The Problem Is Far Worse

By Terry Rosen © 2020

Abstract

This paper examines various relationships between intelligence, education, IQ, behaviorism, basic reasoning, enumerative study vs. analytic study, continuous improvement, and fundamentals of motivational research and the combined effects they incur on youth in American schools. It further implicates corporate and business practices described by W. Edwards Deming in the 1980's.

This paper does not provide solutions to the problems described. It's my view that people do not take the research or the risks seriously, and until they do, descriptive solutions are of no use.

The Problem Is Far Worse

I've been teaching roughly 20 years. I've been learning about Continuous Improvement for roughly 30 years, so I've always had 'improvement' in mind while pursuing my teaching practice. My journey in quality included both Deming and Kohn, as well as Imai, (Kaizen). My master's thesis attempted to describe iterative improvements in the technology classroom, as well as approachable methods for improving teaching practices within a classroom.

Of course, though, the biggest faults in education are institutional in nature, just as they are in business. Through my continued and varied readings I've accumulated background knowledge in psychology(esp. Jaynes), education(esp. Kohn), evolution, (Gould), child-rearing (Kohn and Fay), intellectual thought (DesCartes, C. I. Lewis, basic logic, intermediate statistics), and a host of Applied Technology subjects.

The First Multiple Choice Test

A few years ago, an idea melded in my mind combining features from each of these fields. I began to share the basic concept and found it well received. I started by searching for the first formal use of the multiple choice test.

I routinely ask people when they think this first occurred. Please take a moment to pause and consider the question for yourself. I'll admit up front that I was way off in my guess (which I wrote down before Googling the question). Everyone has been off, usually by some huge amount, except one person who was within 20 years. How far off we are does not matter, but the fact that we are off is material to the discussion.

I came to the conclusion that the IQ test introduced in 1916 was the first use. (Previous versions of the IQ test may have been multifaceted, including features beyond a multiple choice test).

None of the participants in my very informal survey have ever guessed later than 1900. My own guess was the late 1700s to early 1800s. The earliest guess I've received was approximately 800 B.C.

The correlation of the IQ test with the first use of multiple choice testing is both interesting and disturbing.

Enumeration vs. Analysis

When Deming wrote about education in The New Economics in 1986, he cited his own article, On Probability as a Basis for Action. With this article, he essentially proves that counting cannot reveal cause. My interpretive piece based on his article is that ‘no matter how accurately we count the people in Pittsburgh, we can’t say **anything at all** about ‘why’ those people are in Pittsburgh. ’ The idea is much deeper, of course. It forms, I believe, a rule of thought. You cannot count something and make a claim to cause based on your count. This is the fundamental basis of correlation as causation, as a fallacy. I take it further.

It’s common for people to observe an event a single time and then believe they know the cause. It’s also quite common to share the experience with another person, who then, without any actual information, forms the same belief.

A friend of mine told me they got the flu vaccine and then got the flu, and so vowed that they would never again get the vaccine.

1. First, they misunderstand the point of the vaccine. Those who are vaccinated are less likely to get the flu. We get the vaccine in order to reduce the chances of getting the flu.
2. Next, it’s very likely that they had the symptoms of the flu (due to the vaccine) but were not actually sick with the flu. In this case, getting the flu is not the same as getting flu symptoms.
3. They’ve made a decision, based on a single instance (which they misunderstand), to risk their health for the rest of their lives, based on a single event.
4. They then share their experience of this vaccine /flu they got in order to encourage others to avoid the vaccine.

Less serious examples can be seen literally every day. But how does this apply to our question?

We cannot say, by counting (a single item), **anything at all** about what caused the thing. The above example shows that counting a single item and erroneously believing we know the cause is as common as it is incorrect. I call this an enumerative study with a sample of 1.

But the problem is much worse. People commonly ‘rationalize’ a plausible cause and effect relationship, then believe it, and then share it with others, with ZERO evidence. Conspiracy theorists often arrive at such conclusions. It’s as if any and every idea we can form in our mind

provides proof of a conspiracy, even with zero substantiation. Once we form it, we can share it to convince others. I call this an enumerative study of zero.

Even worse, we can form an idea in our heads that we think will convince others, even though we know as a matter of fact that the idea is false. I refer to this as an enumerative study of -1.

We cannot witness a person's act and judge the person (or the nature of the person) as the proximal cause of that act. Yet, when a student scores badly on a test, we judge the student by assigning a grade to that work, a mark which the student carries with them through their life.

Supposed Racial Differences

Nowhere is this more pronounced than in the judgment of race by white academics and scientists for the past 300 years. In Mismeasure of Man, Gould unravels, with sometimes painful historical accuracy, the horrific misattribution of traits to the races through an amalgam of continued misuse of counting and infantile use of statistics. He further displays both the avarice of some and the ignorance of reason held by others as they 'prove' racial differentiation via techniques such as measuring the size of the brain. The reasoning is bereft of validity, even though the basics of reasoning were established thousands of years earlier. Then further, these men without a background of understanding their own unconscious cognitive biases actually believe what they think they've discovered, 'that brain size implies level of intellect'.

The Final Chapter

For discussion here is the evident misuse of IQ described in the final chapter of Gould's book. The benevolent intention of the invention of the IQ metric was quickly supplanted by the appropriation and misuse of the tool to judge people. Its use in this vile endeavor is so well illustrated by Gould that on my first read I had to put it down and walk away from it for a few days before finishing the read, (because it was making me feel sick to my stomach). But, at its foundation, the assumption held in regard to the IQ test is that it accurately measures a person's innate level of intellectual potential. Further, it is believed that the potential that is measured is 'meaningful' and ultimately 'predictive'.

For clarity, IQ tests claim to establish the CAUSE of the score and that the CAUSE is the person's innate intelligence inherited from their parents.

Deming's proof destroys this presumptive conclusion without any consideration of the rationale required. One cannot count a thing and then judge the cause of that count.

Applied to Testing

We can then easily and necessarily apply the concept to standardized testing. The same fallacy applies.

This is done in schools by way of the MAP tests (or CMAS and others). The test is given 3 times in one year. It thus purports to measure change during 3 time periods, which interestingly, can be used to measure loss of skill over a Summer break (or just lack of skill development during that period).

No. It cannot do any such thing. Why not?

Because it is MERELY an enumerative study. Repeating an enumerative study does not make it less enumerative. Does it even measure enumerative change, regardless of causality? To believe that it does requires us to believe that every student is motivated to perform in the same way and for the same reasons. Which is impossible.

It has no claim to the stability of the process. Via Scholtes we know that test scores are created via various, immeasurable causes. So attributing the score itself to the ability of the student, as an accurate measure of any kind, is so ludicrous that it seems difficult to imagine that anyone believes in it at all. But basically, in what universe is a classroom with randomly varying skill levels and socioeconomic roots to be considered a ‘controlled experiment’?

(A section of the original text has been removed. It was an application of Kahneman’s research of Israeli judges, which was later debunked.)

PDCA - 100 Years Later

In 2019, I was blessed to be involved in the re-creation of Construction Trades in Boulder, Colorado. I knew about trades training but was introduced to it more in-depth. In particular, I’m speaking of apprenticeship, especially that of the plumbing and electrical trades.

I’d spent years teaching the design cycle, a direct outgrowth of PDCA, the four-step process that Deming added to Shewhart’s 3-step process. Shewhart’s process derived from the scientific method, hypothesis, experiment, and results. Shewhart renamed this in convention to specification, production, inspection. Shewhart took this process, which was originally linear, and made it iterative. Keep iterating, and the process/quality will keep improving.

It was no surprise then to relate Deming’s PDCA to the process of apprenticeship. The journeyman PLANs what to DO, demonstrates it even, the apprentice does it, the journeyman CHECKs it, the journeyman ACTs on his analysis.

Then, if the apprentice mastered technique, the journeyman moves the apprentice on to the next technique. But ‘if not’, what happens? The journeyman iterates the training. He performs

the cycle, again and again, WITHOUT JUDGEMENT, until the skill is mastered. Here lies the crux. Two cruxes, if you will:

First, an unmastered skill is not neglected.

Second, it is not 'judged'.

The retraining (remediation) is a matter of fact, indeed, expected. Failure is expected, reinforced even. Yes, failure is reinforced as evidence of commitment to improve.

We know, from 10 to 20 years of educational experience (as a student), that failing a quiz does not mean going back and mastering the content so we can pass it. Failing a quiz means our chances of getting a good grade have just diminished. We may have had teachers that remediated, or we may not have. But we have, in essence, a PDCA cycle with the A removed. The teacher plans the content P, delivers it D, the students take an assessment, and the teacher grades it C (I consider this counting, not analysis). The class moves on.

At first glance, it is obvious that 25% of the learning process has been lost. But clearly this is false if we observe the apprenticeship model where the apprentice may do badly the first time, slightly better the second time, and may take 5 or even 10 iterations to reach mastery. Let's observe the iterative loss of just four cycles:

PDCA vs. PDC
PDCA
PDCA
PDCA

Sixteen pieces vs. 3 pieces. This process has eliminated more than 80% of the learning opportunities. This system is literally designed to **create defects** and avoid mastery.

Let's revisit our students who have not done well in their navigation of the system and find themselves in the tenth grade, and with near zero academic success. Now, suppose we give them a MAP test? How motivated are these individuals to complete such a test? Or any test? Or to even be at school at all?

A Hypothetical Test

I like to pose the following hypothetical question to adults (generally other teachers). Presume we have a 100 question, multiple choice test. Each question has four possible answers, only one of which is correct. A student scores 70% on the test. How much of the content of the test did the student understand?

Except for math teachers, (typically math teachers who know me personally), almost all adults will say 70%. I then explain that no the answer is 60%. The students got 60% correct and guessed at the other 40, and got ¼ of them right by chance, or, 60+10=70.

The uncomfortable truth is that teachers believe' that the scores students receive are attributable to only the student (generally work effort, definitely not randomness). Parents, too, believe that the grade represents the student's effort, or in this case, lack of effort. EVEN the student likely believes it's his/her lack of effort. How can they know any better?

Where Did My Intrinsic Motivation Go?

All this plays into a situation with declining **intrinsic** motivation and dependence on **extrinsic** motivation. Kohn provides copious and inarguable research evidence about the effects of grades and gold stars on the reduction or elimination of intrinsic motivation. Why does this happen? Deming mentions this fact in The New Economics. Indeed, intrinsic motivation in the workplace supports many of Deming's views on leadership and supervision.

[Bizarrely, Michael Wu, the de facto king of gamification, has published an exquisite set of explanations in glorious detail about what is or is not extrinsic or intrinsic motivation, but has **no understanding at all** (as of c. 2014) of the risks of overusing extrinsic motivation.]

I believe the apprenticeship process is ubiquitously understood at a basic level by most adults. Many adults would also agree that schools are not doing the job assigned to them. But few would recognize that the extrinsic factors have caused the vast majority of issues in schools. Passing students on without requisite skills serves no one. But keeping them back incurs a social stigma in our culture of dramatically heavy import. It might be horribly demoralizing to a student. This is reflected all over our culture. But 'how' was this shift accomplished?

I now believe that the advent of the IQ test is directly to blame, through no fault of its own. The test gave the impression of hard factual evidence of the innate intellect of the person. And people were told this was true by other people who believed it to be true. (Psychomatricians, as a general rule, still believe it.) In the 1920's IQ was a miracle tool for judging people, especially minorities, and thus justifiably guiding decisions of work and social policy. Invalid as it was, it nevertheless became the mantra of the mind. By the 1940s, we had a generation of parents with a generation of children that lived under this belief. By the sixties, two generations. In 2020, we've had a hundred years of this fallacy, and almost no one alive remembers any different. At this point, the belief is so deeply held that questioning it is akin to heresy and getting rid of it is impossible. [The evidence proving it is invalid is far too numerous to include or summarize here. The best single source for this that I'm aware of is Gould's book, Mismeasure of Man.]

These systems of judgment have, of course, infiltrated every aspect of our lives. Job performance appraisals, tipping wait staff (I tip up front), tithing based on the quality of the sermon (or perhaps how well it matches with our deeply unconscious cognitive biases), racism, xenophobia, criminal justice. And the damage is not yet complete. By adhering to principles of Skinner's behaviorism (Kohn), we can further eliminate people's willingness to learn, or maybe

even to work at all. All we have to do is replace their intrinsic motivation with enough extrinsic motivators and then remove the extrinsic rewards.

Short version:

IQ tests led directly and quickly to loss of remediation (in schools)
Loss of remediation resulted in 20-80% loss of learning
Loss of intrinsic motivation (via punishment or rewards) led to the elimination of interest in learning

All of the above play a role at work too. High school graduates have roughly 12 years of motivation reduction. College graduates, 16 years. Or more.

Then we hire change management experts to 'undo' the trained reticence to change, learn, or improve. Experts are being rewarded to get other people to change, and likely receive rewards for their efforts.

Informal list of sources:

Deming, New Economics, Out of the Crisis, On Probability as a basis for action
Gould, Mismeasure of Man
Kahneman, Thinking Fast and Slow
Kohn, Punished By Rewards, Unconditional Parenting
Scholtes, Team Handbook, Leader Handbook (Video in the Deming Video Library)
Wu, blogs on intrinsic and extrinsic motivation,Lithiasphere

Addendum re. Deming's use of control charts combining enumerative measures with controlled experiments:

There is an exception to Deming's own rule. Shewhart gives us the control chart. Via the control chart, we can measure a process into stability, alter a variable, and measure the process again (presuming 30 iterations for acquisition of the standard deviation). Through such a study, we can use counting (in a controlled environment) to lead us to a conclusion about the cause of 'the change' in the counts.

There is a rationale to Deming's use of dual control charts. The fact is that the two control charts have a controlled experiment between them. If there is a change between the two control charts, AND, both charts are stable, (though different), we can KNOW with utter certainty, that the experiment we performed between them CAUSED the new result in the second chart. [Implied, but not stated, is the fact that if there is NO change, or a relatively very small change, we can be certain that our experiment had NO CAUSAL EFFECT.]

For clarity, the control chart certainty process requires three parts:

CONTROL CHART - EXPERIMENT - CONTROL CHART.