Measurement in Education: Its Lure and Liabilities

Francis Schrag
University of Wisconsin, Madison

Measurement has played an indispensable role in designing and building our modern world. Even in occupations, such as medicine, whose focus is humans rather than inanimate objects, measurement has been a source of progress though its liabilities have not gone unnoticed.[1] On the other hand, when it comes to education, the verdict is decidedly mixed and contested; many thoughtful observers deem measurement a scourge.[2] Is this judgment fair? Not entirely, I shall argue; yet, I shall maintain that there *are* important differences between education and medicine that account for the different conclusions we reach about the role measurement ought to play in each domain.

WHAT IS MEASUREMENT?

Although space forbids an in-depth discussion of the nature of measurement, itself, we must begin by asking what measurement is and what makes it possible. There is an enormous literature on the topic and competing positions have been articulated.[3] One is the view of Joel Michell (inter alia) who claims that measurement is inherently associated with quantity and may be considered "the jewel in science's crown."[4] Michell defines measurement as "*the estimation or discovery of the ratio of some magnitude of a quantitative attribute to a unit of the same attribute*."[5] The measurement of temperature provides an obvious example here of what I shall call measurement in the more restrictive sense (MR).

Luca Mari and co-authors articulate a starkly different view. They deny that quantification is either necessary or sufficient for measurement: "In our view, when one claims to be engaged in 'measurement activities,' one is claiming that one is attempting to develop methods of obtaining high-quality information about a property."[6] What is the meaning of "high-quality" information here? According to the authors:

> the appraisal of quality generally involves evaluation of the degree to which there is a theoretical basis for the procedures that generated the information, *the extent to which the results are stable and can be reproduced, and the usefulness of the information in assisting the accomplishment of tasks* or the answering of questions of interest.[7]

When several judges provide a numerical rating to a student essay or to a gymnastics performance, and the ratings are then averaged, that illustrates measurement in what I shall designate as a more permissive sense (MP). I have neither the background nor the need to adjudicate the dispute between the two positions, but I think it not unfair to say that measurement and quantification are often identified because of the advantages the latter brings not only to science but to myriad practical pursuits from carpentry to rocketry to food preparation.

A minimum condition of an attribute's being measurable (MR) is that it admits of more or less of the attribute. A second condition is that objects be capable of being *ordered* according to *how much* of the attribute they contain. But we do not yet have MR. What is lacking? We must be able to devise a numerical *scale* on which a number can be found corresponding to the magnitude of the attribute possessed by that object. A so-called "interval" scales is needed; here, the quantitative distance between two or more units is the same on every part of the scale.[8]

What advantages does quantification (whether MR or not) provide? Consider a middle school selecting students for an advanced mathematics program and trying to justify those decisions to parents of students who were not selected. Suppose those judgments are made by teachers on the basis of their day-to-day work with children. These judgments may be very objective, but a disappointed parent would be more accepting if scores on a schoolwide math test showed her daughter with a score of 71 when the cutoff for the advanced class was 88. Here, quantification provides a way of making diffuse and private judgments more precise and public, hence less likely to evoke accusations of bias or unfairness. Beyond that, quantification provides opportunities to compare not simply individual students but schools, or entire educational systems

DOI 10.47925/75.2019.488

across time and space. Think of the problem of selecting students for an elite medical school with an acceptance rate below four percent. Do you think that the fairness of the selection process would be enhanced by eliminating applicant scores on the Medical College Application Test (MCAT)—a 7.5 hour test taken by 85,000 students in more than 20 countries?

ARE EDUCATIONAL ATTAINMENTS SUCCEPTIBLE TO BEING MEASURED?

But wait! Can we actually *measure* (MR) a student's capacity for "Critical Analysis and Reasoning Skills" or knowledge of "Psychological, Social, and Biological Foundations of Behavior," two sections of the exam? Well, couldn't we simply formulate a reasonable number of test questions in each domain, and rank the candidates according to the percentage of correct answers? Recall that for MR, we must be able to correctly claim, not only that Jill has a better understanding of the domain than Jack—who, in turn, has a better understanding than Jody—but we must be able to construct an interval scale that permits us to place all three students on that scale. Otherwise, we may not realize that the gap between Jill and the other two is huge, while Jack and Jody are actually quite close in their mastery. And, beyond that, we must be able to say: a.) that any alternative approach to measuring that attribute in the three students would not change the relative placements on the new scale, and b.) if the three students took an equivalent test tomorrow, their scores (adjusting for measurement error) would remain the same. Can these things be accomplished?

I believe it is fair to say that contemporary psychometricians have found means to design and score tests in such a way as to produce measures *that approach interval measures* typically found in the physical sciences. Psychometric approaches depend on using sophisticated statistical analysis and scales in such a way that assessing the student's performance on a test takes the difficulty of the test questions into account. Diverse approaches, including Rasch analysis and Item Response Theory, depend on having adequate samples of students of varying ability levels, preferably a random sample, and an adequate number

of "items" of different levels of difficulty.[9]

Just as a single temperature reading may, for a variety of reasons, not provide an accurate assessment—perhaps the thermometer was not shaken down sufficiently, or was left out in the sun, or perhaps the patient is an outlier with a high (or low) normal—requiring a follow up a bit later, so one score on an MCAT might fail to accurately record an applicant's ability. For that reason, medical school applicants are permitted to take the MCAT up to three times in a single year, and seven times in a lifetime.

While often derided, multiple choice, machine scored tests like the MCAT must be appreciated for what they deliver. The MCAT enables the ranking of tens of thousands of applicants from around the world on a single scale, capturing their readiness for the academic rigors of medical school. Whether or not the MCAT actually predicts success in medical school, it is certainly perceived to be a legitimate screening device.[10] Since the test is given thirty times a year, there must be thirty different versions to avoid cheating, yet all versions must be capable of being equated so that students taking the test in, say, August, have neither an advantage nor a disadvantage over those taking it in February. Moreover, this year's version must be of comparable difficulty to the ones given in previous years despite continual addition and deletion of specific questions. It is in everyone's interest to have tests scored promptly and accurately; indeed, MCAT scores are released a few hours after applicants take the test. All in all, the MCAT (and its counterparts in other fields) must be recognized as remarkable pieces of social technology, essential building blocks of our meritocracy.[11]

I used the MCAT to illustrate the power of psychometrics in education; the MCAT is one of many tests whose basic purpose is to rank individual test takers—who live vast distances apart—for admission to a school or university when, as is often the case, the number of applicants vastly exceeds the capacity of an institution to serve them. In the case of tests like the MCAT, the test publisher provides the rankings of applicants, but each medical school admissions committee decides on whether a particular applicant's performance on the MCAT warrants continuing consideration or elimination from the pool of

DOI 10.47925/75.2019.488

applicants. A person scoring very poorly might turn out to be a good candidate and a person with a high MCAT score may flunk out, but the chances are against it. The meaning and weight to put on the score of a particular applicant in the context of all the other information about her is ultimately a matter of *judgment*, and committee members may and do often disagree. Such judgment may be arbitrary, in a sense, but it is not capricious. Think of it this way: An admission officer favoring admission of a candidate who performed poorly on the MCAT would need to persuade her colleagues that other evidence presented on her behalf is strong enough to override her low MCAT score.

Before discussing the weaknesses of MR in education, let's note that MP in the form of quizzes, semester-end tests, term papers, etc. have always been and rightly continue to be an intrinsic aspect of schooling. When trying to identify the disadvantages of measurement (MR) in education, it is hard not to go off the rails in two directions. First, it is easy to conflate measurement and testing with the way tests are used, with what has been called a testing regime. Many of the complaints against the current testing regime in the United States should not be considered liabilities of measurement itself, only of its misuse. The second way to go off the rails is to adhere to the common view that machine-scored, multiple choice tests can only assess low level skills and facts, not higher-level capabilities like reasoning skills and skill in judgment. Granted that questions dealing with facts and low-level skills are the most common and the easiest to formulate, questions designed to assess complex, sophisticated skills can be accommodated. Consider this question adapted to multiple-choice format from a recent column in the *New York Times*:

> A. Four cards are laid in front of you, each of which, it is explained, has a letter on one side and a number on the other. The sides that you see read: E, 2, 5 and F. Your task is to turn over only those cards that could decisively prove the truth or falsity of the following rule: "If there is an E on one side, the number on the other side must be a 5." Which ones do you turn over?
>
> a.) E and 5; b.) F and 5; c.) E and 2; d.) E and F[12]

Tests of situational judgment have also found favor in certain occupations, such as nursing. Here, "Candidates are asked to identify the appropriateness or effectiveness of various response options from a pre-defined list of alternatives. These response options are designed in advance with a pre-determined scoring key agreed to by subject matter experts."[13]

We must ask: Is there anything *inherently* deleterious in conceptualizing educational accomplishment in such a way as to make it amenable to measurement in the more restrictive sense? To determine this, let's identify the fundamental assumptions that underlie MR in education.

1. That "more" and "less" are appropriate adjectives to describe what people know and can do as a result of learning.

2. That the universe of knowledge and judgment can be broken down into separate, (overlapping) domains, and that each of these can be further analyzed into individual propositions and skills.

3. That an assessment of what individuals know within a domain can be conducted in a reasonable amount of time—measured in hours--by *counting* and analyzing responses to a reasonable number of "items" (30—250) randomly selected from that domain, given:

a. That the items can be ranked in terms of difficulty, and

b. That they admit of "right" (or at least more or less "right") answers *according to the test preparer*.

c. That they are not biased against any one type or group of students.[14]

4. That when multiple forms of a test are needed, statistical techniques are able to render the different forms sufficiently equivalent, given the purpose of the test.

5. That even if skills and proficiencies beyond those being tested are required to correctly answer the items, these will not contaminate the

DOI 10.47925/75.2019.488

score on the domain being assessed.

I want to underline the constraint imposed by the test *format* regardless of domain: the examinee's task is limited to selecting the *most appropriate* answer from the options *provided by the examiner*. One risk here is that inferences drawn from examinee performance may be unwarranted because the scoring algorithm cannot recognize the extent to which that performance tracks command of the *format* rather than mastery of the domain. But my key point can best be perceived by contrasting the previous question with this one:

> B. Formulate three multiple-choice questions of different degrees of difficulty to test students' abstract reasoning skill, and explain how you would validate the levels of difficulty.

In contrast to A, B requires a considerable background in *multiple* domains including the psychology of reasoning, psychometrics, and test design. More important, there is *no correct* answer to the question. The responses do not lend themselves to *measuring* examinees' capabilities, only to *judging* them. Observe, as well, that numerous adjectives might be applied to responses to B but not to A: solid, weak, original, ingenious, unintelligible, simplistic, pirated, elegant, slipshod, and so on. We can, therefore, say that B, unlike A, is both a much more authentic task, and provides access to traits that potentially tell us much more about the examinee, but at a cost of a potential loss of objectivity in judging the responses.

We may now reach the following tentative conclusion: Measurement MR (or something very close to it) is possible in education but can be achieved only by imposing certain constraints on the nature of the tasks presented to examinees, which limits the qualities of mind and character capable of being assessed.

Are these limitations acceptable and do the benefits outweigh the liabilities? This is not a question that can be answered, in general, and without respect to context. I find the use of machine scored tests as one, among several, applicant screening devices to be legitimate. However, we cannot make a balance sheet without considering their liabilities. There are indeed many, and my purpose

here will be served by merely listing them: teaching to the test, narrowing of the curriculum, devaluing originality; conveying a misconception about real-world problems; demoralizing teachers and deflating their status relative to that of psychometricians; encouraging gaming the system, if not outright cheating; putting needless stress on students, undermining the inherent satisfaction of learning; deceiving the public; intensifying competition and ranking among students. All these disadvantages derive from two liabilities*: a shift in attention from the good or end sought to the number taken to represent the amount of that good, and, where multiple goods (and bads) are involved, a shift in attention from the whole panoply to the most easily quantified.*

MEASURMENT IN MEDICINE AND EDUCATION COMPARED

My final question remains: What accounts for the very different balance sheet we feel compelled to draw up in medicine and in education? My answer is in three parts:

First, the amenability to measurement (MR) of the two goods, health and education differs. The idea that health consists in some kind of balance or harmony of diverse substances goes back to the ancient Greeks, of course. For example, humans are unlikely to survive temperatures below 35° or above 42° C. Everything from cholesterol, to hormones, to cells of various types to vitamins, minerals, antigens, amino acids, bacterial counts, to say nothing of bone density, pulse rate, and blood pressure can be measured (MR). These all have been discovered to exhibit a *normal range in healthy human populations*, beyond which typically lies disease or even death.[15] We have seen that some educational outcomes can be manipulated to yield measurable characteristics, but others, not. One person might be said to have more or less understanding of quantum physics, more or less facility with Chinese language, more or less skill as a carpenter, more or less insight into a novelist's intentions, more or less originality as a script writer, more or less stage presence, more or less elegance as a figure skater, but in none of these cases need we say that the qualitative "more" and "less" correlates with or can be explained by some higher or lower *number of*

DOI 10.47925/75.2019.488

measurable units.

Second, there is a potential cost to measurement in the educational domain that I have not mentioned, one that does not have its counterpart in the medical, namely the role measurement plays in contributing to our self-concept. Diagnosis in medicine *can* play a similar role. A person with diabetes may feel, not simply that he has diabetes, but that he *is* a diabetic; this is a part of his or her identity.[16] But people don't identify themselves with their health *measurements*, with the amount of sugar in their blood, their number of heart beats per minute, and the like. Not so with measurements in education; these "create" high and low scorers, or, for example, perfect scorers on the ACT or SAT. Moreover, these identities are always *relative* to others with higher or lower scores; they inevitably create a hierarchy and each test-taker's place in that hierarchy as Michel Foucault contended.[17] I, therefore disagree with psychometrician Daniel Koretz when he asserts: "Tests may 'designate' winners and losers, but they don't <u>create</u> them. There simply <u>are</u> winners and losers."[18] Yes, on most tests, especially those well designed to allocate candidates to scarce opportunities, necessarily some will score higher and others lower. But the meaning of those scores is not inherent in them; we assign those meanings to them. Karl Marx and Friedrich Engels envisioned a society in which social goods were allocated from each according to their abilities, to each according to their needs. If this admittedly utopian ideal were to be realized, the high scorers would not be winners, nor the low scorers, losers.

Third, and most important, the relation between the goods and their quantitative representations is very different in the two domains. The best way to appreciate this is via a simple thought experiment: If forced to choose, which would you want for your son or daughter, excellent health and mediocre "numbers" on a comprehensive blood panel; or excellent "numbers" and mediocre health? Now, ask: If forced to choose, which would you want for your son or daughter aspiring to become a physician, a solid mastery of college subjects and a low score on the MCAT exam; or a top score on the exam but a mediocre grasp of college subjects? My guess is that you will say that the first choice is a "no brainer." Health is what you value. In the second case, my guess is that

you will feel somewhat conflicted. Why? Because blood panel numbers have no value in and of themselves, whereas top MCAT scores have enormous value in facilitating access to income and social status. Neither education nor health are, in themselves, positional goods. Your having either good is no impediment to my having that same good, and might even enhance my ability to get it. Blood panel scores are not positional goods either. Your son or daughter's having very good cholesterol numbers does not put anyone else at a disadvantage. Not so with MCAT scores; your son or daughter can achieve a high score *only if* someone else's son or daughter gets a low score.

Because educational *credentials* are positional goods, there is a strong incentive to focus on the numerical representation of the good, shifting attention from the good, itself, and thereby inviting all sorts of mischief. Historian Theodore Porter summarizes the situation admirably:

> Whoever can exploit the ambiguity of measures to fulfill numerical targets without having to expend resources on the thing measured enters into the domain of funny numbers. Such opportunities will be found wherever approval, payment, or some other desired end is made contingent on a quantitative standard.[19]

While in the case of medicine, there is no doubt that the benefits of measurement outweigh the costs, things are not so clear in education. How could we find out? Before confronting that questions let's ask whether a much more limited role for educational measurement is even feasible in contemporary societies. Indeed, it is, as illustrated by Finland—considered a world leader in education.[20] Students face exams constructed by their teachers during their schooling and only at the end of secondary school do they confront a national, standardized exam, which they must pass in order to graduate from high school. The exam is graded twice, once by students' own teachers and once by independent examiners from the ministry of education. "The grading uses a seven-point-scale adjusted to normal distribution. This means that the number of top grades and failed grades in each exam is approximately 5 %. The questions are largely open ended, most requiring extensive writing. A typical essay question reads: Media

DOI 10.47925/75.2019.488

is competing for audiences—what are the consequences?"[21]

Clearly, Finland has a very different testing regime than the United States, making minimal use of measurement, depending much more on expert judgment. Suppose we were trying to draw up a balance sheet for the two testing regimes, how would we proceed? Let's begin by selecting a state of comparable size and clearly contrasting testing regime, Wisconsin. Now let's ask, could we *measure* the educational success of Finland versus Wisconsin, then compare them? But wait, would we do that via Finnish style or US style testing? United States psychometricians would probably insist on the latter, arguing that the former is too subjective. Finnish testing experts would likely argue that multiple choice testing would not capture the qualities the Finnish education system is trying to cultivate. The contest over how to assess educational outcomes would only be intensified when it came to attempts to measure the social and economic consequences of those outcomes. For example, Wisconsinites might say that mean GNP is the most relevant metric, and here Wisconsin dominates Finland. But Finns likely would reject that measure, favoring indices of income inequality like the Gini coefficient; here, the Finns enjoy substantially lower inequality despite the fact that Wisconsin has one of the lowest levels of income inequality among the fifty states.[22] Of course, we could imagine a joint team of Wisconsin and Finnish experts negotiating and deciding how much to weight each of the many dimensions for which quantitative measures exist. This composite, however, would *not* be MR. My point is not simply that the value of measurement in education cannot, itself, be *measured* (MR). I have a deeper point: The choice of an appropriate yardstick or metric is, itself, influenced by the prevailing educational culture, of which the testing culture is a part. There is no neutral way to identify appropriate educational metrics.

SUMMARY

1. The lure of measurement derives from the aspiration of replacing vague and private judgments by more precise public judgments.

2. The multiple choice, machine-graded test is an impressive piece of social

technology with legitimate use as a screening device.

3. The basis for the many disadvantages resulting from measurement derive from the shift in attention from the goods themselves to their numerical proxies and from the whole panoply of goods to the most easily quantified.

4. In medicine, the benefits of measurement clearly outweigh the costs, although the latter are not trivial; in education, the costs are high relative to the benefits. This is primarily because, in contrast to medicine, the proxies themselves, not what they purport to represent, have social and economic value.

5. Testing regimes cannot be *measured*; the selection of metrics to assess a testing regime, is, itself, influenced by the educational and testing cultures that exist in a society.

Acknowledgements

I would like to acknowledge helpful feedback and suggestions I received from William Boone, Joel Mitchell, Dan Hausman, Daniel Bolt, Denis Phillips, Bill Schwab, and Elliott Sober.

1 See e.g., Peter C. Gøtzsche et al. "Beware of Surrogate Outcome Measures," *International Journal of Technology Assessment in Health Care* 12, no. 2 (1996): 238-246.
2 See e.g., Daniel Koretz, *The Testing Charade: Pretending to Make Schools Better* (Chicago: University of Chicago Press, 2017).
3[3] See Nancy Cartwright and Norman Bradburn, "A Theory of Measurement," in *The Importance of Common Metrics for Advancing Social Science Theory and Research: Proceedings of the National Research Council Committee on Common Metrics* (Washington, DC: National Academies Press, 2011); L. Finkelstein, "Problems of Measurement in Soft Systems," *Measurement* 38, no.4 (2005): 267-274; Giovanni Battista Rossi, "Measurability," *Measurement* 40 (2007): 545-562; Eran Tal, "Old and New Problems in Philosophy of Measurement," *Philosophy Compass* 8, no.12 (2013): 1159-1173.
4 Joel Michell, "Militant Pantometry: Logical Limit of Measurement, and the Prefabrication of Psychometrics," paper presented at Dimensions of Measurement Conference, Center for Interdisciplinary Research, University Bielefeld, Germany, March 14-16, 2013, 1.
5 Joel Mitchell, "Quantitative Science and the Definition of Measurement in Psychology," *British Journal of Psychology* 88 (1997), 358 (italics in original).
6 Lucas Mari, Andrew Amul, David Torres Irribara, and Mark Wilson, "Quantities,

DOI 10.47925/75.2019.488

Quantification, and the Necessary and Sufficient Conditions for Measurement," *Measurement* 100 (2017), 120.
7 Ibid., 120n13 (emphasis added).
8 See Rossi, 555-558; B. D. Wright and John M. Linacre, "Differences Between Scores and Measures," *Rasch Measurement Transactions* 3, no. 3 (1989), 63, http://www.rasch.org/rmt/rmt33a.htm.
9 Chong Ho Yu, "A Simple Guide to the Item Response Theory (IRT) and Rasch Modeling,"
http://www.creative-wisdom.com; Ronald K. Hambleton and Russell W. Jones, *Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development: An NCME Instructional Module* (Washington, DC: Fall 1993).
10 Ellen R. Julian, "Validity of the Medical College Admission Test for Predicting Medical School Performance," *Academic Medicine* 80, no. 10 (2005): 910-917.
11 I am aware of the many criticisms that have been aimed at both the concept and the reality, but ask yourself: Would you prefer to identify physicians by lottery?
12 Manil Suri, "Does Math Make You Smarter?," *The New York Times*, April 13, 2018, http://www.nytimes.com/2018/04/13/opinion/sunday/math-logic
13 Fiona Patterson, Lara Zibarras & Vicki Ashworth, "Situational Judgment Tests in Medical Education and Training: Research, Theory and Practice: AMEE Guide No. 100," *Medical Teacher* 38, no. 1 (2016): 4-5.
14 I owe this point to Denis Philips.
15 Not every statistical outlier requires medical attention, e.g., 7-foot-tall people.
16 See Ian Hacking, "Making Up People," *London Review of Books* 28, no.16 (August 17th, 2006), http://www.generation-online.org/c/febiopolitics2.htm.
17 Michel Foucault, *Discipline and Punish: The Birth of the Prison*, trans. Alan Sheridan (New York: Vintage, 1995), 170-194.
18 Daniel Koretz, *Measuring Up: What Educational Testing Really Tells Us* (Cambridge, MA: Harvard, 2008), 53.
19 Theodore M. Porter, "Funny Numbers," *Culture Unbound* 4, no. 4 (2012), 1.
20 Pasi Sahlberg, "The Brainy Questions on Finland's Only High Stakes Test," *Washington Post* (March 24, 2014) https://www.washingtonpost.com/news/answer-sheet/wp/2014/03/24/the-brainy-questions-on-finlands-only-high-stakes-standardized-test/?utm_term=.90a005f53c28 (I put aside the question of what evidentiary basis supports this judgment).
21 Ibid., 3.
22 "List of U.S. States and Territories by GDP," *Wikipedia*, last modified October 14th, 2019, https://en.wikipedia.org/wiki/List of U.S states and territories by GDP; http://tradingeconomics.com/finland/gdp-percapita.