Copyrighted Material

Introduction

It is rare to know exactly when and where an idea originates. But for me it was Monday, September 7, 1970, at 5:45 P.M. I was on the South Side of Chicago walking north on Dorchester Avenue, between Fifty-sixth and Fifty-seventh streets. I was a new assistant professor at the University of Chicago, and was walking home from my first day of work. Standing in the street was a very large, black motorcycle next to an even larger, fearsome looking, man. He was wearing motorcycle boots and a leather jacket, had long hair and a full beard, and was speaking with a young woman whose apartment was in a building two doors down from mine. In passing I overheard a snippet of conversation. He shook his head in response to something she said and replied, "But that's an epistemological question."

Epistemology was almost a foreign word and concept to me at that time. I had heard the term before. Adjoined to metaphysics and phenomenology, it formed an almost holy triumvirate in the requisite undergraduate introductory course in philosophy, but it had made no impact. However, hearing it used on the street made a difference. Returning home, I looked it up and found "method for gaining knowledge," which might translate into "scientific method," which might then be specialized to "procedures used by actual practicing scientists."

The iconic physicist Richard Feymann provided, as only he could, a clear description of modern epistemology in his famous Messenger Lectures, given at Cornell in 1964 (subsequently anthologized in his book *The Character of Physical Law*):

> In general we look for a new law by the following process. First we guess it. Then we compute the consequences of the guess to see what would be implied if this law that we guessed is right. Then we compare the result of the computation to nature, with

Copyrighted Material

> experiment or experience, compare it directly with observation, to see if it works. If it disagrees with experiment it is wrong. In that simple statement is the key to science.
>
> *It does not make any difference how beautiful your guess is. It does not make any difference how smart you are, who made the guess, or what his name is—if it disagrees with experiment it is wrong. That is all there is to it.* (1965, 156)

I don't know what epistemological question was being discussed on Dorchester Avenue forty years ago, but two key ones are how and when evidence should be used. It was clear that Feynman placed evidence in an exalted position. It vetoed all else. Yet strangely, at least to me, this point of view is not universal.

We hear the term *evidenced-based decision making* in many fields; medicine, education, economics, and political policy, to pick four. Its frequent use implies that this is a new and modern way to try to solve modern problems. If what we are doing now is evidence based, what were we doing previously?[1]

How can we consider the use of evidence in science new? Hasn't evidence been at the very core of science for millennia? The short answer is no. Making decisions evidence-based has always been a tough row to hoe, for once you commit to it, no idea, no matter how beautiful, no matter how desirable, can withstand an established contrary fact, regardless of how ugly that fact might be. The conflict between evidence and faith in the modern world is all around us, even in scientific issues for which faith is not required.[2] So it is not surprising that using evidence to make decisions is taking a long time to catch on. The origination of the formal idea of using evidence as a method for gaining knowledge is often dated, as are so many things, with Aristotle (384 B.C.–322 B.C.) but its pathway thereafter was not smooth, for once one commits to using

[1] When I suggested that evidence-based medicine's predecessor must have been faith-based, my boss, Donald Melnick, corrected me and said that he liked to think that medicine was intelligently designed.

[2] A story is told of a conversation between Napoleon and Laplace in which Napoleon congratulated Laplace on the publication of his masterwork, *Traité de Mécanique Céleste*, but then added that he was disappointed because "no where in this great work was the name of God mentioned even once." Laplace is said to have responded, "I did not need that hypothesis."

Copyrighted Material

evidence to make decisions, facts take precedence over opinion.[3] And not all supporters of an empirical approach had Alexander the Great to watch their backs. Hence it took almost 2000 years before Francis Bacon (1561–1626) repopularized the formal use of evidence, an approach that was subsequently expanded and amplified by British empiricists: the Englishman John Locke (1632–1704), the Irishman George Berkeley (1685–1753), and the Scot David Hume (1711–1776).

Although the development of a firm philosophic basis for incorporating evidence in how we know things was necessary, it was not sufficient. Much more was required. Part of what was needed was a deep understanding of uncertainty. This was recognized and, almost coinciding with the onset of the twentieth century, began the development of statistics, the Science of Uncertainty. Statistics' beginning was primarily mathematical, with a focus on fitting equations to data and making judgments about the legitimacy of the inferences one might draw from them. This changed in 1977 with the publication of John Tukey's (1915–2000) *Exploratory Data Analysis.* Tukey, a towering figure of twentieth-century science, legitimized the practice of the atheoretical plotting of points with the goal of finding suggestive patterns. He pointed out that "the greatest value of a graph is when it *forces* us to see what we never expected." Tukey's key contribution was his legitimization of this sort of empirical epistemology. He likened exploratory analysis to detective work, in which the scientist gathers evidence and generates hypotheses, the guesses that Feyman referred to.

Traditional statistical methods were more judicial in nature, in which the evidence was weighed and a decision was made. The modern scientific world has both the philosophic and mechanical tools to use evidence to generate hypotheses and to test them. Yet the rigorous thinking that the scientific method requires has yet to penetrate public discourse fully. Guesses are made, sometimes from intuition, sometimes from hope, sometimes from dogma. But too often these guesses only make sense if you say them fast—for when they are tested with evidence and logic, they are found faulty.

[3] Bertrand Russell reports that even Aristotle had trouble following the tenets of empiricism in all aspects of his own life, for he maintained that women have fewer teeth than men; although he was twice married, it never occurred to him to verify this statement by examining his wives' mouths.

Copyrighted Material

In the chapters to follow I will show that if we decide to use evidence, we can discover things. Usually these discoveries add to our store of knowledge, and we are happy to have found them. But sometimes what we discover conflicts sharply with our intuition. It is in situations like these that our commitment to empiricism, as a way of knowing things, gets tested. When this experience of contradictory evidence happens—to return to the sidewalk wisdom with which I began this introduction—what we decide to do becomes an epistemological question.

Evidence of success in contemporary education encompasses many things, but principal among them are test scores. When scores are high, we congratulate all involved. When they are low, we look to make changes. When there are differences between groups, ethnic or gender, we are concerned. If these differences shrink we are pleased; if they grow larger we often metaphorically shoot either the messenger (the test) or the educators.

Shooting the messenger as a strategy for dealing with bad news has a long history. And it persists despite its low likelihood of sustainable success. In this book I discuss the use of tests and their associated scores as evidence in making educational decisions. The examples chosen illustrate only a small portion of the range of uses to which tests are put—from traditional uses like making the triage decision about admittance to college (chapters 1 and 2), to awarding scholarships (chapter 3), to allocating educational resources for instruction (chapter 4), to judging the quality of instruction (chapter 9). In the course of these illustrations it seems worthwhile to illuminate some commonsense ideas on the use of tests that, with some thought, we discover are deeply flawed (chapters 5, 6, 7, and 8).

Let us start at the beginning.

The use of mental tests appears to be almost as ancient as civilization itself. The Bible (Judges 12:4–6) provides an early reference in Western culture. It describes a short verbal test that the Gileadites used to uncover the fleeing Ephraimites hiding in their midst. The test was one item long. Candidates had to pronounce the word *shibboleth*; Ephraimites apparently pronounced the initial *sh* as *s*. The consequences of failure were severe, as the Bible records that the banks of the Jordon River were littered with 42,000 bodies of Ephraimites (it is unknown how many of those 42,000 were Gileadites with a lisp).

There is substantial evidence of the beginnings of an extensive testing program in China at around 2200 B.C., predating the biblical

Copyrighted Material

> Jephthah then called together the men of Gilead and fought against Ephraim. The Gileadites struck them down because the Ephraimites had said, "You Gileadites are renegades from Ephraim and Manasseh."
>
> The Gileadites captured the fords of the Jordan leading to Ephraim, and whenever a survivor of Ephraim said, "Let me cross over," the men of Gilead asked him, "Are you an Ephraimite?" If he replied, "No," they said, "All right, say 'Shibboleth.'" If he said, "Sibboleth," because he could not pronounce the word correctly, they seized him and killed him at the fords of the Jordan.
>
> Forty-two thousand Ephraimites were killed at that time.
>
> Judges 12:4–6

program by almost a thousand years. The emperor of China is said to have examined his officials every third year. This set a precedent for periodic exams in China that was to persist for a very long time. In 1115 B.C., at the beginning of the Shang dynasty, formal testing procedures were instituted for candidates for office.

The Chinese discovered the fundamental tenet of testing:

> *a relatively small sample of an individual's performance, measured under carefully controlled conditions, could yield an accurate picture of that individual's ability to perform under much broader conditions for a longer period of time.*

China's testing program, augmented and modified, has lasted almost uninterrupted for more than four thousand years. It was advocated by Voltaire and Quesnay for use in France, where it was adopted in 1791, only to be (temporarily) abolished by Napoleon. It was cited by British reformers in 1833 as their model for selecting trainees into the Indian civil service system—the precursor of the British civil service. The success of the British system influenced Senator Charles Sumner and Representative Thomas Jenckes in their development of the American civil service examination system that they introduced into Congress in 1868. There was a careful description of the British and Chinese systems in Jenckes's report *Civil Service of the United States*,

Copyrighted Material

which laid the foundation for the establishment of the Civil Service Act passed in January 1883.

The use of large-scale testing grew exponentially in the United States after World War I, when it was demonstrated that a mass-administered version of what was essentially an IQ test (what was then called "Army Alpha") improved the accuracy and efficiency of the placement of recruits into the various military training programs. The precursors of what would eventually become the SAT were modeled on Army Alpha.

Testing has prospered over the four millennia of its existence because it offers a distinct improvement over the method that had preceded it.

> *To count is modern practice, the ancient method was to guess.*
>
> *—Samuel Johnson*

Testing also fit comfortably into the twentieth-century meritocratic zeitgeist where advancement was based increasingly on what you knew and could do instead of your lineage and wealth.

But as time has passed and test usage increased, the demands that we have made on test scores have increased, as have the "fineness" of the distinctions we wish to make. As this has happened, tests and how they are scored have improved. These improvements have occurred for three principal reasons:

1. The demands made on the test scores have become more strenuous.
2. We have gathered more and more evidence about how well various alternatives perform.
3. Our eyes have become accustomed to the dim light in those often dark and Byzantine alleys where tests and psychometrics live.

But although deep knowledge of testing has increased, testing's usage has expanded well beyond the cadre of experts who understand it. In the modern world, too often those who use test scores as evidence to guide their decisions are unacquainted with testing's strengths and weaknesses. Instead they often find that test scores are (to borrow Al Gore's evocative phrase) an inconvenient truth. In short they are facts that get in the way of the story that they want to believe. When this happens, either the facts are ignored or their accuracy is maligned. In this section I will lay out some facts and arguments in the hope that

Copyrighted Material

future decision-makers will understand better how to use this marvelous invention to assess the state of the educational enterprise and thence to amend its flaws.

The first three examples all grow from a report published in September 2008 by the National Association for College Admission Counseling (NACAC). The report was critical of the current, widely used, college admissions exams, the SAT and the ACT, and made a number of recommendations for changes in the admissions process. It was reasonably wide-ranging and drew many conclusions while offering alternatives. A description of this report's findings was carried broadly in the media. Although well-meaning, many of the suggestions only make sense if you say them very fast.

Three of its major conclusions were the following:

1. Schools should consider making their admissions "SAT optional," that is allowing applicants to submit their SAT/ACT scores if they wish, but they should not be mandatory. The commission cites the success that pioneering schools with this policy have had in the past as proof of concept.
2. Schools should consider eliminating the SAT/ACT altogether and substituting achievement tests. The report cites the unfair effect of coaching as the motivation for this proposal. Its authors were not naive enough to suggest that because there was no coaching for achievement tests now that, if they carried higher stakes, coaching for them would not be offered. Rather they claimed that such coaching would be directly related to schooling and hence more beneficial to education than coaching that focused solely on test-taking skills.
3. The use of the PSAT with a rigid qualification cut-score for such scholarship programs as the Merit Scholarships should be immediately halted. It should be replaced with a more rigorous screening test without a fixed minimum eligibility score.

I use evidence to examine the validity of these recommendations in the first three chapters. This information provides enough momentum to carry us through the next seven chapters, in which we investigate other uses of tests within the educational system and discuss how they have been misused, as well as what can be done to ameliorate these problems.