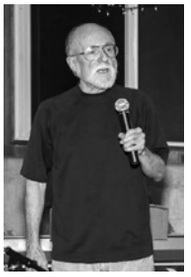
Excerpt from *Qualitative Research & Evaluation Methods*, by Michael Quinn Patton (2015, Sage Publications), pp. 93-95

MQP Rumination # 3

Fools' Gold: The Widely Touted Methodological "Gold Standard"
Is Neither Golden Nor a Standard

I am offering one personal rumination per chapter. These are issues that have persistently engaged, sometimes annoyed, occasionally haunted, and often amused me over more than 40 years of research and evaluation practice. Here's where I state my case and make my peace.



The *Wikipedia* entry for RCTs, reflecting common usage, designates such designs as the "gold standard" for research. News reports of research findings routinely repeat and reinforce the "gold standard" designation for RCTs (e.g., *New York Times*, 2014). Government agencies and scientific associations that review and rank studies for methodological quality acclaim RCTs as the gold standard. For example, researchers at the University of Saint Andrews reviewed "what counts as good evidence" and found that

> when the research question is "what works?", different designs are often placed in a hierarchy to determine the standard of evidence in support of a particular practice or programme. These hierarchies have much in common; randomised experiments with clearly defined controls (RCTs) are placed at or near the top of the hierarchy and case study reports are usually at the bottom. (Nutley, Powell, & Davies, 2013, p. 10)

The Gold Standard Versus Methodological Appropriateness

A consensus has emerged in evaluation research that evaluators need to know and use a variety of methods in order to address the priority questions of particular stakeholders in specific situations. But researchers and evaluators get caught in contradictory assertions: (a) select methods appropriate for a specific evaluation purpose and question, and use multiple methods—both quantitative and qualitative—to triangulate and increase the credibility and utility of findings, *but* (b) one question is more important than others (the causal attribution question), and one method (RCTs) is superior to all other methods in answering that question. This is what is known colloquially as talking out of both sides of your mouth. Thus, we have a problem. The ideal of researchers and evaluators being situationally responsive, methodologically flexible, and sophisticated in using a variety of methods runs headlong into the conflicting ideal that experiments are the gold standard and all other methods are, by comparison, inferior. Who wants to conduct (or fund) a second-rate study if there is an agreed-on gold standard?

The Rigidity of a Single, Fixed Standard

The gold standard allusion derives from international finance, in which the rates of exchange among national currencies were fixed to the value of gold. Economic historians share a "remarkable degree of consensus" about the gold standard as the primary cause of the Great Depression:

> the mechanism that turned an ordinary business downturn into the Great Depression. The constraints of the gold-standard system hamstrung countries as they struggled to adapt during the 1920s to changes in the world economy. And the ideology, mentality and rhetoric of the gold standard led policy makers to take actions that only accentuated economic distress in the 1930s. (Eichengreen & Temin, 1997, pp. 1–2)

The gold standard system collapsed in 1971 following the United States' suspension of convertibility from dollars to gold. *The system failed because of its rigidity*. And not just the rigidity of the standard itself but also the rigid ideology of the people who believed in it: Policymakers across Europe and North America clung to the gold standard despite the huge destruction it was causing. There was a clouded mind-set with a moral and epistemological tinge that kept them advocating the gold standard until political pressure emerging from the disaster became overwhelming.

Treating RCTs as the gold standard is no less rigid. Asserting a gold standard inevitably leads to demands for standardization and uniformity (Timmermans & Berg, 2003). Distinguished evaluation pioneer Eleanor Chelimsky (2007) has offered an illuminative analogy:

> It is as if the Department of Defense were to choose a weapon system without regard for the kind of war being fought; the character, history, and technological advancement of the enemy; or the strategic and tactical givens of the military campaign. (p. 14)

Indeed, while inquiries into the effectiveness of new pharmaceutical drugs or agricultural production techniques may benefit from RCTs, such designs are both inappropriate and misleading for inquiring into the complex, dynamic

(Continued)

(Continued)

phenomena that characterize much of the human condition. The problem is that treating experimental designs as the gold standard cuts off serious consideration of alternative methods and channels millions of dollars of research and evaluation funds into support for a method that has not only strengths but also significant weaknesses. The gold standard accolade means that funders and policymakers begin by asking, "How can we do an experimental design?" rather than asking, "Given the state of knowledge and the priority inquiry questions at this time, what is the appropriate design?" Here are examples of the consequences of this rigid mentality:

- At an African evaluation conference, a program director came up to me in tears. She directed an empowerment program with women in 30 rural villages. The funder, an international agency, had just told her that to have the funding renewed, she would have to stop working in half the villages (selected randomly by the funder) in order to create a control group going forward. The agency was under pressure for not having enough "gold standard evaluations." But, she explained, the villages and the women were networked together and were supporting each other. Even if they didn't get funding, they would continue to support each other. That was the empowerment message. Cutting half of them off made no sense to her. Or to me.
- At a World Bank conference on youth service learning, the director of a university program that placed student interns in rural villages in Cambodia offered her program for an exercise in evaluation design. She explained that she carefully selected 40 students each year and matched them to villages that needed the kind of assistance the students could offer. *Matching students and villages was key*, she explained. A senior World Bank economist told her and the group to forget matching. He advised an RCT in which she would randomly assign students to villages and then create a control group of qualified students and villages that did nothing to serve as a counterfactual. He said, "That's the only design we would pay any attention to here. You must have a counterfactual. Your case studies of students and villages are meaningless and useless." The participants were afterward aghast that he had completely dismissed the heart of the intervention: matching students and villages.
- I've encountered several organizations, domestic and international, that give bonuses to managers who commission RCTs for evaluation to enhance the organization's image as a place that emphasizes rigor. The incentives to do experimental designs are substantial and effective. Whether they are appropriate or not is a different question. In effect, it is *gold from those who enforce the gold standard to those who implement it*. One senior economist told me, "We're applying the basic economic principle of giving rewards for what we want. We pay for gold standard designs. We promote those who commission and conduct gold standard designs. So we get gold standard designs. You get what you reward." This explanation of how the world works rolled off his tongue like an oft-told story, which I suspect it was. He was smiling broadly at his cleverness, awash in his wisdom and self-congratulatory certainty. What I heard was rigidity, narrow-mindedness, misguided and distorted incentives, and an utter incapacity to detect even a hint of a problem.

Those experiences, multiplied 100 times, are what have generated this rumination.

Heaven or Gold Standard

Professor David Storey (2006) of Warwick Business School, University of Warwick, has offered a competing metaphor to replace the gold standard. He has posited "seven steps to heaven" in conducting evaluations, where heaven is a randomized experiment. So materially oriented and worldly evaluators are admonished to aspire to the gold standard, while the more spiritually inclined can aspire to follow the path to heaven, where heaven is an RCT.

In contrast, the metaphors of naturalistic inquiry are more along the lines of *staying grounded*, looking at *the real world* as it unfolds, *going with the flow*, *being adaptable*, and *seeing what emerges*.

Evidence-Based Medicine and RCTs

Medicine is often held up as the bastion of RCT research in its commitment to evidence-based medicine. Buthere again, gold standard designation has a downside, as observed by the psychologist Gary Klein (2014):

> Sure, scientific investigations have done us all a great service by weeding out ineffective remedies. For example, a recent placebo-controlled study found that arthroscopic surgery provided no greater benefit than sham surgery for patients with osteoarthritic knees. But we also are grateful for all the surgical advances of the past few decades (e.g., hip and knee replacements, cataract treatments) that were achieved without randomized controlled trials and placebo conditions. Controlled experiments are therefore not necessary for progress in new types of treatments and they are not sufficient for implementing treatments with individual patients who each have unique profiles.
>
> Worse, reliance on EBM can impede scientific progress. If hospitals and insurance companies mandate EBM, backed up by the threat of lawsuits if adverse outcomes are accompanied by any departure from best practices, physicians will become reluctant to try alternative treatment strategies that have not yet been evaluated using randomized controlled trials. Scientific advancement can become stifled if front-line physicians, who blend medical expertise with respect

> for research, are prevented from exploration and are discouraged from making discoveries.

Parachutes and RCTs

Gordon C. S. Smith and Jill P. Pell (2003) published a clever and widely disseminated systematic review of RCTs on parachute use in the *British Medical Journal*. They arrived at the following conclusions:

- No RCTs of parachute use have been undertaken.
- The basis for parachute use is purely observational (qualitative).
- "Individuals who insist that all interventions need to be validated by a randomized controlled trial need to come down to earth with a bump" (p. 1460).

On the other hand, they added that RCT's parachute research could still be done: "We feel assured that those who advocate evidence-based medicine and criticize use of interventions that lack an evidence base will not hesitate to demonstrate their commitment by volunteering for a double blind, randomized, placebo controlled, crossover trial" (p. 1460).

RCTs and Bias

RCTs aim to control bias, but implementation problems turn out to be widespread:

> Even in the most stringent research designs, bias seems to be a major problem. For example, there is strong evidence that selective outcome reporting, with manipulation of the outcomes and analyses reported, is a common problem even for randomized trails. (Chan, Hrobjartsson, Haahr, Gotzsche, & Altman, 2004, p. 2457)

The result is that "a great many published research findings are false" (Ioannidis, 2005).

Methodological Appropriateness as the *Platinum Standard*

It may be too much to hope that the gold standard designation will disappear from popular usage. So perhaps we need to up the ante and aim to supplant the gold standard with a new platinum standard: *methodological pluralism and appropriateness.* To do so, I offer the following *seven-point action plan*:

1. Educate yourself about the strengths *and weaknesses* of RCTS. (See resources below.)
2. Never use the gold standard designation yourself. If it comes up, refer to the "so-called gold standard."
3. When you encounter someone referring to RCTs as the gold standard, don't be shy. Explain the negative consequences and even dangers of such a rigid pecking order of methods.
4. Understand and be able to articulate the case for methodological pluralism and appropriateness, to wit, adapting designs to the existing state of knowledge, the available resources, the intended uses of the inquiry results, and other relevant particulars of the inquiry situation. (See the resources listed below.)
5. Promote the *platinum standard* as higher on the hierarchy of research excellence (and even closer to heaven).
6. Don't be argumentative and aggressive in challenging gold standard narrow-mindedness. It's more likely a matter of ignorance than intolerance. Be kind, sensitive, understanding, and compassionate, and say, "Oh, you haven't heard. The old RCT gold standard has been supplanted by a new, more enlightened, Knowledge-Age platinum standard." (Beam wisely.)
7. Repeat Steps 1 to 6 over and over again.

Resources

- For 10 limitations of experimental designs, see Patton (2008b, pp. 447–450).
- For the case against a rigid methodological hierarchy, see Ornish (2014), Nutley, Powell, and Davies (2013), Chen, Donaldson, and Mark (2011), Donaldson, Christie, and Mark (2009), Patton (2008b), Scriven (2008), and Julnes and Rog (2007).
- For official statements advocating methodological pluralism, see American Evaluation Association (2003), "Scientifically Based Evaluation Methods," and European Evaluation Society (2007), The Importance of a Methodologically Diverse Approach to Impact Evaluation."
- For alternative and mixed methods, see *Broadening the Range of Designs and Methods for Impact Evaluations* (Department for International Development, 2012), *Social Research Methods: Qualitative and Quantitative Approaches* (Bernard, 2013), *The SAGE International Handbook of Educational Evaluation* (Ryan & Cousins, 2009), and the alternative approaches for impact evaluation described in DFID (2012).
- *In platinum standard references, the designation "platinum standard"* has been used by Deborah Lowe Vandell, chair of the Department of Education at the University of California at Irvine, to describe an e*valuation design that uses a range of data collection app*roaches (e.g., observations, interviews, surveys) to collect qualitative and quantitative data on program implementation and outcomes (*Evaluation Exchange*, 2010). See also "Toward a Platinum Standard for Evidence-Based Assessment by 2020," which incorporates *case studies, comparative methods, triangulation, and alternative causal approaches in recognition of the wide array of goals and methodologies that are appropriate for assessing programs and policies in a dynamic and globalizing world (*Khagram & Thomas, 2010); "Evaluating Outside the Box: Mixing Methods in Analyzing Social Protection Programmes" (Devereux & Roelen, 2013); and Scriven's (2008) article, which refers to a higher standard than gold for causal research, a platinum standard (p. 18).