 2, and 5 are likely to comprise most spending.

somewhere in the region of USD $140 billion per annum is spent in this category. This number is vast—greater than the combined GDP of Luxembourg and Oman—and yet it constitutes a tiny proportion of the whole.

But if this 4% is spent wisely and if, over time, there is also greater clarity of thought about how the other 96% is expended, then locally and globally we would expect to see remarkable things happening in education.[2] A well-spent 4% could be the proverbial "golden bullet" for education.

The trouble is, we're not seeing enough of those wonderful things. Global inequality in education outcomes is very far from being solved. Even in highly developed countries, large numbers of students are not graduating from secondary education with appropriate certification (2016 noncompletion rates were 33.1% in England, 27% in Australia, and 16.8% in the United States, according to the U.K. Department for Education, 2018; Australian Bureau of Statistics, 2018; and U.S. National Center for Education Statistics, 2018). The challenges in developing countries are far greater and almost too depressing to document. According to the United Nations Educational, Scientific and Cultural Organization (2014), at least 250 million of the world's 650 million primary school children are unable to read, write, or do basic mathematics. Most of these children are in developing countries and more than half have had at least four years of schooling.

Many have argued that this is a failure of society rather than of the quality of education systems (see Chudgar & Luschei, 2009), and they are right, to a point. The trouble is that we have also witnessed first-hand and through secondary research countless examples of schools operating in challenging situations that are making a real difference (Ofsted, 2009). So, we know that while the problem is societal, it can be solved through education—if we invest in unlocking and effectively implementing the right stuff. Surely, if it is not solved through education, then we need to question why we bother with schools at all!

[2]Of course, effective implementation is also crucial and we explore this in an upcoming paper, "Going for Gold."

Insomnia-Producing Realities

The issue that keeps us awake at night is the fear that the 4%, or USD $140 billion, is being spent on all the wrong areas and that this is why the equity gap has not yet been addressed. Our fear is that it is being spent on shiny toys that, on the surface, look like effective educational interventions but that hold no promise of helping us find education gold (see *What Doesn't Work in Education: The Politics of Distraction* by John Hattie [2015] for an overview of some of those cults of distraction).

We are all for diversity in teacher professional learning, curriculum materials, and student resources but that diversity must come with evidence of impact. The challenge for teachers and school leaders is like the one that many of us face on the weekend when we go to the supermarket for our weekly shopping. When we arrive at the supermarket, the product array is vast and we have relatively little time and information to make decisions. We do the best we can in the time available and often we fall into the habit of buying the same items over and over—because everyone else seems to buy those items, the packaging looks nice, we recognize the brand, and there's a risk that the alternative products might be worse.

The Marketplace: Pedagogy, Passion, and Profit

One of us gets one to three email requests a week to endorse a new education product (book, app, resource, etc.). When we ask two questions (Has it been deeply implemented outside of your class or school? Have you any evidence of impact on students?), 99% of the products fail and too few of the remaining 1% have valid and reliable evidence of impact. This is depressing indeed.

Our intuition is that, like the products in the supermarket, many of the items that educators use in schools or the training programs they undertake have been selected almost at random or because they look shiny and well packaged. If they work, that's great, but how do we really know they have a strong theory of change? How do we know that the product developers have evaluated their offerings to the highest standard, or that they have redeveloped their product or logic model based on any less than glowing testimonials? Too often, "they work" is assessed in terms of the author or developer's conviction and classroom experience, and perhaps teacher satisfaction, rather than a broad enough impact on students. Product developers and training providers always point to some evidence of impact. This is smart marketing on their part. Who has ever heard of someone trying to sell a product with the line "We think it probably works. We haven't got any tangible evidence, but other teachers say they like it"?

Ultimately, the case we want to make is that when you scratch beneath the surface, many of the claims made by educational product and service providers are no better than the quote above, albeit that they have more marketing finesse. Most of these are distractions or pyrite ("fool's gold"), and they should be buried where they can no longer damage the educational process.

We know these are strong words, and we also want to recognize that many education product and service providers work with scholars, researchers, and classroom practitioners who are deeply, passionately behind the pedagogy and are not merely out to make a buck.

A tough reality is that many quality education companies are not high-profit endeavors. To fully test a practice or intervention is time-consuming and very expensive, so they are caught between a rock and a hard place, erring on the side of getting the product out and pursuing proof of efficacy later.

We also want to acknowledge that many educational product developers do conduct in-house evaluations of their offerings, but these are often small-scale efforts and prone to bias. Neither of us can think of any example of an education service provider that has published or celebrated research showing that their product is bunkum.

Lastly, there are indeed big education companies who make vast sums on programs and products tied to standardized testing, curriculum, and content. In these cases, it's often the slick marketing and quest for shareholder value that drives the publicity and resulting frenzy for their products. These are the shops that deserve to sink to the ocean floor first—unless they redouble their efforts at collecting and evaluating their evidence of impact.

2. A Glitch in the Matrix

The stampeders flocked to Seattle not because they lacked common sense, but because common sense is less common than we might think. We all develop belief systems to survive in our busy, buzzing world, but sometimes our beliefs are ill founded or even utterly irrational.

The research supporting this comes largely from behavioral economics and particularly from the work of Amos Tversky, Daniel Kahneman, Herbert Simon, Cass Sunstein, and Richard Thaler (see the references for suggested further reading). During the 1970s and 1980s, they questioned a central tenet of economics: that human beings are rational and that we make decisions by carefully and explicitly calculating the positive and negative outcomes of each course of action.

The behavioral economists, whose research methods straddled into applied psychology, concluded that economists were probably the only rational humans and only because they were explicitly trained to be! Largely everyone else made decisions on the fly, with limited information, and tended to rationalize bad choices after they were made (often referred to as *cognitive dissonance*).

Our Inherent Biases

During the last forty years, a growing database of cognitive biases or glitches in our human operating system have been catalogued and confirmed through laboratory experiments and psychometric testing.

The research suggests that biases afflict all of us, unless we have been trained to ward against them. To date, behavioral economists have recorded more than eighty cognitive biases.

In Table 2, we summarize some of the inherent biases that, if left unchecked, can result in us stampeding to educational toys that look like wonderous glittering gold but that tarnish very quickly. These biases or negative mind hacks are significant hurdles to educators relentlessly reviewing and testing their assumptions about the impact that they are having on learning in the classroom and in selecting the right things in which to invest the precious 4%.

For educators to overcome these cognitive biases and fallacies, they need to further develop their skills of logic and rationality without losing their passion for teaching. Educators need not act like desperate stampeders. This requires the development of mindframes that enable educators to have an unrelenting focus on impact and to continually ask themselves *whether they are having the greatest impact that they could* and *how do they really know*?

Table 2: Inherent Cognitive Biases

Cognitive Bias Category	Description	References
Authority Bias	The tendency to attribute greater weight and accuracy to the opinions of an authority figure—irrespective of whether this is deserved—and to be influenced by it. **EDUCATION: Don't be swayed by famous titled gurus. Carefully unpack and test all of their assumptions, especially if they are making claims outside their specific area of expertise. Be particularly suspicious of anyone who writes and publishes a position paper.**	Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67*(4), 371–378.
Confirmation Bias **Post-Purchase Rationalization** **Choice-Support Bias**	The tendency to collect and interpret information in a way that conforms with rather than opposes our existing beliefs. When information is presented that contradicts current beliefs, this can transition into **Belief Perseverance** (i.e., where individuals hold beliefs that are utterly at odds with the data). **EDUCATION: We tend to select education approaches, products, and services that accord with our worldview and we will often continue to do so, even when convincing evidence is presented that our worldview may be distorted. Be prepared to go against the grain and to question sacred assumptions.**	Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220.
Observer Expectancy Effect **Observer Effect** **Hawthorne Effect** **Placebo Effect**	The tendency for any intervention, even a sugar pill, to result in improved outcomes—mainly because everyone involved thinks the intervention will work and this creates a self-fulfilling prophecy. **EDUCATION: If educational "sugar pills" can generate positive effect sizes, then well-crafted education "medicines" should generate a double whammy of effect plus placebo turbo boost—so opt for the latter.**	Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases, 32*(1–2), 51–63.
Ostrich Effect	The tendency to avoid monitoring information that might give psychological discomfort. Originally observed in contexts where financial investors refrained from monitoring their portfolios during downturns. **EDUCATION: The importance of collecting robust and regular data from a range of sources about the implementation of new interventions and analyzing this ruthlessly. Collect evidence to know thy impact.**	Galai, D., & Sade, O. (2006). The "Ostrich Effect" and the relationship between the liquidity and the yields of financial assets. *Journal of Business, 79*(5), 2741–2759.

Cognitive Bias Category	Description	References
Anecdotal Fallacy	The tendency to take anecdotal information at face value and give it the same status as more rigorous data in making judgments about effectiveness. **EDUCATION: Do not take spurious claims about impact at face value and do not invest in training based on participant satisfaction testimonials alone.**	Gibson, R., & Zillman, D. (1994). Exaggerated versus representative exemplification in news reports: Perception of issues and personal consequences. *Communication Research*, *21*(5), 603–624.
Halo Effect	The tendency to generalize from a limited number of experiences or interactions with an individual, company, or product to make a holistic judgment about every aspect of the individual or organization. **EDUCATION: Sometimes the whole is less than the sum of its parts. Just because an educational support organization has world-leading expertise in area A does not mean it has the same level of expertise in area B.**	Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*(4), 250–256.
Invented Here vs. Not Invented Here	The tendency to avoid using a tried-and-tested product because it was invented elsewhere—typically claiming "but we are different here." **EDUCATION: Be open to using and adapting existing ideas. Avoid reinventing the educational wheel—unless you work in terrain where wheels are useless (you probably don't).**	Piezunka, H., & Dahlander, L. (2014). Distant search, narrow attention: How crowding alters organizations' filtering of suggestions in crowdsourcing. *Academy of Management Journal*, *58*, 856–880.
IKEA Effect	The tendency to have greater buy-in to a solution where the end user is directly involved in building or localizing the product—like assembling an IKEA bookcase. **EDUCATION: Make the effort to localize and adapt tested solutions. This will generate greater emotional buy-in than standardized deployment.**	Norton, M. I., Mochon, D., & Ariely, D. (2011). The IKEA effect: When labor leads to love. *Journal of Consumer Psychology*, *22*(3), 453–460.
Bandwagon Effect **Illusory Truth Effect** **Mere Exposure Effect**	The tendency to believe that something works because a large number of other people believe it works. **EDUCATION: It might work and it might not. Test all claims carefully and don't blindly join the bandwagon to keep up with the Joneses.**	Mehrabian, A. (1998). Effects of poll reports on voter preferences. *Journal of Applied Social Psychology*, *28*(23), 2119–2130.

Cognitive Bias Category	Description	References
Clustering Illusion **Cherry Picking**	The tendency to remember and overemphasize streaks of positive or negative data that are clustered together in large parcels of random data (i.e., seeing phantom patterns). **EDUCATION: Are the claims made by educational researchers or service providers based on longitudinal data with a common long-term pattern or from a small snapshot that could have been cherry-picked?**	Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295–314.
Conservativism	The tendency to revise one's beliefs insufficiently when presented with information that contradicts our current beliefs. **EDUCATION: If the evidence is robust, it just might be true. There was a time when people who declared that the earth wasn't flat were burned as heretics. Carefully test all claims and evaluate evidence.**	Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York, NY: Cambridge University Press.
Courtesy Bias	The tendency to give an opinion that is more socially palatable than our true beliefs. **EDUCATION: Participant satisfaction scores from training events in some cultural contexts may be a grade or higher than the scores people would give if they were less polite.**	Morrison, L. J., Colman, A. M., & Preston, C. C. (1997). Mystery customer research: Cognitive processes affecting accuracy. *Journal of the Market Research Society*, *39*, 349–361.
Law of the Instrument	"If you have a hammer, everything looks like a nail" (Maslow, 1966, p. 15). The tendency to only address problems for which you already have a potential solution. **EDUCATION: Start with the problem or "wicked issue" you are trying to solve and then work backward to instruments, rather than searching for problems to which you have the solution.**	Maslow, A. H. (1966). *The psychology of science: A reconnaissance*. New York, NY: Harper & Row.
Bike-Shedding	The tendency to avoid complex projects, like world peace, to focus on projects that are simple and easy to grasp by the majority of participants—like building a bike shed. **EDUCATION: Build a bike shed if the world really needs bike sheds. If it doesn't, then fix what needs fixing most.**	Parkinson, C. N. (1958). *Parkinson's law, or the pursuit of progress*. London, England: John Murray.
Sunk Cost Fallacy	The tendency to continue with a project that is not bearing fruit, simply because so much has been invested in it already and withdrawal would be an admission of failure. **EDUCATION: Review implementation of new approaches regularly and set clear kill parameters/hurdles that must be achieved for the project to stay live. Ruthlessly prune anything that does not pass the hurdle test.**	Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Process*, *35*(1), 124–140.

3. How Do We Really Know?

We advocate an approach to education that is built on reason rather than intuition alone. This involves systematic collection of data on students' learning experiences in the classroom and the ways in which teachers and product developers can accelerate this learning. From data, we can inform intuitions and judgments and build theories. From theories, we can build structured processes—continually testing and refining these, too.

Toward a Unified Theory of Education?

The mathematical physicist Sir Roger Penrose (1989) developed a four-quadrant framework to categorize the various theories of science. He distinguished between the following:

- **Superb Theories**—which have been phenomenal in their range and accuracy.
- **Useful Theories**—which have either a narrower range of application or more imprecise predictive capability.
- **Tentative Theories**—which are similar to Useful Theories but without any significant experimental support (i.e., they seem to make sense but more evidence is needed).
- **Misguided Theories**—which are theories without experimental support and where there are lots of other competing and equally/more plausible theories (but Penrose refused to name the theories that should be placed in this quadrant).

In the following table, we have tentatively added to Penrose's original list; the items in bold are our additional suggestions.

Superb Theories	Useful Theories
Newtonian Mechanics	Quark Model
Einstein's Special Relativity	Big Bang Theory
Einstein's General Relativity	Darwin's Theory of Evolution
Quantum Theory	**Tectonics**
	Weather Systems

Tentative Theories	Misguided Theories
String Theory	**Cold Fusion**
Super Gravity	**Flat Earth Theory**
Grand Unified Theory	**Lamarckism**
	Æther
	Alchemy
	Astrology

To date, physics is probably the only discipline to field theories that would rank in the superb camp. Our intuition is that education is unlikely to have anything in the superb category until the neuronal structures of the brain have been fully mapped and simulated and then related to learning. Any superb theory of education is likely to be a tripartite arrangement between cognitive psychology/neuroscience, computer science, and education. This might be a bridge too far right now and it may never come.

Our contention is that most contemporary theories in education straddle the tentative and misguided categories, perhaps with some wiggling into the useful category. Most lack significant empirical evidence (beyond the intuitions of educators), have limited (successful) predictive capability, and, in many cases, are pitted against competing theories with an equally slim evidence base.

Even Visible Learning, which is arguably among the most comprehensive syntheses of meta-analyses of what works best in education, is very far from meeting the criteria of a superb theory. At best, it would likely straddle the bottom of the useful category, but most likely, it would sit firmly in the tentative category. To make it into the top two categories, it is probably necessary for a more explicit set of theoretical tenets to be elaborated and subjected to falsification.

Proxies for Learning

The human brain is the most complex machine in the known universe. Housed beneath 7-millimeter-thick bone plating, its operations remain one of the enduring mysteries of science.

In the absence of hard evidence from the brain itself, we have to infer student learning and the impact of educational products and teacher development programs indirectly. Some of the main indirect proxies for learning that we have at our disposal are as follows:

- **Lesson Observation:** watching and listening to the interaction between learners and teachers.
- **Assessment:** using the outcomes of standardized high-stakes tests to infer something about the quality of learning and teaching.
- **Meta-Analyses:** collating the findings from multiple research projects, conducted in many different ways, and aggregating them to draw holistic conclusions about what works more, what works less, and what doesn't work at all.

In the sections that follow, we recap some of the inherent challenges with each of these approaches in helping us identify education gold.

The Limits of Lesson Observation

In many education systems, it is a mandatory requirement that every teacher undergoes at least an annual observation by his or her school leader. Heads and principals generally use some form of rubric or scoring sheet and rate their teachers against this. At our last count, we located more than 120 observation forms that had been published with some evidence about their reliability and validity.

These observations are often used for performance management purposes, to identify who are the "good" and "less good" teachers, and by national inspectorates, to make more holistic judgments about whether a school is outstanding, good, or poor. They are also used for developmental purposes, such as when teachers peer review each other's lessons so they can offer one another advice and harvest good practice to apply back in their own classrooms. Finally, these observations can be used to sift education pyrite from education gold, by observing the impact of a new education product or teacher development program in the classroom.

But we should ask ourselves an important question: can you actually see, hear, and sniff a good lesson? Are our five senses any good at measuring outstanding, adequate, and poor? Can we see the impact of the teacher on each student in a class? Do we watch the teacher's performance or do we watch the impact on the students from this performance? And, what if the performance is spectacular but the impact is of little consequence?

If we rephrase these questions as binary yes/no choices, then the answer to whether we can make meaningful and rigorous observations is a resounding *yes*. Here are some examples:

- Is the teacher in the classroom?
- Is the teacher talking to the class?
- Are the children all awake?
- Has homework been set and marked?

It's relatively straightforward to establish a sampling plan for each of these questions and any two observers will have a high degree of consistency in their observations (with minimal training), even if they are not educationalists.

So, for these kinds of binary questions about performance, we can see, hear, and sniff reasonably reliably. We could probably stretch from asking binary questions to inquiring about frequency, such as how often something occurred (e.g., Were all the students awake, all the time, during the lesson?).

But when we want to use observation to determine whether the teacher delivered a high-quality lesson, we ask questions like these:

- Did the teacher deliver a "good" lesson?
- Did all the students "achieve" the learning objectives?
- Were the learning objectives worthwhile, appropriate, and sufficiently challenging for the students?
- Was the classwork a "good" fit with classroom-based activity?
- Did the teacher provide "good" feedback on the classwork?
- Were the education products "effective"?
- Did the teacher-training program deliver "impact" in the classroom?

With these questions, we open a huge can of worms. Who decides what "good" is, and who decides what "impact on students" means?

To answer these questions, observers rely on proxies for learning. A proxy measure is when we use one thing that's quite easy to get data about to tell us about something else, which is much more difficult to get data about. For example, doctors rely on blood tests, blood pressure measurements, and heart rate analyses to tell them whether a patient is fit and well. Generally, these tests work relatively well. However, it's possible to have a rare type of illness that does not show up on these types of tests, which means that you might be given a

clean bill of health by the doctor but actually be at death's door.

It's the same with lesson observations. It is possible that when we measure with our eyes, we are looking in the wrong areas. We may see busy, engaged students in a calm and ordered classroom where some have supplied the correct answers, and we conclude that a heck of a lot of learning is going on. Yet it is quite possible that absolutely nothing of any significance is being learned at all (as in the good old days where teachers practiced their lessons before the inspector came).

We know, too, that much of what goes on inside the classroom is completely hidden. The late great Graham Nuthall, in his seminal work *The Hidden Lives of Learners* (2007), theorizes that there are three separate cultural spheres at play in the classroom: the **Public Sphere**, which in theory is controlled by the teacher; the **Social Sphere** of the students, which the teacher is often unaware of; and the **Private Mental Worlds** of the students themselves, which both the teacher and the other students are unable to directly access. In short, most of what goes on in the classroom is inaccessible to the teacher and less still to a third-party observer.

Confounding this, the evidence from neuroscience suggests that of the vast array of data collected by our various senses each second, very little is actively processed by the conscious mind. So even within the Public Sphere that we have direct access to as observers, it's likely that we see very little. As we focus narrowly on some aspects of classroom practice, we miss the stooge in a gorilla suit dancing across the room. As observers, we have our own lens, our own theories, and our own beliefs about what we consider is best practice and these can bias the observations, no matter how specific the questions in any observation system. Most observations of other teachers end up with us telling teachers how they can teach like us!

The challenge with observation is that often we end up seeing what we want to see and we can be guided by our cognitive biases. The process of observing is like interpreting a **Rorschach Image**, one of those inkblot images that psychiatrists show their patients, where some say they can see their mother and others JFK.

There has been quite a lot of research into the problem of lesson observation in the last few years. The strongest dataset comes from the Measures of Effective Teaching (MET) project, which was funded by the Bill & Melinda Gates Foundation (2013). The MET study concluded that a single lesson observed by one individual, where the purpose was to rate teacher performance, has a 50% chance of being graded differently by a different observer. In the best-case scenario, where a teacher undergoes six separate observations by five separate observers, there is "only" a 72% chance that there is agreement and thus a 28% chance that their judgments are misaligned to the lesson observation rubric.

Now that's a whole lot of observation for still almost a 1/3 chance of error.

Observers frequently disagree about what they are observing, even with a well-established observation schedule. In assessment, we call this the inter-rater reliability problem.

The MET project found that:

1. Observers rarely used the top or bottom categories ("unsatisfactory" and "advanced") on their observation instrument. (**Courtesy Bias**)
2. Compared to peer raters, school leaders differentiated more among teachers. The standard deviation in underlying teacher scores was 50% larger when scored by school leaders than when scored by peers (i.e., leaders were more likely to be harsh). (**IKEA Effect**)
3. But school leaders rated their own teachers higher than leaders from other schools. (**Invented Here vs. Not Invented Here**)
4. When an observer formed a positive impression of a teacher, that impression tended to linger, even if the teacher's performance had declined. (**Halo Effect**)

In short, the whole process of lesson observation (when used to measure the impact of training, a new product, or the effectiveness of a teacher) is riddled with many of the cognitive biases that we described in section 2 ("A Glitch in the Matrix").

We know that observations work much better for frequency questions, such as how often something happens (e.g., *How often does the teacher promote, set goals, review, repeat comments, deepen understanding, make connections, and use open and closed questions?*). In our work, we use frequency questions and an automatic coding system to observe class lessons and can achieve very high levels of reliability. The same automated system can ask students about their learning (e.g., *My teacher explains difficult things clearly. In this class we learn to correct our mistakes. When I am confused, my teacher knows how to help me understand. My teacher checks to make sure we understand what is being taught*).

This allows, at least, a perspective from both the teacher and students. Such information can be useful for teachers to see their impact through the eyes of the students, have dependable information about what they actually did, and have comparisons to normative information from many thousands of teachers on these observed behaviors.

The research on micro-teaching also suggests that the act of video recording lessons and then peer reviewing those recordings can have significant impact. This is powerful for teacher development, but it is a step too far to then use this information for product or teacher evaluation (although teachers may choose to use aspects of the observations as part of their claims about effectiveness of their teaching approach, provided it is interpreted alongside other triangulated information).

The Curve Before the Plateau

Much of the research into teacher development tells us that educators have a very steep learning curve during their first few years in the profession (see Henry, Bastian, & Fortner, 2011). Indeed, they learn half of what they end up knowing about how to teach in their first year, half as much again in their second, and then it gets reasonably flat after that (and note that they learn hardly anything from initial teacher-training programs!). Perhaps this curvilinear growth reflects why the teacher pay growth is often similar (pay flattens out after a few years) and why those who fail to make this quick increase are more likely to leave the profession within the first five years. During those early years, teachers are engaging in what Daniel Kahneman calls *slow thinking*. Their learning is deliberate, effortful, and tiring.

Thinking Fast and Thinking Slow

Daniel Kahneman (2011) distinguishes between two types of thinking:

- **Slow Thinking**—which is deliberate, reflective, and effortful. We employ this type of thinking when we are learning new skills, such as how to speak a foreign language, drive a car, or "teach like a champion."
- **Fast Thinking**—which is automatic, reflexive, and effortless. We draw on this type of thinking when we have mastered a skill and where it no longer makes our brains hurt to exercise it.

After about three years in the job, teachers often shift to fast thinking. The steep learning curve has plateaued and their actions become more automatic and less reflective. In the early years, they are much more open to evidence of what is working and what is not—they have not developed routines that they apply and are more willing to learn from what is and what is not working with students. When they move to fast thinking, teachers can stop learning, stop reflecting, stop self-evaluating, and stop improving their own performance. They believe their methods work but the students must have some faults when they do not respond to their tried-and-tested methods (e.g., It worked for other students, so why not these ones?).

Deliberate Open-Mindedness

The fact that teaching experience and "wisdom" don't necessarily lead to continually deepening,

improving practice is a bitter pill to swallow. We want it to be true; it seems that it should be true, right? The sweet spot is when teachers engage in meaningful peer lesson observation for development purposes and watch not the teacher, but the *impact* of the teacher. This helps keep the fires of enthusiasm and experimentation burning. Lesson studies focused on the impact of the lesson on students also can help increase openness to new ways to impact learning outcomes.

But we want to reiterate that if our goal is to measure education quality to definitively test which interventions will lead us to gold, lesson observations alone will not give us a robust answer, because we can't see everything with our eyes. There is no such thing as immaculate perception (Nietzsche, 1891).

Assessment

High-stakes assessment has been an important rite of passage throughout much of human history. Many ancient cultures and tribal societies required their young to undertake risky and painful quests to mark the transition to adulthood. For the Australian Aboriginals, this involved boys surviving unaided in the outback for up to six months, using the skills that they had been taught during childhood. For some African tribes, it involved successfully hunting a lion. In some South American communities, the transition to adulthood involved being able to demonstrate a very high threshold for pain, including the imbibing of neurotoxins.

The ancient Chinese were possibly the first to develop a national written assessment system. This was called the Imperial Examination and it was used as a mechanism to select administrators for government posts (Fukuyama, 2011). The system originated in 605 AD as a way of avoiding hereditary appointments to government office. Candidates would be placed in individually curtained examination cells to undertake the written assessment, which lasted for several days. At night, their writing board doubled as a bed.

It is this rite of passage that we continue to deploy in the form of national school-leaver examinations, such as the SAT and the International Baccalaureate (IB), today. Modern educational assessments are high stakes but without the physical risk of the tribal tests (although they can invoke high levels of stress). Different times, different measures. The SAT, A Levels, IB, and other assessments signal to employers and training providers that school leavers have acquired the required skills for the next stage of their journey.

These assessments can tell us, often with relatively high levels of accuracy, a student's level of competence in mathematics, literacy, foreign languages, and science and about the depth and breadth of knowledge the student has acquired across a range of curriculum areas. From this, we can also make inferences about a student's readiness for university studies and life beyond school, albeit with less precision (as we may need to also include the proficiency to learn, address challenges, be curious, feel a sense of belonging in more open learning environments, achieve financial security, and gather support from others).

Navigating by the Light of the Stars

The outcomes of high-stakes summative assessments are also often used to make inferences about the quality of schools (e.g., school league tables), school systems (e.g., PISA, Trends in International Mathematics and Science Study [TIMSS], and Progress in International Reading Literacy Study [PIRLS]), and individual teachers and about whether certain education products and programs are more effective than others. In other words, they are often used in the quest to find education gold.

In this context, high-stakes assessments are blunt instruments—akin to piloting your boat by the stars on a cloudy night, rather than a GPS system. We can infer something about which schools are higher and lower performers, but we need to carefully tease out background variables like the starting points and circumstances of the learners and multiple other important outcomes, so that we can measure the distance traveled rather than the absolute end point in one set of competencies. Indeed, all too often we find that the greatest variability in learning outcomes is not between different schools but between different teachers within the same school (McGaw, 2008). The key unit of analysis should be the teacher rather than the school, and many high-stakes assessments may not be attributable to a particular school.

In the context of individual teachers (provided there is a direct link between the teacher and the particular content assessed), the outcomes of high-stakes assessments can tell us quite a lot about which teachers are more or less effective—particularly where the pattern of performance holds over several years. Again, care is needed, as it is not only the outcomes of the assessments but the growth from the beginning to end of the course that should be considered. Otherwise, those teachers who start with students already knowing much but growing little look great, and those who start with students who know less at the beginning but grow remarkably look poor—when it should be the other way around.

But unless the outcomes of high-stakes student assessments are reported back to schools at the item level (i.e., how well students did and grew on each component of the assessment, rather than just the overall grade), teachers are left in the dark about which elements of their practice (or third-party products) are more or less effective or completely ineffective. They just know that overall, by the light of the stars, they are navigating in the right or wrong direction. And even where they are navigating in the wrong direction, there are likely some elements of their tradecraft or product kitbag that are truly outstanding but are missed.

Even where teachers are able to access item-level data from high-stakes evaluation, the inferential jump that they must make to systematically map this back to specific elements of their tradecraft or the impact of specific training programs or pieces of educational technology is too great to do with any meaningful fidelity. In other words, the outputs of high-stakes examinations are not reported at high enough resolution to tease out, with high confidence, the educational cargo cults from education gold. So, often, they are an event (two to three hours on one day) and the inference from this event to the teaching and learning is too great a leap.

Navigating With a GPS System

The only way we can use student achievement data with any sense of rigor to sift out the education gold is by collecting data (formatively) at the beginning, middle, and (summatively) end of the journey to systematically measure the distance traveled by individual students and groups of learners. By experimentally varying very narrow elements of teacher practice or aspects of educational products and programs, we can see whether this results in an upward or downward spike in student performance. It is as important to know about the efficiency and effectiveness of the journey as it is to reach your destination. This is one of the benefits of GPS systems.

Summative vs. Formative Evaluation

Too often, teachers see summative assessment as "bad" and formative assessment as "good" when this is nonsense; some see summative as needing to be highly reliable but with formative, the measurement rigor can be less. If formative is more powerful, then it, too, needs to be based on highly valid measures and observations.

We prefer to use the terms *formative* and *summative evaluations* and abandon the misleading terms formative and summative assessments. Our arguments and analysis in this section have principally been about the use of summative evaluation as a systematic mechanism to make inferences about what's education gold. But we want to stress that it's more often about what it is used for than the mechanism of data collection itself. That is, the same assessment instrument can be used both formatively and summatively. As Bob Stake puts it: when the cook tastes the soup, it is formative; but when the guest tastes the soup, it is summative.

Within the context of the individual teacher in the individual classroom, we know that formative

evaluation is educational gold in and of itself (Hattie & Timperley, 2007). The most effective approach to formative evaluation contains three components:

- **Feed-up:** Where am I going?
- **Feed-back:** How am I doing?
- **Feed-forward:** What is my next step?

What is important is not the testing itself but the way that it is incorporated into the cycle of challenging goals to support learners in unlocking the skill, will, and thrill to learn.

The challenge, of course, is that "everything seems to work somewhere and nothing everywhere" (Wiliam, 2014). So, even where this analysis is conducted systematically, we cannot be completely certain that the educational approach, training program, or technology intervention that resulted in education gold in one context will not end up being pyrite in quite another.

We need repeated evaluation projects that investigate the same approaches across many different contexts to give us much greater confidence in the fidelity of our findings. And once we have these data, we face the challenge of vacuuming them up from disparate sources and in drawing the common threads to build a compelling narrative about what's gold. We can then ask not only about overall effects, but under what conditions and for which students programs work best. Thankfully, a great deal of progress has been made here through the use of meta-analysis and we discuss this in the next section.

Meta-Analysis

Before we describe meta-analysis, we want to be overt and lay out a potential conflict of interest. One of us (the older one) has spent the best part of 30 years collecting and aggregating the findings from meta-analyses, which in 2009 was crystalized into *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement* (Hattie, 2009). Given what we have said earlier about the power and hold of cognitive biases on thought processes, you might want to bear in mind the Sunk Cost Fallacy (Arkes & Blumer, 1985), which suggests that we humans tend to continue with a project even if it's not bearing fruit, simply because so much has been invested already and withdrawal would be an admission of failure.

We want to assure you that the Sunk Cost Fallacy is not at play in this instance (although we would say that, wouldn't we?). In any case, we urge you to read on and decide for yourselves.

Thus far, we have outlined some of the challenges involved in using lesson observation and student assessment data to distinguish educational gold. Now we take you on a brief tour of the pitfalls in meta-analysis; we explore the topic in much greater depth in the forthcoming paper, "Real Gold vs. Fool's Gold: The Visible Learning Methodology for Finding What Works Best in Education."

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. These studies can range in size and scope from a few days of action research with a single school to longitudinal study involving several hundred schools. Education researchers use a variety of methods and measures to do their work, comparing a program with others, relating one program with various attributes of students, teachers, and schools, and comparing students over time.

Researchers can then use many statistical methods to make these comparisons (*t*-tests, ANOVA, regression, correlations). Each of these can be

converted into a common metric (an effect size) that provides a measure of the magnitude or size of the effect. Since the early 1980s, many quantitative educational researchers have habitually included the effect size scores in their research outputs. This means that there is currently effect size data from more than 90,000 studies involving more than 300 million students.

But making sense of all these data is extremely hard. To collect, sift, and sort the more than 90,000 education research studies that include effect size data requires a process. Gene Glass, an educationalist, invented a method called meta-analysis in the 1970s that provided educational mavens with a process for collecting and categorizing primary research studies (Glass, 1976; Glass et al., 1981). (Many wrongly believe meta-analysis was invented in medicine and adopted into education, but here is a case of the opposite.) Most importantly, the method provided a way to weight the different pieces of research based on their respective methodologies and to then aggregate the disparate effect size scores into an overall score.

Exploring Common Challenges

The meta-analyses approach, like all other educational research methods, is not free from challenges or criticism. Some of the more common challenges with the meta-analysis approach are as follows:

1. **One number cannot summarize a research field.** The criticism is that meta-analysis focuses on the holistic summary data and that it ignores the fact that the treatment effect may vary very widely from study to study.

   However, if the average effect is reasonably consistent across meta-analyses on the same topic, we can have some confidence in the consistency of the effect. Where there is inconsistency, this is worthy of deeper investigation because it can reveal important information about where, when, and how the influence may vary. A win-win.

2. **Meta-analyses suffer from the "file drawer problem."** This is the argument that education researchers are only likely to publish data that show positive findings and that, because of this, the meta-analyses are likely to present a high proportion of false positives.

   This is one of the reasons that Visible Learning sets the effect size bar so high (i.e., $d > 0.40$). This helps weed out false positives (which are more likely to have lower cumulative effect size values) and focuses everyone's attention onto the interventions with the highest probability of impact.

   There are also statistics for estimating the number of papers still stuck in someone's file drawer that could lead to the decisions being nullified. But we also need a global register of educational research projects that researchers sign up for before their project begins and with whom they register their findings, even if these are negative (as is now done routinely in medicine).

3. **The primary data are Western-centric and some of them are quite old.** Here, the argument is that most of the original research that the meta-analyses draw on was conducted in English-speaking developed countries and thus it cannot be applied with confidence to other contexts.

   All reviewing of literature is rear-view mirror research (recall that *research* means re-search, or searching again), but try driving forward ignoring the rear-view mirror. Ouch. The research can be used with much greater confidence to distinguish educational gold in the contexts of developed countries. This does not mean that the current research has nothing to say about Sub-Saharan or other developing contexts, but higher levels of caution should be applied.

   It is likely that, for now, we should constrain inferences to countries where the between-school variance is much smaller than the within-school variance (which is more unlikely in developing countries). We also need a globally coordinated movement that proactively identifies gaps in the research

and "crowdsources" data collection through affiliated Ministries of Education or research institutes, particularly in developing countries.

4. **Meta-analyses don't help you implement solutions.** Here, the quite reasonable argument is that although meta-analyses provide a useful overview at 40,000 feet about what works, they become much less valuable at 5,000 feet, let alone 5 feet.

   There is currently no sorting house that maps productized educational offers and approaches directly to the evidence of what works. Currently, teachers and leaders are left without any maps or guideposts to help them divine the good, average, and poor bets for learning. We have hardly any theories about implementation methods, often leaving this to the chance of each school leader. The issue today is probably not that there is a lack of evidence; rather, there is a lack of evidence about effective implementation of this evidence, which we explore further in "Real Gold vs. Fool's Gold."

5. **Meta-analyses are a very reactive research approach.** This is the argument that analysts are passive collectors and aggregators of research and that they can only analyze what others chose to research. This means that there are potentially major gaps—some areas have been over-mined, others lightly mined, and yet others not mined at all.

   To date, the process has been almost entirely a reactive search and consolidation strategy. However, some research funders, like the Education Endowment Foundation in the United Kingdom, are starting to explicitly map the knowledge gaps and prioritize these for research funding.

6. **Meta-analyses come in various hues of quality.** This is indeed the case for meta-analyses and for the original studies on which they are based. Since day one, there have been methods for asking about the effects of lower-quality studies and whether they should be omitted (yes, if the lower-quality studies adversely affect the overall effect size).

   The challenges listed above do not mean that meta-analysis should not be used; rather, these issues need attention when meta-analysis is used. The message is about looking forward (i.e., out of the driver's window), while taking into account the rear-view mirror perspective. And the benefit of meta-analysis is that it is able to harvest the data from various studies that have used different research methodologies (including lesson observation and analysis of student achievement data) and synthesize them into a more definitive account of the interventions, products, and training that have the most impact in the classroom.

4. Conclusion

We are driven by the desire to give teachers, school leaders, and policymakers good-quality tools to distinguish education pyrite from education gold, so that they can use the 4% of educational resources effectively. In this paper, we have highlighted some of the challenges with unmediated use of lesson observations, student assessment data, and meta-analysis as homing beacons to identify, with precision, what works best.

It's not that these tools are blind alleys or akin to reading tea leaves. It is more that they must be used and interpreted carefully, and that often there is more than one possible interpretation and more than one causal theory.

Teachers, school leaders, and policymakers are all busy people with incredibly difficult but rewarding day jobs to undertake. But the ways in which each accesses information about what works more, and what works less, in the classroom can be random and riddled with cognitive bias.

Most of the killer research is trapped behind paywalls or subscription services and written in language that is often inaccessible. And, by contrast, quite a lot of the research that is publicly available is written in pursuit of a particular agenda (to convince other academics!). Busy teachers and busy policymakers rarely have the time to find and sift these data with the rigor and tenacity required. There just aren't enough hours in the day. Hence, there is the tendency to fall back on our heuristics, cognitive biases, and hunches when making decisions about what works in the classroom.

A good system for educators to follow in sifting education pyrite from education gold may be found in the work of Daniel Willingham, who tackles this problem head-on in *When Can You Trust the Experts? How to Tell Good Science from Bad in Education* (2012). He suggests a four-step process to sifting through the claims to unleash the gold:

1. **Strip it.** Clear away all the verbiage in the marketing materials and vector in on the *actual* claims. What, specifically, is the claim suggesting an educator should do, and what outcome is guaranteed or, at least, promised?
2. **Trace it.** Ask who created the product, program, or idea and what have other experts said about it? It's common to believe something because an authority confirms it, and this is often a reasonable thing to do. In education research, however, this can be an extremely weak indicator of truth. (**Authority Bias**)
3. **Analyze it.** Why, exactly, are you being asked to accept that the claim is true? What evidence is proffered (and is this good evidence from systematic independent studies or merely anecdotal)? And how does this claim square with your own experience? (Be careful you are not being guided by **Confirmation Bias**.)
4. **Should I do it?** You are not going to blindly adopt every educational program that is developed by a trusted expert or that has a strong theory of change and robust evidence of impact, because these specific offerings may be solutions to problems you simply do not have. And, very occasionally, it may make sense to adopt a program that has not been scientifically evaluated, because you have an urgent need. But you need to *very* carefully consider the likelihood of impact before handing over your share of the precious 4%.

If only the Seattle stampeders had a similar method for evaluating the tools and resources they thought they needed on their journey for gold—and, of course, for evaluating whether the journey was really worth their time and effort in the first place!

In this paper, we have lamented that the large global investment in education is having insufficient impact. Too much is being invested in shiny things that look great but have too little evidence that they are delivering on their promises.

Our argument is that policymakers and educators must be more discerning in how they collectively spend the USD $140 billion that we estimate is expended annually on educational resources, technology, and teacher professional learning. If this funding is focused with more laser precision on effective interventions, there is a much greater probability that every learner will be able to fulfill his or her full potential.

To make the right kinds of investments, policymakers and educators need to be aware of their cognitive biases and the ways in which these can drive us all to covet and privilege the wrong things. They also need to understand the limitations of lesson observations and student achievement data in making cast-iron inferences about what works best, and they should consider the potential benefits of meta-analysis.

However, we appreciate that policymakers and educators are busy folk with limited free capacity to explore claim and counterclaim about what works best. This is why we continue to harvest, synthesize, and disseminate meta-analyses, helping ensure that the collective wisdom is spread far and wide.

We can't stampede toward the distractions that keep us from focusing our pursuit on education gold. Instead, we must privilege evidence of impact and we must use this evidence to ensure that every learner gets a year's growth for a year's input.

References

Arkes, H., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Process*, *35*(1), 124–140.

Australian Bureau of Statistics. (2018). Schools Australia. Retrieved from https://www.abs.gov.au/ausstats/abs@.nsf/mf/4221.0

Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. MET Project Policy and Practice Brief. Seattle, WA: Author.

Chudgar, A., & Luschei, T. F. (2009). National income, income inequality, and the importance of schools: A hierarchical cross-national comparison. *American Educational Research Journal*, *46*(3), 626–658.

Fukuyama, F. (2011). *The origins of political order: From prehuman times to the French Revolution*. New York, NY: Farrar, Straus and Giroux.

Fullan, M. (1982). *The meaning of educational change*. New York, NY: Teachers College Press.

Galai, D., & Sade, O. (2006). The "Ostrich Effect" and the relationship between the liquidity and the yields of financial assets. *Journal of Business*, *79*(5), 2741–2759.

Gibson, R., & Zillman, D. (1994). Exaggerated versus representative exemplification in news reports: Perception of issues and personal consequences. *Communication Research*, *21*(5), 603–624.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295–314.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

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

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

Hattie, J. A. C. (2015). *What doesn't work in education: The politics of distraction*. Open Ideas at Pearson. Retrieved from https://www.pearson.com/hattie/distractions.html

Henry, G. T., Bastian, K. C., & Fortner, C. K. (2011). Stayers and leavers: Early-career teacher effectiveness and attrition. *Educational Researcher*, *40*(6), 271–280.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* New York, NY: Cambridge University Press.

Maslow, A. (1966). *The psychology of science: A reconnaissance*. New York, NY: Harper & Row.

McGaw, B. (2008). How good is Australian school education? In S. Marginson & R. James (Eds.), *Education, science and public policy: Ideas for an education revolution* (pp. 53–77). Carlton, Victoria, Australia: Melbourne University Press.

Mehrabian, A. (1998). Effects of poll reports on voter preferences. *Journal of Applied Social Psychology*, *28*(23), 2119–2130.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*(4), *371–378*.

Morrison, L. J., Colman, A. M., & Preston, C. C. (1997). Mystery customer research: Cognitive processes affecting accuracy. *Journal of the Market Research Society*, *39*, 349–361.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*(4), 250–256.

Norton, M. I., Mochon, D., & Ariely, D. (2011). The IKEA effect: When labor leads to love. *Journal of Consumer Psychology*, *22*(3), 453–460.

Nuthall, G. (2007). *The hidden lives of learners*. Wellington, New Zealand: NZCER Press.

Ofsted (Office for Standards in Education, Children's Services and Skills). (2009). *Twenty outstanding primary schools excelling against the odds*. London, England: Crown Copyright.

Parkinson, C. N. (1958). *Parkinson's law, or the pursuit of progress*. London, England: John Murray.

Penrose, R. (1989). *The emperor's new mind: Concerning computers, minds, and the laws of physics*. New York, NY: Oxford University Press.

Piezunka, H., & Dahlander, L. (2014). Distant search, narrow attention: How crowding alters organizations' filtering of suggestions in crowdsourcing. *Academy of Management Journal*, *58*, 856–880.

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, *32*(1–2), 51–63.

Simon, H. A. (1983). *Reason in human affairs*. Stanford, CA: Stanford University Press.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

U.K. Department for Education. (2018). School and college performance tables. Retrieved from https://www.gov.uk/government/organisations/department-for-education

United Nations Educational, Scientific and Cultural Organization. (2014). *Teaching and learning: Achieving quality for all*. Retrieved from https://en.unesco.org/gem-report/report/2014/teaching-and-learning-achieving-quality-all

United Nations Educational, Scientific and Cultural Organization Institute for Statistics. (2013). Data for the Sustainable Development Goals. Retrieved from http://uis.unesco.org

U.S. National Center for Education Statistics. (2018). *Digest for education statistics: 2016*. Retrieved from https://nces.ed.gov/programs/digest

Wiliam, D. (2014). *Why education will never be a research-based profession and why that's a good thing*. Retrieved from https://www.dylanwiliam.org/Dylan_Wiliams_website/Presentations_files/2014-09-06%20ResearchED.pptx

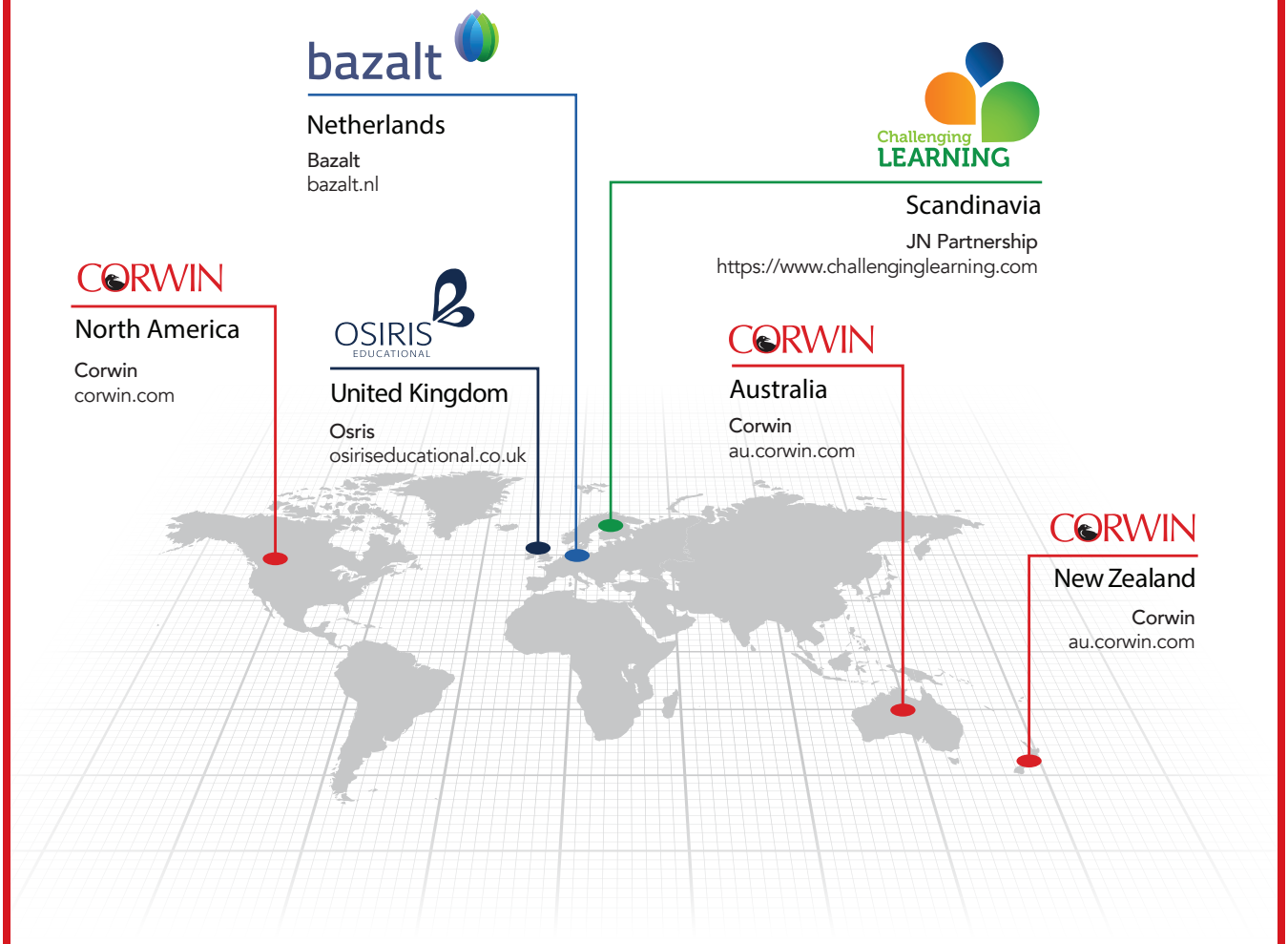
Willingham, D. T. (2012). *When can you trust the experts? How to tell good science from bad in education*. San Francisco, CA: Jossey-Bass.

World Bank. (2017). World Development Indicators Database. Retrieved from https://datacatalog.worldbank.org/dataset/world-development-indicators

CORWIN

To learn more

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



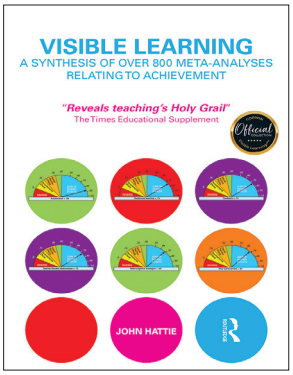
CORWIN Visible Learning$^{plus®}$

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

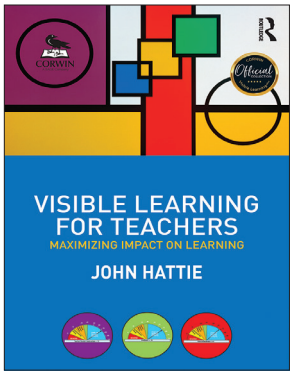
CORWIN Visible Learning plus®

Build your Visible Learning™ library!

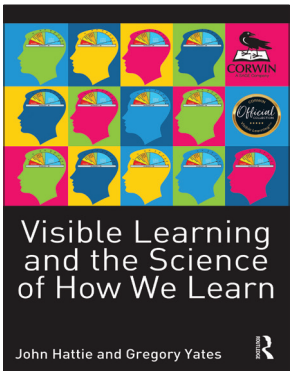
Foundation Series



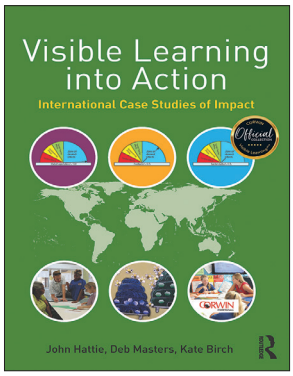
VISIBLE LEARNING



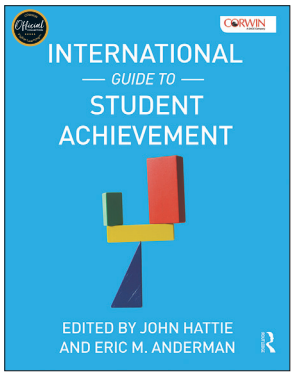
VISIBLE LEARNING FOR TEACHERS



VISIBLE LEARNING AND THE SCIENCE OF HOW WE LEARN



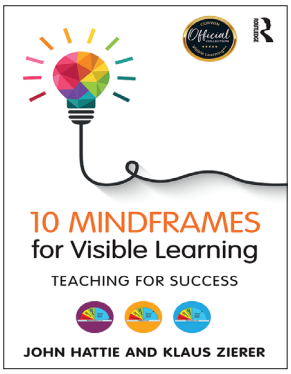
VISIBLE LEARNING INTO ACTION



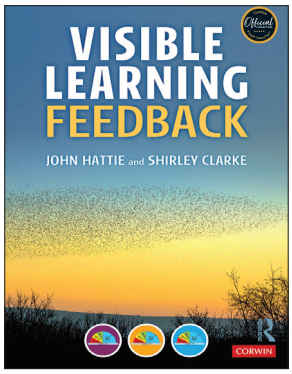
INTERNATIONAL GUIDE TO STUDENT ACHIEVEMENT

Visit corwin.com/vlbooks

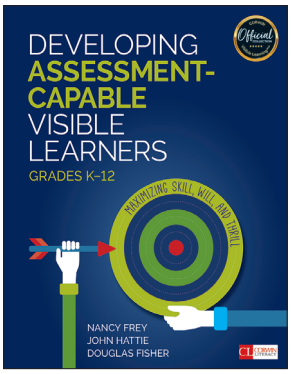
Impact Series



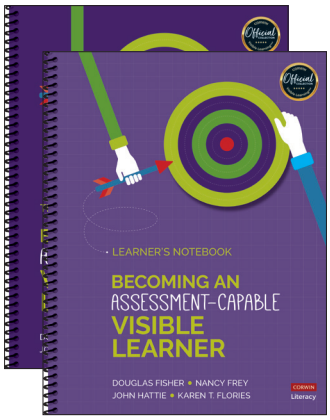
10 MINDFRAMES FOR VISIBLE LEARNING



VISIBLE LEARNING FEEDBACK

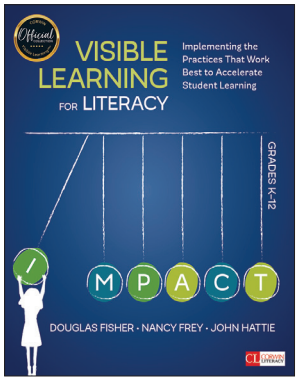


DEVELOPING ASSESSMENT-CAPABLE VISIBLE LEARNERS, Grades K–12

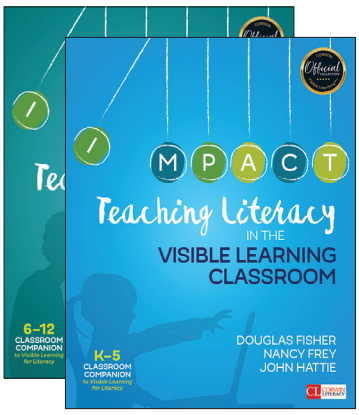


BECOMING AN ASSESSMENT-CAPABLE VISIBLE LEARNER Teacher's Guide & Learner's Notebooks Grades K–2, 3–5, & 6–12

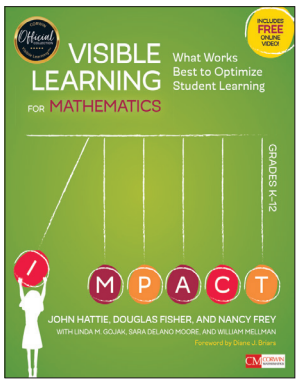
Practice Series



VISIBLE LEARNING FOR LITERACY, Grades K–12

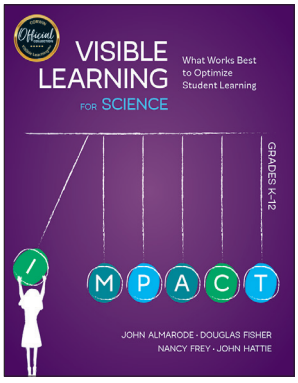


TEACHING LITERACY IN THE VISIBLE LEARNING CLASSROOM, Grades K–5 & 6–12



VISIBLE LEARNING FOR MATHEMATICS, Grades K–12

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



VISIBLE LEARNING FOR SCIENCE, Grades K–12

CORWIN Visible Learning plus®

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.

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.