# Education Cargo Cults *Must Die*


John Hattie
Arran Hamilton

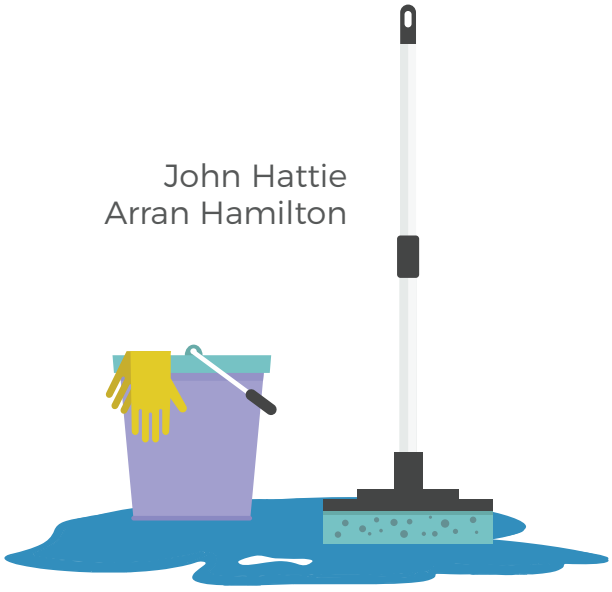
SEPTEMBER 2018

CORWIN
A SAGE Publishing Company

# Contents

# Education Cargo Cults *Must Die*

John Hattie
Arran Hamilton



# 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 up educators around the world. Visible Learning means that students 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.

# 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 of Teaching and School Leaders. His areas of interest are measurement models and their applications to education's problems, and models of teaching and learning. He has published 31 books, published and presented over 1000 papers, and supervised 200 thesis students.

**Dr. Arran Hamilton** is Group Director of Strategy 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 focused on supporting Ministries of Education and corporate funders to improve learner outcomes.

# Acknowledgments

We would like to thank John Almarode, Peter DeWitt, James Nottingham, Ainsley Rose, Ray Smith, Chris Henderson and Julie Smith for their critique of early drafts of this publication.

# Introduction

Back in 1960, the great British filmmaker David Attenborough visited Tanna Island in the South Pacific to document the lives of the islanders. What he witnessed was both fascinating and bizarre.

Prior to the 1940s, the islanders had largely lived in splendid isolation, used stone-age technology and fed themselves through subsistence farming. With the onset of World War II, that era of splendid isolation evaporated, as the Pacific become a major theater of war.

Tanna became a base of operations for the Allied army and the island was quickly overrun with soldiers. For the first time, the islanders saw tinned food, chocolate, rifles, modern textiles, Jeeps and propeller-driven aircraft. The soldiers traded some of this cargo for the cooperation of the islanders.

When the Pacific war ended in September 1945, the soldiers left the island and returned to their previous lives. The islanders, of course, remained, but life without the cargo was wanting.

Soon they began to strategize how they could make the cargo return. They thought hard about the behaviors and rituals of the soldiers and built what we might call a theory of change. They analyzed their current situation, compared it to that of the soldiers, and built an explanatory framework to theorize why the soldiers had access to the delicious cargo, while they did not.

The islanders concluded that the soldiers must have had very powerful ancestors, and that it was these ancestors who bestowed the gifts, which were sent from the spirit world. Ergo, if the islanders replicated the rituals of the soldiers, the cargo would again begin to flow. They cleared runways in the jungle, made rifles and telescopes from bamboo and crafted radio headsets from coconuts. The islanders then formed into platoons and marched up and down the runways—regularly looking up in expectation that an aircraft would land, laden with precious cargo. But nothing happened.

The logic of the islanders could not be faulted, but it was built on flawed premises. Without an understanding of capitalism, mass production, aerodynamics and the internal combustion engine—they were not well

positioned to build a strong theory of change. Instead, they focused on the features of the world that they could see and assumed that everything they could see was everything there was. When the cargo didn't come, they assumed it was because they were marching wrong, or because the runway was the incorrect shape, or their salutes not high or long enough.

The Nobel Prize winning physicist Richard Feynman, during his 1974 commencement address at the California Institute of Technology, made a powerful parallel between the thought processes of the Tanna Islanders and bad science—coining the term *Cargo Cult Science*.

Richard Feynman cautioned that, to avoid falling into the same cargo cult trap as the Tanna Islanders, scientific researchers must be willing to question their own theoretical assumptions and findings, and they must be able to investigate all possible flaws in an experiment or theory. He also argued that scientists should adopt an unusually high level of honesty, especially self-honesty: much higher than the standards of everyday life.

**In this paper, we apply Feyman's cargo cult concept to education. We argue that, much like the Tanna people, we have been fooling ourselves. We examine the seductive factors that have lured us all to embrace false premises, and describe the hallmarks of the education "gold" that is worth our time and investment.**

# Education Cargo Cults: Distinguishing Between Real and Fool's Gold

# Education Cargo Cults: Distinguishing Between Real and Fool's Gold

Governments around the world collectively spend in the region of USD $140 billion each year on teaching resources, education technology, curriculum materials and teacher professional development. But we witness with despair that much of this investment is having insufficient impact. Too much is being invested in shiny things that look good but deliver little. We call these shiny things education cargo cults and our core message is that education cargo cults must die.

We have organized our thinking into seven sections, summarized below. Each is followed by a "HEALTH WARNING" comment, so you can monitor your blood pressure, take deep breaths, or fix a cup of strong tea or an iced adult beverage, as you wish. Whatever it takes to keep yourself in the range of normal.

## 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 professional learning. In section one, we argue that, despite this investment, the returns are way too low.

**HEALTH WARNING:** May make you despair at the continued inequalities in education outcomes– despite the high levels of funding.

## 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. In section two, we argue that educators and policymakers must fight hard to overcome their cognitive biases and to become true evaluators of their own impact.

**HEALTH WARNING:** May make you question how rational your decision-making really is.

## 3. How Do We Really Know?

In section three, we explore some of the key challenges with theory and evidence generation in education—including the limitations of using lesson observations and student achievement data to distinguish convincingly between education cargo cults and education gold.

**HEALTH WARNING:** May make you question how much you can trust what you see with your eyes.

## 4. Meta-analysis

Section four introduces meta-analysis as a mechanism to systematically harvest inferences from student achievement data and to build our collective understanding of what works and how to implement it.

**HEALTH WARNING:** May be the equivalent of drinking a 20 oz. Green Sweetie at your local juice bar (green apple, kale, spinach, celery, cucumber). That is, we are both strong advocates of meta-analysis: you should absorb these claims carefully.

## Conclusion

Finally, we conclude that education cargo cults must die. Our hope is that you will understand that we must worship evidence of impact.

**HEALTH WARNING:** May ultimately make you feel as though you can trust what you see again, because you'll have a framework for identifying evidence and being more skeptical of initiatives and resources that just don't have sufficient backing.

# 1

## Where is the 4% Going?

# 1. Where is the 4% Going?

Cargo cults develop when organizations or individuals spend their meager resources on the wrong things, declare success and congratulate themselves on a job well done—despite strong 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. 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, three years or more can be squandered in this manner, not to mention the money involved.

## The 4 Percent Well-Spent Could Be the Silver Bullet

Globally, the resources expended on education aren't so meager. According to the World Bank, Gross World Product (GWP), which is the sum of Gross Domestic Product (GDP) for every nation on Earth, currently exceeds USD $75 trillion per annum.[1]

Of this USD $75 trillion, approximately 4.7% is spent on education, which in hard currency is **USD $3.5 trillion per annum**.[2] 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.

[1] World Development Indicators database, 2017

[2] United Nations Educational, Scientific, and Cultural Organization ( UNESCO ) Institute for Statistics, 2013.

Figure 1: Breakdown of Education Expenditure

| No. | Expenditure Area | Estimated Percentage* | Estimated Total Global Expenditure (Rounded) |
|---|---|---|---|
| 1. | Facilities Operation, Maintenance and Build | 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 | 60% | USD $2100 Billion |
| 6. | Education Products and Resources e.g., Books; Education Technology etc. | 3% | USD $105 Billion |
| 7. | In-Service Teacher Learning | 1% | USD $35 Billion |

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 ICT], administration, transportation, teacher training, professional learning and resources, it starts to get a little murky and we need to make some careful assumptions:

**HEALTH WARNINGS:** 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 expenditure profile of developing and fragile states, whereas items 1, 2 and 5 are likely to comprise most spending.

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 in grey on the table in Figure 1). In short, for better or for worse, we are stuck with these grey costs and must make sure that the buildings,

transportation and, most importantly, teachers are primed for most effective use in their core task: educating young people.[3]

The area where there is most flex is the estimated 4% of global education budgets in the blue zone in Figure 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, somewhere in the region of USD $140 billion p.a. is spent in this category. This is both vast, greater than the combined GDP of Luxemburg and Oman, and an equally 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.[4] A well-spent 4% could be the proverbial "silver bullet" for education.

The trouble is, we're not seeing enough of those remarkable 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 (Non-completion rates 2016: England 33.1%; Australia 27%; US 16.8%).[5] The challenges in developing countries are far greater and almost too depressing to document. According to UNESCO, at least 250 million of the world's 650 million primary school children are unable to read, write or do basic mathematics.[6] 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 firsthand 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!

---

[3] It is also worth noting that when Purchasing Power Parity Adjusted (PPP) education expenditure per country is cross-tabulated to performance in the PISA assessments—there is no cast iron 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.

[4] Of course, effective implementation is also crucial and we explore this in an upcoming sister publication

[5] England data – Department for Education School and College performance tables; US data—National Center for Educational Statistics; Australia—Australian Bureau of Statistics

[6] UNESCO, Teaching and Learning: Achieving Quality for All (2014)

## 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, beneath the surface, are nothing more than education cargo cults (see Hattie's *Politics of Distraction* 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 at the weekend when we go to the supermarket for our weekly shop. 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 ones, the packaging looks nice, we recognize the brand and because there's a risk that the alternative products might be worse.

## The Marketplace: Pedagogy, Passion, and Profit

One of us gets 1-3 emails a week to endorse a new education product (book, app, resource) and, when asked two questions (Has it been deeply implemented outside of your class or schools? Have you any evidence of impact on students?), 99% 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 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 and that the product developers have evaluated their offers to the highest standard? Or that they have redeveloped their product or project logic model based on any less than glowing testimonials? Too often "they work" is assessed in terms of the author's 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 they have more marketing finesse. Most of these are education cargo cults, and they should be stopped so 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 at all. 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 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 creates the cargo cult. These are the ships 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

# 2. A Glitch in the Matrix

The Tanna Islanders believed in their cargo cult, not because THEY were mentally deficient, but because we are ALL mentally deficient. If any one of us were completely rational then the world would be a strange buzzing confusion—especially when those we meet often act irrationally. We all develop belief systems to survive in our busy, buzzing world; sometimes developing beliefs that "get us through" and some of these rules of thumb or heuristics allow us to unconsciously traverse all manner of situations. But sometimes these beliefs are contrary to good practice.

The research supporting this comes largely from behavioral economics and particularly from the work of Amos Tversky, Daniel Kahneman, Herbert Simon, Richard Thaler and Cass Sunstein (see bibliography 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 post-rationalize bad choices after they were made (often referred to as *cognitive dissonance*).

## Cognitive Bias: If Only We Were Impartial as Judges

During the last 40 years, a growing database of cognitive biases, or glitches in our human operating system, have been catalogued and confirmed through laboratory experiment and psychometric testing.

The research suggests that biases afflict all of us, unless we have been trained to ward against them. More than 80 cognitive biases have been recorded by behavioral economists.

In the table below, we summarize some of the inherent biases that, if left unchecked, can result in education cargo cults—that is, unrestrained intuition over reason that drives us all to pursue products and practices with insufficient scrutiny. 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 this precious 4%.

| Cognitive Bias Category | Description | References |
|---|---|---|
| **Authority Bias** | 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.<br><br>**EDUCATION:** Don't be swayed by famous titled gurus. Carefully unpick and test of all their assumptions–especially if they are making claims outside the specific area of expertise. Be particularly suspicious of anyone that writes and publishes a white paper [!!!]. | Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology, 67* (4), 371 - 378. |
| **Confirmation Bias**<br><br>**Post-Purchase Rationalization**<br><br>**Choice-Support Bias** | The tendency to collect and interpret information in a way that conforms with, rather than opposes, our existing beliefs.<br><br>And when information is presented which contradicts current beliefs this can transition into Belief Perseverance i.e. where individuals hold beliefs, that are utterly at odds with the data.<br><br>**EDUCATION:** We will tend to select education approaches, products and services that accord with our worldview and 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. (1998, June). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2* (2), 175 - 220. |

| Cognitive Bias Category | Description | References |
|---|---|---|
| **Observer Expectancy Effect**<br><br>**Observer Effect**<br><br>**Hawthorne Effect**<br><br>**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.<br><br>**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.<br><br>**EDUCATION:** Understanding 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. |
| **Anecdotal Fallacy** | The tendency of take anecdotal information at face value and giving it the same status as more rigorous data in making judgments about effectiveness.<br><br>**EDUCATION:** Do not take spurious claims about impact at face value and do not invest in training based on participant satisfaction testimonials alone. Beware of testimonial porn. | 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** | 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.<br><br>**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 that they are also world leading in area B. What's good for the goose isn't always good for the gander. | Nisbett, R. & Timothy, D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35* (4), 250 – 56. |

| Cognitive Bias Category | Description | References |
|---|---|---|
| **Not Invented Here** | Tendency of avoiding using a tried-and-tested product because it was invented elsewhere—typically claiming "but we are different here."<br><br>**EDUCATION:** Be open to using and adapting existing IP. Avoid reinventing the educational wheel—unless you work in terrain where wheels are useless [you probably don't]. | |
| **Ikea Effect** | Tendency to have greater buy-in to a solution where the end-user is directly involved in building or localizing the product.<br><br>**EDUCATION:** Make the effort to localize and adapt tested solutions. This will generate greater emotional buy-in than standardized deployment. | Norton, M., Mochon, D., & Ariely, Dan (2011). The IKEA effect: When labor leads to love.. *Journal of Consumer Psychology, 22*, 453 – 460. |
| **Bandwagon Effect**<br><br>**Illusory Truth Effect**<br><br>**Mere Exposure Effect** | Tendency to believe that something works because a large number of other people believe it works.<br><br>**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. |
| **Clustering Illusion**<br><br>**Cherry Picking** | 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).<br><br>**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 cherrypicked? | Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*, 295 – 314. |
| **Conservativism** | The tendency to revise ones' beliefs insufficiently when presented with information that contradicts our current beliefs.<br><br>**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 evidence claims. | Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* New York, NY: Cambridge University Press. |

| Cognitive Bias Category | Description | References |
|---|---|---|
| **Courtesy Bias** | The tendency to give an opinion that is more socially palatable than our true beliefs.<br><br>**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. | |
| **Law of the Instrument** | If you have a hammer, everything looks like a nail.<br><br>**EDUCATION:** Start with the problem or 'wicked issue' you are trying to solve and then work backwards to instruments– rather than searching for nails to bang. | Maslow, A. (1996). *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.<br><br>**EDUCATION:** Don't be a busy fool. 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.* New York, NY: Penguin. |
| **Sunk Cost Fallacy** | 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.<br><br>**EDUCATION:** Review implementation of new approaches regularly and set clear parameters/hurdles that must be achieved for the project to stay live.<br>Ruthlessly prune anything that does not pass the hurdle test. | Arkes, H., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Process, 35*, 124 – 140. |

For educators to overcome these cognitive biases and fallacies, they need to develop their logical-rational skills without losing their passion for teaching. Educators need not act like Tanna Islanders. This requires the development of mind frames 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?

# 3 How Do We Really Know?

# 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 judgements and build theories. And, from theories, we can build structured processes—continually testing and refining these too.

## Towards 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:

- **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** – similar to Useful theories, but without any significant experimental support (i.e. seem to make sense, but more evidence needed)
- **Misguided Theories** – those without experimental support and/or 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 table on the next page, we have tentatively added to Penrose's original list; the items in bold are our additional suggestions.

| Superb Theories | Useful Theories |
|---|---|
| Newtonian Mechanics<br>Einstein's Special Relativity<br>Einstein's General Relativity<br>Quantum Theory | Quark Model<br>Big Bang Theory<br>Darwin's Theory of Evolution<br>**Tectonics**<br>**Weather Systems** |
| **Tentative Theories** | **Misguided Theories** |
| String Theory<br>Super Gravity<br>Grand Unified Theory | **Cold Fusion**<br>**Flat Earth Theory**<br>**Lamarckism**<br>**Æther**<br>**Alchemy**<br>**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 would sit firmly in tentative. 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.

## The Scarecrow Was Only Partially Right

"I could while away the hours/conferrin' with the flowers/consulting with the rain…if I only had a brain," sang Ray Bolger in the classic movie *The Wizard of Oz*. But, as it turns out, the affable Scarecrow wouldn't have solved every problem with a noggin, because the brain is, in some respects, getting in the way of education—and certainly getting in the way of developing a superb theory of education/learning.

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

Through Functional Magnetic Resonance Imaging (fMRI), we can map and monitor the flow of blood in the brain and, through Electroencephalography (ECG), we can measure voltage fluctuations resulting from ionic current within the neurons. These tools are the equivalent of holding a lit match in a large cave: they illuminate some things, but far from everything.

We are still a very long way from being able to see individual neurons firing and wiring and estimates for this accomplishment range from 2030 to never (Bostrom, 2014). We are further still from a theory of consciousness or cross-brain processing that is taken seriously in the scientific community (See Dennett, 1991; Chalmers, 1996; and Searle 1997 for the most plausible accounts that still remain current).

At the bleeding-edge of technology, neuroscientists are exploring the possibility of creating flotillas of nano-scale robots that can swim freely within the brain and attach themselves, like fridge magnets, to the membrane of a neuron or a synapse (Shanahan, 2015). These nanobots would sit at their respective junctions, intercepting the signals of the brain and broadcasting them to the outside world where they could be analyzed by neuroscientists and educationalists. But, right now, this is little more than a thought experiment.

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:

- **Lesson Observation:** watching and listening to the interactions 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 this 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 on some of the inherent challenges with each of these approaches in helping us to sort education cargo cults from educational gold.

# LESSON OBSERVATION

## The Limits of Lesson Observation

In many education systems, it is a mandatory requirement that every teacher undergoes at least an annual observation by their 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 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, with teachers peer-reviewing each other's lessons so they can offer one another advice and harvest good practice to apply back in their own classrooms. Finally, they can be used to sift education cargo cults 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 a teacher in a class of students? Do we watch teacher performance or do we watch the impact on the students? What if the performance is spectacular, but the impact of little consequence?

If we phrase the question as a binary yes/no choice, then the answer to whether we can make meaningful and rigorous observations is a resounding *yes*. And, by binary, we mean questions where there is a clear yes/no answer, like:

- Is the teacher in the classroom?
- Are they 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 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 the performance, we can see, hear

and sniff reasonably reliably. We could probably stretch from binary to asking questions about frequency—how often something occured (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 and ask:

- 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?

We open a huge can of worms. Who decides what "good" is and who decides what "impact" means?

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 and heart rate analyses to tell them whether a patient is fit and well. And, generally, these work relatively well, but 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. When we see busy, engaged students in a calm and ordered classroom where some students have supplied the correct answers and we conclude that a heck of a lot of learning is going on, 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 [in theory 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 that is 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 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 with us telling the teacher 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 ink blot images that psychiatrists show to their patients—where some say they can see their mother and others JFK.

The image below, popularized by the philosopher Ludwig Wittgenstein (1953), provides a similar conundrum. When we undertake lesson observations, do we see a waddling duck or do we see a bounding rabbit? The data is the same, but we can interpret and re-interpret it in more than one way.

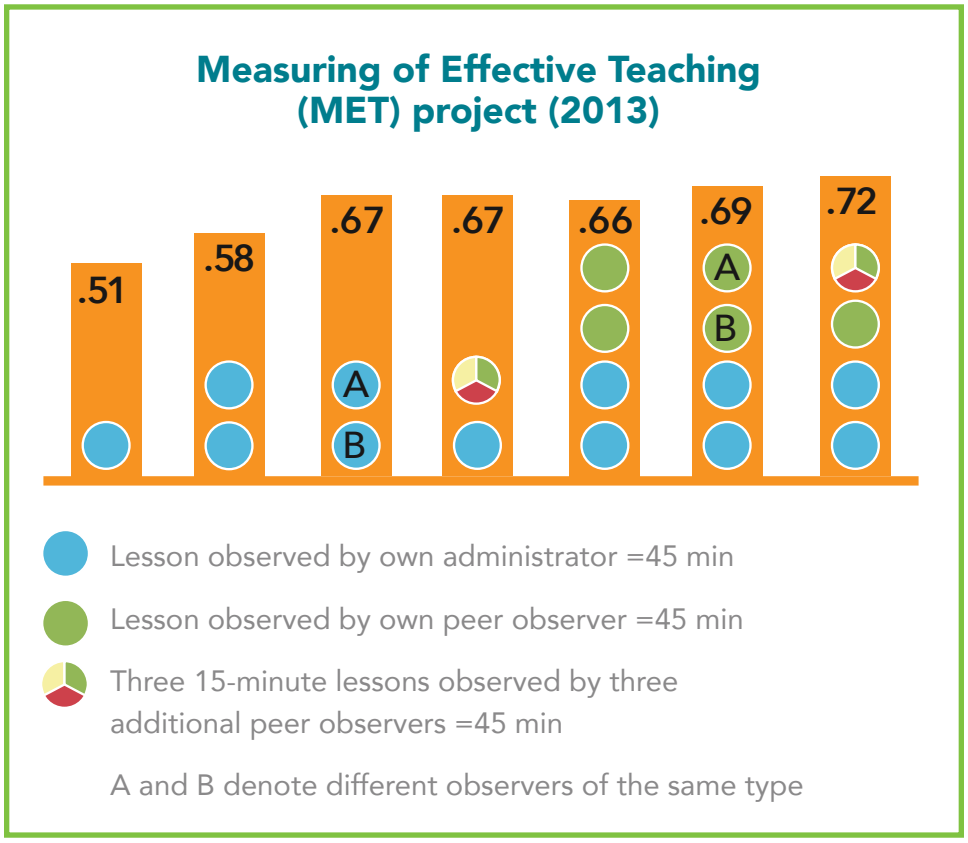
## A RABBIT OR A DUCK?

There has been quite a lot of research into the duck-rabbit problem of lesson observation in the last few years. One of the strongest datasets comes from the Measuring of Effective Teaching (MET) project (2013), which was funded by the Gates Foundation.

The MET project 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. On the graph below, you can see different combinations of observations by different numbers of individuals for different durations of observations. Let's cut to the chase: on the far right-hand side of Figure 2, we can see that, where a teacher undergoes six separate observations by five separate observers, there is "only" a 72% chance there is agreement, thus 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.

Figure 2



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—which is another way of describing the duck-rabbit conundrum.

### 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 teacher effectiveness) is riddled with many of the cognitive biases that we described in section two.

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 have started to automate the coding of 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 teachers 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 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 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 et al, 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 [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, stressful 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, like a foreign language, driving a car or how to "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 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 and 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" doesn't necessarily lead to continually deepening and 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 to keep the fires of enthusiasm and experimentation burning. Lesson studies focused on the impact of the lesson on the students also can help the 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 are cargo cults and which are gold, lesson observations alone will not give us a robust answer about what works best in education reform, because we can't see everything with our eyes. There is no such thing as immaculate perception (Nietzsche, 1891).

# ASSESSMENT

## Assessment: Measuring the Teaching & Learning Journey

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 and, 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 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. SATs, A-Levels, the International Baccalaureate 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, science and about the depth and breadth of knowledge they have 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, financial security and 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; TIMSS; PIRLS], 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 distinguish educational gold from education cargo cults and to validate the former over the latter.

In this context, high-stakes assessments are blunt instruments—akin to piloting your boat by the stars on a cloudy night, rather than GPS. We can infer something about which schools are higher and lower performers, but need to carefully tease out background variables like the starting points and circumstances of the learners, the multiple other important outcomes, so that we can measure 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 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 and programs] are more/less effective. 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 assessments, 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 assessments 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 or 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 tease out the education gold is by collecting it (formatively) at the beginning, middle and (summatively) end of the journey to systematically measure distance travelled by individual students and groups of learners and by experimentally varying very narrow elements of teacher practice to 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 Assessment

Education researchers often make the distinction between summative and formative assessment. The typical distinction is:

- **Summative Assessment** – where the purpose is to compare student performance at the end of a program or study against an agreed benchmark or standard. The outcomes of this assessment can often count in full or in part to a formal qualification and is often used as an external accountability driver.
- **Formative Assessment** – which is a range of formal and informal feedback mechanisms conducted by teachers during the learning process to gather information about the effectiveness of the learning episode and where to go next.

However, note that Michael Scriven (1967), who invented these notions, never used these terms; he introduced them as formative and summative *evaluation*. He claims, as do we, that any assessment can be interpreted formatively (during) or summatively (at the end). The distinction is WHEN—not the nature of the assessments. Any test could be so interpreted; it is the evaluative focus, not the test, that distinguished formative from summative.

Too often, teachers see summative as "bad" and formative as "good" when this is nonsense, and some see summative as needing to be highly reliable, but formative less so. But, 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 an education cargo cult and what's education gold, but we want to stress that the difference between the two is more often about what it's used for than the mechanism of data collection itself. That is, the same assessment instrument can be used both formatively and summatively. As educator 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 looking like a cargo cult 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 this data, we face the challenge of vacuuming it up from disparate sources and in drawing the common threads to build a compelling narrative about what's a cargo cult and 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.

# 4 Meta-Analysis

# 4. Meta-Analysis

Before we set out our stall and 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-analysis, which in 2009 was crystalized into *Visible Learning: A Synthesis of over 800 Meta-Analysis Relating to Achievement*. 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 & Bulmer, 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 to decide for yourselves.

Thus far, we have outlined some of the challenges involved in using lesson observation and student achievement data to firmly distinguish between education cargo cults and educational gold. Now we take you on the tour of meta-analysis.

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

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) which provides a measure of the magnitude or size of the effect. We can also ask about whether the effect is statistically significant and different from zero (no change or no effect and we know this is valuable and also majorly affected by the size of the sample), and we can ask about the magnitude or size of the effect. We need to ask both if the outcome is different from chance and what the size of the effect is.

## Effect Size: Does it Pack the Punch of a T-Rex, a Boxer, or a Cat?

Many of the quantitative studies, as a matter of course, have sufficient data to calculate an effect size. Rather than telling you whether something works or not, it quantifies on a universal scale how powerful (or how weak) the intervention is. In other words, if something works, does it pack the punch of a T-Rex, a heavy-weight boxer, of us, or a small cat?

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 post-intervention with a single group or the means from an experimental and control group.

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

$$d = \frac{\overline{x}_1 - \overline{x}_2}{SD}$$

In plain English, this 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 the pooled standard deviation (SD) for the dataset. The output of this calculation is a numerical value which shows the gain or decline in performance from the intervention compared to a control group as a proportion of a standard deviation. So, an effect size of 0.20 means that the second basket of scores were 20% of one standard deviation higher than the first basket of scores. Yes, there are two methods (pre-post and intervention comparison) and they can lead to different interpretations—in our work, we check to see if there are meaningful differences between them before we make interpretative comparisons.

The beauty of the effect size statistic is that it is a form of universal translator. No matter what testing instrument the researcher uses and no matter how the scoring is done, so long as there are sufficient numerical outputs and at least two sets of scores (means and SD), it's possible to calculate the effect size—then, they are comparable in many ways independent of the sample, the measure and the context.

We should, of course, worry about the quality of the instrument, and reliability and information in the scores and scoring, and focus very much on the most defensible interpretation of the effect size.

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 this data is extremely hard. Imagine that you visit a giant book warehouse in search of the various works of and about William Shakespeare. When you open the door and peek inside, you are surprised to find that all the racks are empty and that the books are randomly piled on the floor neck-deep. Undeterred, you drive straight in and begin the mammoth task of sifting the tens of thousands of randomly assorted books. After a fashion, you sift through the cookery books, political biographies, murder mysteries and start to get close to what you are looking for: literary criticism related to Shakespeare; then a yellowing copy of *Romeo and Juliet*; and, finally, you begin to track down copies of all his other works, including the sonnets and songs. You then sift through the 300 Shakespeare-related works that you have harvested, throwing the rest back on the pile.

To collect, sift and sort the 90,000 plus education research studies that include effect size data requires a similar 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. (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 these mavens with 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.

## Building Gleaming Glass Cities with Education Data

Glass also showed how it is possible to see if there are critical variables that may affect any overall conclusion—in the jargon, these are *moderators* (e.g., do the overall findings apply similarly to 4 year and 20-year-old, to math and music, to US and Australian students, etc.). Glass said that researchers throw stones onto a pile, whereas meta-analysts take those stones and build houses.

The work of these educational mavens is crucial in providing a meaningful pathway through the research. To date, there have been more than 1,500 separate meta-analyses undertaken in different categories of education intervention relating to achievement outcomes (and many more with other outcomes). In this case, the mavens visit the "warehouse" to sift and sort the works on things like assessment for learning programs, thematic curriculums, homework, etc., rather than the collected works of Shakespeare. To extend Glass' claims, synthesizing many meta-analyses is like building a city. To gild the lily of the metaphor, synthesizing brings both clarity and structure and potential future function to millions of tiny "sand grains" of data.

*Visible Learning* (2009) represented an early (but not the first) attempt at mega-mavenry—that is, the collection and collation of all the educational meta-analyses into a single meta-meta-analysis. We believe that both the meta- and meta-meta-analyses (or first- and second-order meta-analysis) provide a useful compass to educators and policymakers in distinguishing education cargo cults from educational gold. And we believe that Visible Learning represents the nearest thing we have in education that might aim to fit Sir Roger Penrose's category of a useful theory—albeit that it would likely hover towards the top of the tentative category. We accept that no current approach in education (Visible Learning included) is ever likely to rank in Penrose's (1989) superb category, but the key here is that it is the INTERPRETATION and STORY that help explain the findings that is a contender for a useful theory; not the data.

## A Q & A to Explore Common Challenges

We also recognize that the meta-analysis approach, like all other educational research methods, is not free from challenge or criticism. In a sister publication, we will explore these challenges in detail.

Some of the more common challenges with the meta-analysis approach are:

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 widely from study to study.

   **Response:** If there is little or modest dispersion across all the effect-size data for a particular influence, then we can have a higher level of confidence in the synthesis of the research findings. But, if there is substantial dispersion, then the search for moderators is undertaken in earnest. This search was discussed in more detail in *Visible Learning* (2009), and many critics have ignored this search. Just because it was hard to find many moderators (there were some for some influences) does not mean the search was not undertaken, or that we should not continue to search for moderators. BUT we also need to increase the resolution or dots per square inch in our analysis.

2. **The File Drawer Problem.** This is the argument that education researchers are only likely to publish data which shows positive findings and that, because of this, the meta-analyses are likely to be presenting a high proportion of false positives.

   **Response:** We agree. This is one of the reasons that Visible Learning sets the effect size bar so high (i.e. $d > 0.40$). This helps 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 highest probability of impact. There are also statistics for estimating the number of papers still stuck on 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 to 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 is Western-centric and some of it is 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 that it cannot be applied with confidence to other contexts.

   **Response:** Again, we agree. All reviewing of literature is rear-view mirror research (recall research means re-search; searching again), but try driving forward ignoring the rear-view mirror. Ouch. The research can be used with much greater confidence to distinguish between cargo cult approaches and educational gold in developed country contexts. 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 which crowd sources coordinated data collection through affiliated Ministries of Education or research institutes—particularly in developing country contexts.

4. **How do we implement.** Here the quite reasonable argument is that, whilst meta-analyses provide a useful overview at 40,000 ft. about what works and what is a cargo cult, they become much less valuable at 5,000 ft., let alone 5 ft.

   **Response:** Yet again, we agree. 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 map 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, but that there

is a lack of evidence about effective implementation of this evidence (we are soon to release a sister publication on this).

5. **Meta-analyses are a very reactive research approach.** This is the argument that the mavens 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.

   **Response:** Again, yes. We really need that globally coordinated approach to the coordination and collection of education research that was alluded to under the third point. Meta-analyses come in various hues of quality. This is also the case for meta-analyses and original studies they are based on. 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).

There are many other criticisms of meta-analysis and syntheses of meta-analysis and we will return to this in a future white paper. This (and the future) paper will not argue that these issues should mean that the methods should not be used; but, when used, these issues need attention. Too many critics cite criticisms and then falsely assume that they have not been acknowledged or investigated in the Visible Learning work.

We also note the major message from Visible Learning is "Know thy impact." Knowing one's impact not only begs the moral purpose questions about what you mean by impact, but also means the focus needs to be on the impact you are having on your students, which students are exhibiting this impact and whether this impact is sufficient for the investment—hopefully, after implementing the high-probability interventions identified in the Visible Learning research. The message is about looking forward (i.e. out of the driver's window), while taking account of the rear-view mirror perspective.

## Towards a Global What-Works Repository

We are driven by the desire to give teachers, school leaders and policymakers good quality tools to distinguish education cargo cults 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/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 through this data with the rigor and tenacity required. There just aren't enough hours in the day. Hence the tendency to fall back on our heuristics/cognitive biases/hunches when making decisions about what works in the classroom.

We accept that meta-analyses or the interpretations from meta-analyses are not foolproof either. After all, they aggregate data from achievement tests of various types only. We note a German team aggregating data from motivational and emotion studies for a yet-to-be-published study. And one of us recently completed a meta-synthesis on how we learn (Hattie & Donoghue, 2016)—and another is needed on health and physical outcomes, and another on willingness to reinvest in coming to school and so on. But, as noted above, various measures of the quality of the original data and the quality of the meta-analysis can help address whether quality studies lead to similar results as lesser quality studies.

# Conclusion

# Conclusion

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 provide 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 on educational resources, technology and teacher professional learning each year. 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 their 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 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 have established the Visible Learning global network as the definitive vehicle for widely disseminating what works best in education and how to implement it to great effect. The Visible Learning research was first released in 2009 with 800 meta-analyses, 150 effects, and six domains of influence. Today, it represents more than 1440 meta-analyses, 250 effects, seven domains of influence, and 27 subdomains. But the research and evaluation of evidence continue, as does implementation and continuous improvement through the Visible Learning$^{plus}$ school change model of professional learning. The power of this research, which we will make readily accessible on an ongoing basis, lies in helping policymakers and educators understand which factors have the highest impact on student achievement so they can make strategic decisions based on evidence to maximize their time, energy and resources.

Education cargo cults must die. Instead, we must privilege evidence of impact and we must use this evidence to ensure that every learner gets at least a year's growth for a year's input.

## The Role of Worship: The Prince Philip Movement

The oral history of Yaohnanen people of Tanna Island contains a tale about the son of a mountain spirit who travelled to a distant place. Once there, he was said to have married a powerful woman but it was prophesied that he would one day return home.

In 1974, Queen Elizabeth II and her husband Prince Philip made an official visit to Tanna Island. The Yaohnanen people, seeing the level of respect accorded to Queen Elizabeth by the various government officials, concluded that Prince Philip must be the son of the mountain spirit referred to in their legends. The Cult of Prince Philip was born.

For us, there is no harm in worship—as long as we bow down to the right things. The right kind of cult worships evidence on the impact of learning on students. Evidence is the rich jam at the heart of the whole education enterprise: it is to be relished and spread far and wide.

# Bibliography


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

Bostram, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford. UK: Oxford University Press.

Cantrell, S. & Kane, T. J. (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]. Bill & Melinda Gates Foundation.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory.* New York, NY: Oxford University Press.

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.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112* (1), 155 – 159.

Dennett, S. (1991). *Consciousness explained.* Boston, MA: Little, Brown and Company.

Feynman, R. (1985). "*Surely you're joking, Mr. Feynman": Adventures of a curious character*. New York, NY: W.W. Norton.

Feynman, R. (1974). *Cargo cult science: Some remarks on science, pseudoscience, and learning how to not fool yourself,* [Caltech's 1974 commencement address]. Retrieved from http://calteches.library.caltech.edu/51/2/CargoCult.htm

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

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*, 295–314.

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

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

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

Hattie, J. A. C. & Donoghue, G. M. (2016). Learning strategies: A synthesis and conceptual model. *Nature Partner Journals Science of Learning,* Review Article Number 16013.

Hattie, J. A. C. & Yates, G. (2014). *Visible learning and the science of how we learn.* Oxford, UK: Routledge.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement.* Oxford, UK: Routledge.

Hattie, J. A. C. (2012). *Visible learning for teachers: Maximizing impact on achievement.* Oxford, UK: Routledge.

Hattie, J. A. C. (2015). *What does work in education: The politics of collaborative action. Open Ideas at Pearson*.
Retrieved from https://www.pearson.com/hattie/distractions.html

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

Hattie, J. A. C., Masters, D., & Birch, K. (2016). *Visible Learning into action.* Oxford, UK: Routledge.

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. (1996). *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, Vic.: 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. *The Journal of Abnormal and Social Psychology, 67* (4), 371 – 378.

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

Nisbett, R. & Timothy, D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35* (4), 250 – 56.

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

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

Ofsted (2009). *Twenty outstanding primary schools excelling against the odds.* London: Crown Copyright.

Parkinson, C. N. (1958). *Parkinson's law: Or the pursuit of progress.* New York, NY: Penguin.

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

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

Sawilowsky, S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods, 8* (2), 467–474.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39-83). Chicago, IL: Rand-McNally.

Searle, J., Dennett D., & Chalmers, D. (1997). *The mystery of consciousness.* New York, NY: New York Review of Books.

Shanahan, M. (2014). *The technological singularity.* Boston, MA: MIT Press.

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

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

Wiliam, D. (2014). *Why teaching will never be a research-based profession and why that's a good thing.* Retrieved from http://www.dylanwiliam.org/Dylan_Wiliams_website/Presentations.html

Wittgenstein, L. (2009). *Philosophical investigations* (4th ed.). Hoboken, NJ: Wiley-Blackwell.

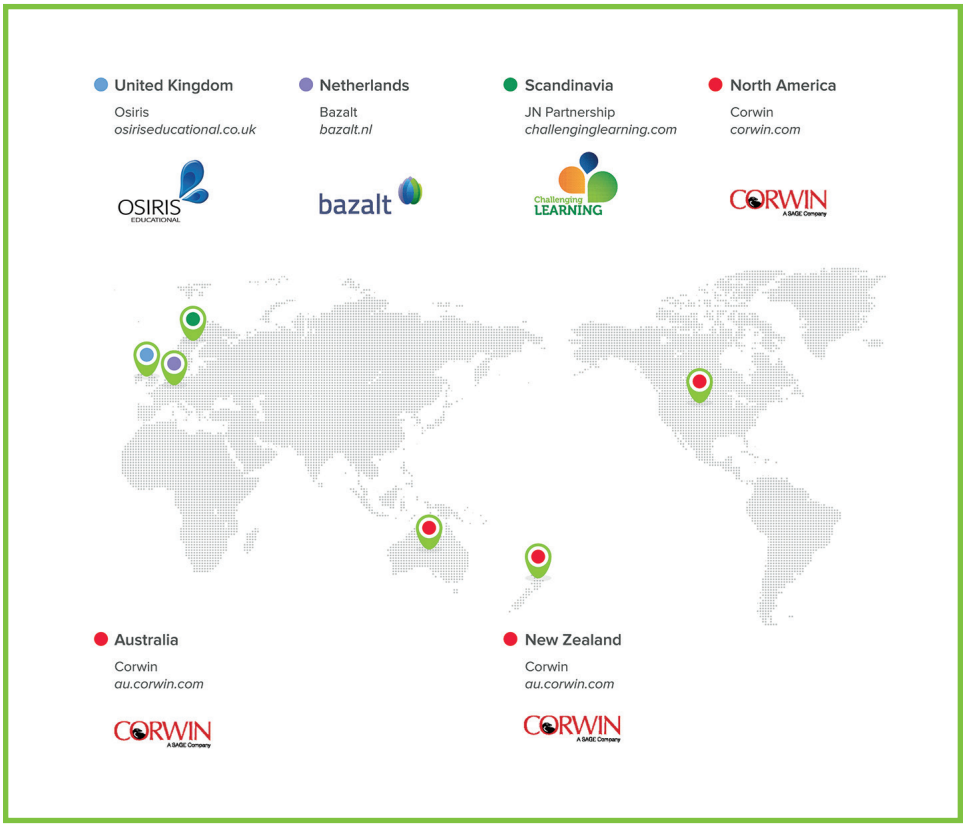
To learn more and get involved in the

# Visible Learning Global Network

## Contact Us:

**www.visiblelearningplus.com** | **www.corwin.com**

Email: **visiblelearning@corwin.com** | Twitter: **@VisibleLearning**



CORWIN
A SAGE Publishing Company