Studies in History and Philosophy of Science xxx (2017) 1–7

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

Clinical outcome measurement: Models, theory, psychometrics and practice

Leah McClimans [a,*], John Browne [b], Stefan Cano [c]

[a] Department of Philosophy, University of South Carolina, Columbia, SC, USA
[b] Department of Epidemiology and Public Health, University College Cork, Cork, Ireland
[c] Modus Outcomes, Letchworth Garden City, UK

ARTICLE INFO

Article history:
Received 9 December 2015
Received in revised form
1 June 2016
Available online xxx

Keywords:
Classical test theory
Health outcomes
Measurement
Models
Psychometrics
Rasch measurement theory

ABSTRACT

In the last decade much has been made of the role that models play in the epistemology of measurement. Specifically, philosophers have been interested in the role of models in producing measurement outcomes. This discussion has proceeded largely within the context of the physical sciences, with notable exceptions considering measurement in economics. However, models also play a central role in the methods used to develop instruments that purport to quantify psychological phenomena. These methods fall under the umbrella term 'psychometrics'. In this paper, we focus on Clinical Outcome Assessments (COAs) and discuss two measurement theories and their associated models: Classical Test Theory (CTT) and Rasch Measurement Theory. We argue that models have an important role to play in coordinating theoretical terms with empirical content, but to do so they must serve: 1) as a representation of the measurement interaction; and 2) in conjunction with a theory of the attribute in which we are interested. We conclude that Rasch Measurement Theory is a more promising approach than CTT in these regards despite the latter's popularity with health outcomes researchers.

1. Introduction

One thread in the contemporary literature in philosophy of measurement emphasizes the role that models play in measuring outcomes. With some notable exceptions (Boumans, 2015), this discussion has proceeded largely within the context of the physical sciences (Mari, 2000). Models, however, also play qualitative and quantitative roles in psychology. Our interest in this paper is with the methods used to develop instruments that purport to quantify health phenomena, specifically Clinical Outcome Assessments (COAs). The umbrella term for the methods used to develop these instruments is "psychometrics".

In this article, we discuss two psychometric theories and their associated measurement models: Classical Test Theory (CTT) and Rasch Measurement Theory. We argue that models have a role to play in coordinating theoretical terms with empirical content. To play this role, models must: 1) serve as a representation of the measurement interaction; and 2) in conjunction with a theory of the attribute of interest, e.g., one that supplies theoretical quantity terms, explain relationships among theoretical terms. We further argue that the Rasch Model provides a representation of the measurement interaction, while the CTT model does not. In the context of COAs, both measurement theories generally fail to utilize an attribute theory. Despite this failure, and CTTs popularity in health outcomes research, we conclude that health researchers should explore the use of Rasch Measurement Theory.

2. Philosophy, models and classical test theory

2.1. Measuring time

Although physics enjoys more powerful explanatory theories than psychology (Borsboom, Mellenbergh, & van Heerden, 2004; Taagepera, 2008), physics can provide a useful baseline for thinking about the role of models in coordinating measurement. We provide such a baseline with an exploration of the application of theoretical and statistical models in making inferences about the relationship between measurement indications and measurement outcomes in the context of time (Tal, 2011). Similar to psychological constructs, such as intelligence and physical functioning, time is not observable. Moreover, the definition of the unit of time is ideal. The 'second' is defined as the duration of exactly 9,192,631,770 periods

* Corresponding author.
E-mail addresses: mccliman@mailbox.sc.edu (L. McClimans), j.browne@ucc.ie (J. Browne), stefan.cano@modusoutcomes.com (S. Cano).

http://dx.doi.org/10.1016/j.shpsa.2017.06.004
0039-3681/

of the radiation corresponding to a hyperfine transition of cesium-133 in the ground state (BIPM (2014). No actual cesium atom ever satisfies this definition, nor do we have a complete understanding of what it would take to satisfy it.

What we do have are "realizations", i.e., atomic clocks, or more accurately cesium fountains, that approximately satisfy the definition of the 'second' (Tal, 2013). In these clocks, cesium atoms are funneled down a tube where they pass through radio waves. During this process, the atoms' electrons move between two specific energy levels. The frequency of the radiation released when the electrons transition can be used as the basis for measuring duration (similar to the swinging of a pendulum). But, unlike the atoms in the ideal definition, real atoms are subject to extrinsic influences that result in measurement uncertainty and bias. In order to "realize" the referent of the definition of the second, metrologists must model their clocks according to the individual sources of uncertainty and bias they take to affect them.

In discrete steps, metrologists identify ways that the cesium fountains systematically diverge from the theoretical ideal. The example that Tal (2011) provides is gravitational redshift. The definition of the standard second assumes that cesium is in a flat space-time, i.e., gravitational potential of zero. Primary standards, however, exist on earth where the gravitational potential is greater than zero. General relativity theory predicts that the cesium frequency will be red-shifted depending on the altitude of the laboratory where the particular primary standard is located. Redshifts thus indicate measurement bias. The de-idealization process provides a magnitude for the predicted redshift, this correction plus an estimate of uncertainty is added to the primary standard's outcome. This de-idealization process is considered adequate when: 1) the outcomes of a primary standard converge on the outcomes from the other standards within the uncertainties ascribed to each clock, and; 2) the ascribed uncertainties are derived from appropriate theoretical and statistical models of each realization (Tal, 2011).

This example from physics illustrates how models contribute to the coordination of the mathematical formalism of the second as a unit of time with its related empirical content. In the de-idealization process models function to apply theoretical terms to a specific primary standard, e.g., gravitational red shift to the altitude of a particular clock. In serving this function these models represent the measurement interaction, i.e., a representation of the phenomenon being measured, the instrument measuring it and the environment in which the measurements take place (Tal, 2016). But, in doing so, the model serves in conjunction with theory, in this case general relativity theory, which provides the theoretical terms and motivation for the de-idealization.

2.2. Clinical Outcome Assessments (COAs) and classical test theory (CTT)

COAs refer to measuring instruments—typically in the form of a questionnaire—that can be influenced by human choice, judgment or motivation. Thus they are measuring instruments in which people (i.e., patients, clinicians or observers) provide the data that become the measurement indications, e.g., categorical judgments to yes/no questions. The measurement outcomes from these instruments are used to support evidence of the impact of disease or treatment benefit. But how do researchers obtain measurement outcomes from the measurement indications? Put differently, how do researchers justify the inference from the data patients, clinicians or observers provide to outcomes that express a knowledge claim about the quantity of interest, e.g., physical functioning?

The answer to this question differs depending on the measurement theory that researchers use to analyze empirical data (Wilson, 2013a,b). The dominant measurement paradigm in COA development is classical test theory (CTT) (Borsboom, 2006; Cano & Hobart, 2011). CTT embodies three ideas from early in the twentieth century, first, the recognition of error in measurement, second, the conception that error can be conceptualized as a random variable and third, the notion of correlation and how to index it. The development of CTT begins in 1904 with Charles Spearmans' demonstration of how to correct a correlation coefficient due to measurement error and reaches maturity with Melvin Novick's discussion in 1968 (Traub, 1997). CTT turns on a simple model where an observed score (O), i.e., the empirical data acquired after someone fills out a questionnaire, is equal to a person's true score (T) plus uncertainty, commonly termed random error (E), thus $O = T + E$.

When using CTT the value of the true score is taken to be a theoretically unknown value which is assumed to be constant, and the observed score is assumed to be a random variable which produces a bell-shaped curve around the true score. The error score is taken to have an expectation value of zero. The idea here is that as the number of observations, i.e., administrations of the questionnaire, increases, the random errors will tend to cancel one another out, thus the mean of the observations is taken as an estimate of the true score. To acquire an empirical value for T in the context of COA a person must be measured repeatedly on a scale (fill out the items of a questionnaire) and each observation (individual items or repeated administration of the same questionnaire) must be independent of the others (Hobart & Cano, 2009).

In some contexts, CTT makes sense. Borsboom (2005, pp. 14–5) provides an example from astronomy. Imagine that we want to locate the position of a planet that is far enough away that its position can be considered constant. We take multiple careful measurements, but they do not yield identical results. We can interpret the deviations in measurements as random error, the result of weather, shaky hands, etc. Moreover, in this context such measurements usually produce a bell-shaped curve around the true score. But in the context of the behavioral sciences CTT makes less sense.

First, unlike the position of a distant planet repeated administrations of a questionnaire are not independent of one another. Respondents remember the questions from previous administrations and reevaluate their answers in light of them. Second, COAs do not function as a "series of repeated measurements" (Borsboom, 2005, p. 15), rather they function as "measurements on a single occasion" (Borsboom, 2005, p. 15). In a series of repeated measurements the true score should remain the same from one administration of the questionnaire to another. But COAs do not function in this way. Apart from the fact that respondents will remember their answers from previous questionnaires, it is also the case that patients' health can change over the course of administrations of the questionnaire. Third, the interpretation of the observed score as an estimate of the true score significantly depends on the assumption of a continuous variable, e.g., distance, with a normal probability distribution. But many of the variables in the context of COAs are categorical rather than continuous, as the responses elicited from respondents to individual questions can only take a limited number of values (e.g., strongly agree, agree, disagree, strongly disagree).

These difficulties, as well as others, are well known (Borsboom, 2006; Cano & Hobart, 2011). Typically, the first two are managed through a thought experiment: imagine the person filling out the questionnaire is brainwashed in-between a series of administrations (Lord & Novick, 2008). This thought experiment renders administrations of a questionnaire independent of one another and enables us to interpret the administrations as a series of measurements. The third difficulty is often dealt with by simply

Please cite this article in press as: McClimans, L., et al., Clinical outcome measurement: Models, theory, psychometrics and practice, Studies in History and Philosophy of Science (2017), http://dx.doi.org/10.1016/j.shpsa.2017.06.004