Measurement Meanderings: Response to Francis Schrag

Tone Kvernbekk
University of Oslo

Francis Schrag grapples with measurement; a topic that is important in education, which is becoming more widespread and more used, and therefore even more important. As he observes, the verdict concerning measurement in education is mixed; some are for and some are against. I like and agree with Schrag's balanced approach and conclusion: measurement is good for some things but not for others; it has a role to play but it also has its limitations. My contribution will be to further complicate the issue. Some of these complications will entail problematization of some of the assumptions Schrag identifies as underlying measurement.

Let us begin with the problem of what measurement *is*, since that is intimately connected to what it can and cannot do for us. As Schrag observes, the literature is huge, several definitions and positions are available. He cites two, I will contribute a third one. Education is full of concepts that are not directly observable, such as motivation, understanding, intelligence, personality, Bildung, learning. We all use such terms to hypothesize about various unobservable entities and constitents of the world. How do we get access to them so that we can study them? Simply put, we decide what should count as *indicators* of the term (not every observable sign counts as an indicator). Thus, a leading handbook in education research methods defines measurement as the process of linking concepts to indicants.[1] This is a broad definition which conspicuously and purposefully leaves out any mention of numbers. Accordingly, anything can be measured, including sex (we assign to the variable sex the values m, f, other …). *But*: to make sensible use of measurements, you have to

PHILOSOPHY OF EDUCATION 2019 | Kurt Stemhagen, editor

evaluate their quality and adequacy, and one thing to look at is concept/construct validity—do we measure what we think we measure? Construct validity is the agreement between the theoretical definition of the concept in question, say intelligence, and the indicants we have chosen. Are the indicants the rights ones? Do they cover the definition? Too many? Some missing? Quantification does not really come into it—measurements can be made and expressed with or without numbers, but the validity problem remains the same. So also in the case of the MCAT: clearly it runs on the assumption that it is valid, that it actually measures what it is supposed to measure. This assumption can be added to Schrag's list.

But things are philosophically complex here. Roughly, there are two views on the ontological status of theoretical terms, realism and anti-realism, and the one you adopt has thoroughgoing effects for what you think about measurement.[2] For anti-realists (e.g., the positivists), theoretical terms are "derived" talk, a sort of short-hand language, for observation. No construct validity problem occurs because the theoretical term is thought to be identical to the proposed indicants. Intelligence becomes what is measured on intelligence tests, nothing more, nothing less. If you are a scientific realist, you think that terms such as Bildung, understanding, anxiety, etc. refer to (existing) inner processes in persons, and the issue of construct validity immediately arises. How can I be sure that the indicators are true indicators of the construct? I guess that most practicing educational researchers are realists; the constructors of the MCAT probably believe that their test measures some inner quality/ability. And I think we can add a couple of more assumptions: that the quality measured not only belongs essentiallty to the individual (it is not contextual), but that it is stable and can be counted on to remain so.

As a corollary: a brief comment on assumptions concerning the nature of knowledge, given its measureability. This has been the focal

point of big discussions among British philosophers of education. I specifically mention Andrew Davis here as relevant to Schrag's second assumption.[3] Davis argues that assessment standards amount to prioritizing consistency in testing over what you really want to test, which is "rich knowledge and understanding." Rich knowledge is irreducibly holistic in nature, Davis claims. It consists of a large web of knowledge and beliefs that hang together and are dependent on each other. Aiming at a reliable, consistent, system of assessment for large groups of students is logically incompatible with the assessment of rich knowledge and must be at the expense of the validity of what one aims to assess. Testing necessarily only dips into a tiny part of this, and nationwide comparisons force schools to focus on separate, superficial bits and pieces. This speaks to the problem of what knowledge and beliefs (understanding, attitudes, abilities, capacities) *are* that they can be meaningfully measured (that is, their measurability), and to what you think your measurement results tell you about the nature of the inner quality in question.

The MCAT test is a screening test with a multiple-choice format. While screening tests may use right and wrong answers, this is no assumption of other forms of measurement, and Schrag is not right to generalize the assumptions of screening tests to all forms of measurement. Take, for example, measurement in randomized controlled trials (RCTs), where you run an intervention and then compare output measures across study group and control group. This is a mapping of results, not a counting of right and wrong answers. RCTs are quantitative and express results in sometimes very sophisticated statistical terms. But the use of numbers does not per se constitute the problem of measurement. Numbers concern accuracy. In passing, the use of numbers in result reports might make the results look more objective and undisputable. But surely objectivity does not reside in numbers, but in the possibilities for replication, for independent evaluation, in the general care that goes

DOI 10.47925/75.2019.501

into the gathering and handling of data. The problem—the link between our indicants and the concepts we purport to study—remains whether you use numbers or not.

RCT measurements are good for highlighting another issue, namely the role of context and design and what kind of inferences your measurements allow you to make, and about whom. RCTs, because of the research design, allow you to make strong inferences about the cause of some measured outcome in the study group. But this comes at the expense of generality. RCTs show what works in the study group, and *only* that, Nancy Cartwright and Eileen Munro argue; they do not tell you what works in general.[4] So the measurements have a narrow range of application. It is unclear who else such result might apply to; if "similar" children or situations are recommended, just how similar should they be and in what respect?

Measurements abound. One worry is their quality and validity, another is their subsequent use (or abuse). What do measurement results tell you? About your inner qualities? About the role of contextual factors? What you should study at university? Do they diagnose your problem? Will they tell you anything about possible future achievement? RCT results do not tell a teacher what to do in her own classroom, for example. There exist statistics for practically every aspect of human life. Today many new simple tests are devised so that parents and professional staff can make judgments about the normality of a child—the individual measurements (from circumference of head at birth to understanding of numbers at age four) are seamlessly absorbed into vast statistics showing where the child stands as compared to average values. Discrepancies are likely to make parents worry—worries that by and large are completely unnecessary, but which may trigger interventions to cure problems that will likely disappear by themselves given time. In my country the PISA

results (measurements of student achievements) have had large impact on schools.[5] The PISA measurements shape the public image of the school. They teach us to think that schools are mediocre and that competition is the remedy. The PISA results are the premises for today's policy and curriculum planning.

So, measurement is one thing. Using it wisely is quite another.

1J.P. Keeves, ed., *Educational Research, Methodology and Measurement: An International Handbook (Advances in Education)* (Oxford: Pergamon Press, 1988).
2 Stephen Norris, "The Inconsistencies at the Foundation of Construct Validity Theory," in *Philosophy of Evaluation,* ed. Ernest R. House (San Francisco: Jossey-Bass, 1983).
3 For example, Andrew Davis, "Criterion-referenced Assessment and the Development of Knowledge and Understanding," *Journal of Philosophy of Education* 29, no. 1 (1995): 3-21.
4 Nancy Cartwright and Eileen Munro, "The Limitations of Randomized Controlled Trials in Predicting Effectiveness," *Journal of Evaluation in Clinical Practice* 16, no. 2 (2010): 20-266.
5 Svein Sjøberg, "PISA-syndromet. Hvordan norsk skolepolitikk blir styrt av OECD" [The PISA syndrome. How the OECD governs Norwegian educational policy], *Nytt Norsk Tidsskrift* 31, no. 1 (2014): 30-43.

DOI 10.47925/75.2019.501