On Trial: the Compatibility of Measurement in the Physical and Social Sciences

On Trial: the Compatibility of Measurement in the Physical and Social Sciences

S J Cano[1], T Vosk[2], L R Pendrill[3], A J Stenner[4]

[1]Modus Outcomes, Spirella Building, Letchworth Garden City, SG6 4ET, UK, stefan.cano@modusoutcomes.com; [2]8105 NE 140th Pl, Kirkland, WA 98034, tvosk@comcast.net; [3]SP Technical Research Institute of Sweden, Metrology, Box 857, SE-50115 Borås, Sweden, leslie.pendrill@sp.se; [4]MetaMetrics, Inc., 1000 Park Forty Plaza Drive, Suite 120, Durham, NC 27713, USA & The University of North Carolina, Chapel Hill, North Carolinam USA, jstenner@Lexile.com.

Abstract

In this paper, we put social measurement on trial: providing two perspectives arguing why measurement in the social and in the physical sciences are incompatible and counter with two perspectives supporting compatibility. For the case 'against', we first argue that there is a lack of definition in the social sciences. Thus, while measurement in the physical sciences is supported by empirical evidence, calibrated instruments, and predictive theory that work together to test the quantitative nature of properties, measurement in the social sciences, in the main, rests on a vague, discretionary definition of measurement that places hardly any restrictions on empirical data, does not require calibrated instruments, and rarely articulates predictive theories. The second argument for the case 'against' introduces the problem associated with psychometrics, including different approaches, methodologies, criteria for success and failure, and considerations as to what counts as measurement. Making the first case 'for', we highlight practical principles for improved social measurement including units, laws, theory, and metrology. The second argument 'for' introduces the exemplar of the Lexile Framework for reading that exploits metrological principles and parallels the paths taken by, for example, thermometry. We conclude by proposing a way forward potentially applicable to both physical and social measurement, in which inferences are modelled in terms of a measurement system, where specifically the output of the instrument in response to probing the object ('entity') is a performance metric, i.e. how well the set-up performs the assessment.

1. Introduction

In the social sciences, measurements are frequently made using rating scales, questionnaires or ability tests, which are types of 'instruments', but different from the devices, such as voltmeters, used for measurement in physics and engineering. This is because the variables, or constructs, 'social measurement' seeks to capture (e.g., educational ability, psychological function, cognitive function, disability, and fatigue) are not generally quantifiable using the instrumentation familiar to physical scientists [1]. In contrast to the access to directly observable variables physicists and engineers are often presumed to have, social scientists are typically seen as inferring their measurements from manifestations of hidden variables assumed to be observable only indirectly [2]. In this paper we put social measurement on trial: providing two perspectives arguing why measurement in the social and in the physical sciences are incompatible and counter with two perspectives supporting compatibility. We conclude with a proposal of a potential way forward. It is important to flag that, we do not include discussions surrounding model-based theories (e.g., see references in [3]), "replication crisis" in the social sciences [4], or statistical models used in complex hierarchical systems, as these are beyond the scope of this article.

2. The Case 'Against' Part 1: A Lack of Definition

Just as there are many instruments[a] measuring physical properties in the physical sciences, there are many instruments purporting to measure educational, psychological and health-related variables in the

[a]Measuring instrument: "device used for making measurements, alone or in conjunction with one or more supplementary devices" (§3.1 VIM)

social sciences [5]. However, there are crucial differences between the physical and social sciences in what they require from their instruments. Historically, psychological effects initially entered into measurement science in the field of psychophysics in the 19th Century in attempts to relate abstract human sensations to quantifiable physical external stimuli (such as touch pressure, and sound pitch). Psychometrics developed thereafter to include other mental attributes (such as attitude, knowledge, empathy) which are not simply responses to physical stimuli, and are termed 'latent traits' [6].

While measurement in the physical sciences is supported by empirical evidence, calibrated instruments, and predictive theory that work together to test the quantitative nature of properties, measurement in the social sciences is, in the main, made possible only by a vague, discretionary definition of measurement that places hardly any restrictions on empirical data, does not require calibrated instruments, and rarely articulates predictive theories [7]. The key requirement of a scientific theory is to make predictions that can be tested empirically and that can be relied upon to be reproducible in the future [7,8]. This applies to theories relating different variables as much as it applies to the theory that a particular variable exists as a quantitative attribute. However, in the social sciences, the proposal of a latent variable is hardly ever framed as a theory, the hypothesis of a unit amount is rarely tested, and very few predictions are made that would indicate the quantitative character of the attribute.

Despite a long history of clearly grounded and well-articulated measurement theory and practice [7], measurement in the social sciences is usually completely detached from the ontological claim of a quantitative latent variable. Instead, Stanley Smith Stevens famously proposed that measurement be defined as 'the assignment of numerals to objects or events according to rules' [9]. However, Stevens did not specify the rules explicitly but referred to permissible transformations: 'In what ways can we transform its values and still have it serve all the functions previously fulfilled?' [9; p. 680]. In practice, this reasoning results in a circularity as the scale level implies permissible transformations, which, in turn, determine the scale level [10]. In case of metric scales, numerals are interpreted as numbers that 'represent aspects of the empirical world' [9; p. 677].

The first case against the social and physical sciences being compatible is that measurement in the latter is supported by empirical evidence, calibrated instruments, and predictive theory that work together to test the quantitative nature of properties. In contrast, measurement in the social sciences, in the main, rests on a vague, discretionary definition of measurement that places hardly any restrictions on empirical data, does not require calibrated instruments, and rarely articulates predictive theories.

3. The Case 'Against' Part 2: The Problem of Psychometrics

The methodology underpinning social measurement is known as psychometrics. There are three major psychometric paradigms[b]: Classical test theory (CTT [11]), item response theory (IRT [12]), and Rasch Measurement Theory (RMT [13]). For nearly a century, the dominant paradigm has been the traditional psychometric methods that are underpinned by CTT. This theory of measurement stemmed from seminal works of scientists including Galton, Spearman, and Pearson [7]. However, the last half-century has seen the development of psychometric methods that have refocused the way we can and should measure. These methods (e.g., RMT) have significant conceptual and empirical advantages over traditional approaches [14]. However, social measurement researchers have been slow to accept, embrace, understand, and apply them in their work [7].

We propose three central challenges to social measurement and psychometric methods. The first challenge is to arbitrate the differences between the three main psychometric paradigms (i.e., CTT, IRT, RMT). Each prioritizes paradigmatically incommensurable approaches, criteria for success and failure, and considerations as to what counts as measurement. Might there be scientific, mathematical, aesthetic, moral, and practical reasons why one paradigm is superior to the others? The second challenge, and in part a sequela of the first, is to determine whether there are any practical methods for evaluating in common terms the quality of the many thousands of measurement instruments in the social sciences.

[b] It could be argued that the rarely-used, but historically important Guttman paradigm, is a fourth paradigm.

Third, irrespective of the psychometric paradigm chosen to develop or evaluate instruments, statistical analyses alone do not and cannot inform us as to what is being measured. Statistical adequacy does not guarantee valid measurement. As such, psychometric statistics can be misleading when considered in isolation, and cannot be expected to produce consistently meaningful results when considered apart from qualitative scale content evaluations.

To sum up the second case against, measurement, whether physical or social, to be compatible ought to have similar definitions (i.e., the ratio of two magnitudes of the same *thing* in which the denominator is the unit, providing for invariant comparisons) and the same broad goal (i.e., quantification of meaningful variables). Social measurement neither aspires to nor approximate physical measurement [15], probably because it is not achievable [16]. However, the role of social measurement methods in high-stakes decision making for education, psychology, and health, highlights the need for the field to advance. In particular, there is an urgent need for: 1) testable psychometric theories; 2) testable construct theories, and 3) an experimental hypothesis-testing approach. In the two cases for compatibility which follow, we pick up on these issues.

4. The Case 'For' Part 1: Practical Principles for Improved Social Measurement

Though a large number of significant contributions to more fully scientific measurement for psychological and social studies have been made over the last several decades, they have not yet cohered into an identifiable paradigm with an associated community of practice. Advances fall under four primary headings: units, laws, theory, and metrology.

4.1. Units

In physics, units are fixed at the same size along the entire range of variation, they are defined linearly as a ratio to a larger magnitude, they are invariant across instruments and samples, and reliability is typically conceived in terms of individual measures' uncertainty and precision. In the social sciences, fixed unit sizes, ratio definitions, linearity and individual precision estimates have been achieved, but are not typically a priority. Units in the social sciences usually reflect the ordinality of the underlying scale, vary nonlinearly across instruments and samples, with reliability conceived in terms of group-level correlations. Contrary to popular opinions concerning the supposedly deterministic structure of concrete entities in physical nature, Newtonian physics, elementary number theory, and arithmetic are now recognized as involving irreducible stochastic complexity [17,18]. This parallel motivation for stochastic models across the physical and social sciences notwithstanding, significant challenges must be met to improve the quality of units in the social sciences, not least of which involve improved methodological training and standards.

4.2. Laws

In his work in the 1950s on the measurement of reading ability, Rasch intentionally structured his measurement models so they would have the same form as Newton's laws of motion, following the method first definitively articulated by Maxwell [19-21]. Maxwell's method takes advantage of the fact that a large number of natural laws take the same form, with ratios or products of any two of three parameters equal to the third [4]. The Rasch Reading Law then relates reading ability to text complexity and comprehension rates in the same form as the Combined Gas Law [22].

The goal, however, is not a mere mechanical and superficial copying of parameters from one domain to another. Instead, Maxwell focuses on interactions, with the aims of 'constructing, modifying and merging source analogies in light of the constraints of the problem domain under investigation' [19]. Detailed construct theories, and not just model parameters, are needed.

4.3. Theory

"What I cannot create, I do not understand." This statement from Richard Feynman [23] frames a perspective on validity that focuses on the capacity to recreate or synthesize the construct in the laboratory. This fundamental criterion must be met before claims to understand a construct enough to make valid statements about it are warranted. As was the case with units and laws, there are marked

contrasts between physics and the social sciences in the ways theory is used and developed. In physics, predictive control over constructs makes theory-based measurement highly efficient and effective, mass-produced instruments are not calibrated on data but in terms of theoretical properties, and constant improvements to theory drive down costs and enhance precision in primary associated industries.

In the social sciences, predictive control over constructs has been demonstrated [22, 24], but has had little impact. Further, widely-used instruments are almost always calibrated on data, high stakes test items are increasingly expensive, and a general lack of effective theory undercuts improvement efforts, allowing costs to constantly spiral higher in education, health care, social services, etc. When item calibrations and measures derived from them can be predicted from theory, the kind of reasoning processes and linguistic structures involved in all discourse and applicable in a wide variety of disciplines can be employed [25]. Instead of seeing measurement, modelling, and science as derived from physics, it may prove more accurate to see physics, the social sciences, and all other fields of inquiry as fundamentally rooted in the same kinds of cognitive processes.

4.4. Metrology

Scientific measurement is achieved with simplicity, elegance, and parsimony when an invariant, portable unit is defined via a mathematical law, predicted from theory to a useful degree of precision, and efficiently deployed in practical applications as a universally uniform metric traceable to a reference standard. In the social sciences, in contrast, units are not traceable to reference standards; communities of stakeholders do not set consensus unit standards, no resources are spent setting and maintaining uniform unit standards; and education, health care, and social service markets are accordingly highly inefficient due to the very high transaction costs incurred. Established successes with additive units, conformity to scientific laws, and predictive theory suggests metrology is feasible and would be highly desirable in the fields employing social measurement methods.

We would proffer that the most important and pressing issue for the social sciences is need for theory. And thus, to begin to employ model-based reasoning in social measurement we must adopt a more thoroughly experimental approach [26]. To that end, two key requirements are needed to advance our understanding of what instruments measure in the social sciences: explicit theories of the constructs; and explicit methods to test those theories. As such, instrument development should be from the ground up (i.e., clear definitions), rather than the traditional top down approach (i.e., arbitrary statistically-driven methods of grouping items). This necessitates the complementary roles of qualitative and quantitative methods to ensure that construct definitions and subsequent theory determines instrument content, and psychometric analyses are used to assess the validity of the resultant construct theories [27].

When item difficulties and by implication estimates of person measures are under control of a construct theory, it becomes possible to develop a construct specification equation [28]. This is the formal expression of the theory regarding the observed regularity in a set of observations generated by a given measurement process. Therefore, the equation articulates a theory of item-score variation and simultaneously provides the vehicle for confirmation or falsification of the theory. Construct specification equations are developed by regression analysis of item locations on selected item characteristics. They provide a test of fit between scale-generated observations and theory. In essence, the greater the proportion of variation in item location explained by the selected item characteristics, the greater the support for the proposed construct theory, the greater the evidence for scale validity, and the more meaningful the interpretation of person measures [29].

In summary, the first case for the compatibility of measurement in the social and physical sciences is based on the widely used practical principles for improved definition in social measurement's units, laws, theory, and metrology. An exemplar is presented in the second supportive case.

5. The Case 'For' Part 2: The Exemplar of Reading Measurement

Reading is a process in which text and reader act together to produce meaning. Reading provides the most advanced form of theory-referenced measurement [28, 29]. The central premise is to shift the

focus from studying people, which is the usual focus in social measurement, to studying people relative to items. Five levels of evolving construct theory culminate in a causal model relating reader ability, text complexity and comprehension. The model is conjoint, enabling reader ability and text complexity to be denominated in a common unit and denoted as a numeric value typically ranging from 300L to 2000L [30].

It is rare in the social and behavioural sciences to find instrumentation that exploits metrological principles and copies the paths taken by, for example, thermometry [31]. The 30-year history of developments in reading measurement exemplifies an attempt to do just that. As such, a key feature of causal Rasch models [32] is that they take a form similar to some three variable physics equations like D = RT or F = MA. A crucial feature of such equations is the trade-off or cancelation property, which states that, for example, a change in mass can be traded-off for a change in acceleration to hold force constant or a change in rate can be cancelled by a change in time to hold distance constant. Michell [33] has argued that successful trade-offs and cancelations like those above are evidence for the quantitative status of attributes figuring in such equations. To sum up the second argument for the compatibility of measurement in the social and physical sciences, reading measurement stands as the exemplar of the evolution of measurement in social science being played out, and provides a potential template for constructs beyond 'reading ability'.

6. Conclusion: A Way Forward

As increased stringency is required, for instance in measurements in education, healthcare and many contemporary applications where human perception is key, one seeks the same kind of quality-assurance of social measurements as is established in physics and engineering [34, 35]. Specifically, this means assurance with objective metrological comparability ('traceability') and declared measurement uncertainty, since the decisions that need to be made about student ability or patient health should be comparable with known uncertainties, based on sound measurements. In seeking this stringency, we should at the same time be realistic, acknowledge limitations in quantifying social measurement, and avoid suffering from what has been called 'physics envy' [35].

In considering the philosophical foundations of social measurement, Maul et al. [36] recall various approaches, including empiricism, pragmatism and realism. The philosophical realism behind physical metrology assumes, as in physics, that there is an objective reality, which exists even when we do not perceive or have instruments to measure it. One might argue that there is seldom objective reality in what is measured in social science (e.g., the challenge of a task) without our actually perceiving or measuring it. Lacking an independent objective reality might lead to measurements in the social sciences providing no unique 'right' answer. This makes metrology in social measurement challenging, since: (i) independent reference standards used in metrology to ensure the comparability of different measurements would be difficult to establish separately from the actual measurement process, and (ii) measurement uncertainty would often be very large, since each new measurement set-up would produce definitions divergent from others.

The argument that **empirical operationalism** is irreconcilable with general scientific practice and vocabulary [38] (since each unique set of operations would be associated with distinct definitions) does not hold, since in physical and engineering metrology it is accepted[c] that 'a reference can be a measurement unit, a measurement procedure, a reference material, or a combination of such.'

7. References

[1] Thorndike E L 1904 *An introduction to the theory of mental and social measurements* (New York: The Science Press)
[2] Thurstone L L 1928 *Am J Sociol* **33** 529-554
[3] Pendrill L, Fisher W 2015 *Measurement* **71** 46–55

[c] VIM 2015, *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms* , §1.1 Note 2, http://www.bipm.org/en/publications/guides/

[4] Open Science Collaboration Estimating the reproducibility of psychological science. *Science*, 349 (6251), aac4716, 2015.
[5] Kempf-Leonard K 2005 *Encyclopedia of Social Measurement* (Oxford: Elsevier)
[6] Tesio L 2003 *J Rehabil Med* **35** 105-115
[7] Wright B 1997 *Educ Meas: Issues Practice* **16** 33-52
[8] Popper K 1959 *The Logic of Scientific Discovery* (London, Routledge)
[9] Stevens S S 1946 *Science* **103** 677-680
[10] Duncan O D 1985 *Synthese* **42** 21-34
[11] Nunnally J C, Bernstein I H 1994 *Psychometric Theory*, (New York: McGraw-Hill)
[12] Lord F M, Novick M R 1968 *Statistical theories of mental test scores* (Reading, Massachusetts, Addison-Wesley)
[13] Rasch G 1960 *Probabilistic models for some intelligence and attainment tests* (Chicago, University of Chicago Press)
[14] Lord F, Novick M 1968 In: Lord F, Novick M, eds. *Statistical Theories of Mental Test Scores*, (Reading, Mass.: Addison-Wesley) pp. 13-26
[15] Michell J 1997 *Br J Psychol* **88** 355-383
[16] Michell J 1999 *Measurement in Psychology* (Cambridge, UK: Cambridge University Press)
[17] Chaitin G 1994 *International J Bifurcation Chaos* **4** 3-15
[18] Chaitin G J 2003 *The Limits of Mathematics* (London: Springer-Verlag)
[19] Nersessian N 1996 *Philos Sci* **63** 542-546
[20] Nersessian N 2002 In: Malament D, ed. *Reading Natural Philosophy* (Lasalle, Illinois: Open Court) pp. 129-166
[21] Fisher W 2010 Rasch, Maxwell's method of analogy, and the Chicago tradition. *Probabilistic Models for Measurement in Education, Psychology, Social Science and Health: Celebrating 50 years Since the Publication of Rasch's Probabilistic Models*, University of Copenhagen School of Business, FUHU Conference Centre, Copenhagen, Denmark, June 13-16, 2010.
[22] Burdick D, Stone M, Stenner A 2006 *Rasch Measurement Transactions* **20** 1059-1060
[23] Nersessian N 2008 *Creating Scientific Concepts* (Cambridge, Massachusetts: MIT Press)
[24] Stenner A, Stone M 2010 *J Appl Meas* **11** 244-252
[25] Fisher W, Stenner A 2011 *International J Multiple Research Approaches* **5** 89-103
[26] Stenner A J, Smith M, Burdick D 1983 *J Educ Meas* **20**, 305-316
[27] Stenner A J, Smith M 1982 *Percept Motor Skills* **55** 415-426
[28] Stenner A J, Burdick H, Sanford E E, Burdick D S *J Appl Meas* **7** 307-322
[29] Hawking S 2001 *The Universe in a Nutshell* (New York: Bantam Books)
[30] Stenner A, Stone M, Burdick D 2009 *Rasch Measurement Transactions* **23** 1204-1206
[31] Chang H 2007 *Inventing Temperature: Measurement and Scientific Progress* (Oxford: Oxford University Press)
[32] Stenner A, Fisher W, Stone M, Burdick D 2013 *Frontiers in Psychology*, **4** 1-14
[33] Michell J 1990 *An Introduction to the Logic of Psychological Measurement* (Hillsdale, New Jersey: Lawrence Erlbaum Associates)
[34] Pendrill L Petersson N 2016 *Measurement Science & Technology* (in press)
[35] Ruhm K 2016 *Measurement*, **79** 276-284
[36] Nelson R 2015 *Issues in Science and Technology* **31** 71-78
[37] Wilson M 2013 *Measurement* **46** 3766-3774
[38] Maul A, Torres Irribarra D, Wilson M 2016 *Measurement* **79** 311-320
[39] Bond T, Fox C 2015 *Applying the Rasch model* (New York, Lawrence Erlbaum Associates)
[40] Pendrill L 2014 *Metrologia* **51** S206