Meeting Feynman: Bringing light into the black box of social measurement

Thomas Salzberger
Senior Lecturer Vienna University of Economics and Business, Vienna, Austria
E-mail: thomas.salzberger@gmail.com

Abstract. The social sciences largely have not achieved successes of the same stature as the natural sciences. This may be due in large part to failures in the theory and practice of measurement. Rather than modelling the mechanism of measurement, which remains a black box, assigned numerals are interpreted as measures. More advanced methods have been available for many years but have been adopted slowly and often been applied inadequately. While some models may indeed shed light on the black box, more elaborate conceptual theories are required also in order to eventually lift social measurement to the level of physical measurement.

1. Introduction
Measurement in the social sciences (social measurement) and measurement in the natural sciences (physical measurement) certainly differ in terms of how measurement is carried out. In physical measurement, humans generally act as observers, while objects interact with measurement instruments directly [1]. In social measurement, humans interact with instruments for what we aim to measure resides in the mind of the human respondent. While modern physical measurement results in measures expressed in traceable units associated with an estimate of uncertainty as the two key elements of metrology, the situation in social measurement is different. Social measurement at first leads to qualitative observations. A typical procedure involves the administration of a questionnaire, consisting of statements (items) associated with a response scale, to the respondent. The questionnaire is a measurement instrument that is to quantify a variable that is characterised by its items. The respondent is assumed to process the wording of the item, recall relevant cognitions, form a mental response, and finally choose the response option that best fits the mental response. The outcome of this process therefore is a comparison of a statement that characterises, or reflects, the variable to be measured with the respondent’s mental positioning with respect to the same variable.

Applying the concept of measurement to this process implies the hypothesis that a quantitative latent variable exists which allows for locating both the constituents of the measurement instrument, i.e. its items, and the subject responding to the instrument, on an underlying quantitative continuum. Response formats usually offer multiple categories, which are thought to represent ordered amounts of the property the items are proposed to assess. The observed response as a choice of one response category is transformed into a numerical score. Thus, the score is an ordinal statistic provided the categories adequately reflect an increasing amount of the variable to be measured.

Traditionally, social scientists have interpreted item scores as measures. In classical test theory (CTT [2]), properties of these item measures are investigated with focusing on means, variances, covariances and correlations. Respondent measures are formed as weighted or unweighted sum scores across items. These measures are essentially justified by observed convergence of individual item measures. With respect to their interpretation and their processing, the measures are treated as if they were on a par with measures in the natural sciences.

From a statistical point of view, the most obvious limitation is that scores that represent order cannot be meaningfully added up to sum scores that are linear, equal-interval measures. In terms of measurement philosophy, a more serious limitation arises from the fact that traditional social
measurement attaches numerals to observations, which represent the outcome of the process of measurement, and treats those numerals as numerical measures. This type of social measurement clearly omits the very processes that are thought to imply measurement. The hypothesised response process is not modelled at all. It remains a black box. Obviously, social measurement does not start with the observation of scores. After all scores are assigned to responses to instruments that first need to be developed. However, a thorough conceptualisation of the variable to be measured and development of the measurement instrument must not mask the fact the social measurement omits the core of the mechanism that yields observations which may allow for inferring measures.

The nonchalance with which social scientists have obscured or simply ignored the Achilles’ heel of their measurement procedures have met with harsh and devastating criticism from natural scientists. In a discussion about pseudoscience, Richard Feynman famously said “social science is an example of a science which is not a science; they don’t do [things] scientifically, they follow the forms … [but] they don’t get any laws, they haven’t found out anything” [3].

2. Bringing light into the black box

Today, the question is can the social sciences do better? As a matter of fact, approaches to model the comparison of items and respondents have been available for many years. Item response theory [4] explicitly refers to this comparison as the fundamental process of social measurement. Nevertheless, the status of measurement in the social sciences in general has improved only marginally. The reasons are twofold. First, the process of adopting these models has been very slow with CTT still being the predominant paradigm in most social science disciplines. Second, IRT models are often applied with the same, or a similar, philosophical mind-set that underpins CTT. Both phenomena reflect the intricacies of a paradigm change. CTT produces measures based on statistical procedures that presume rather justify measurement. By contrast, IRT specifies item and person parameters that represent the two entities that are crucial in the comparison of statements in a questionnaire and the respondent. It is from this comparison observations arise that allow us to infer measures. A fundamental limitation of IRT lies in its disregarding fundamental principles of measurement when it comes to the concrete modelling of the response as a function of item and person parameters. The key requirement in this respect is invariance. Item properties must no depend on the respondents that are instrumental in their estimation; and, vice versa, respondent characteristics must not depend on the specific items used. In the context of the natural sciences, it is evident that characteristics of a measurement device have to be stable and independent of the objects measured. In other words, the instrument and the measurement object are regularly separated, for example when determining the mass of a weight in terms of the calibrated response of a weighing instrument [5].

In the Rasch model (RM; [6]), Rasch accounted for invariance by what he called specific objectivity [7]. Specific objectivity is a property of the Rasch model by virtue of parameter separation (see [8] on the Rasch model for measurement in the social sciences). From a mathematical point of view, the Rasch model is an IRT model. However, since it is unique in terms of providing specifically objective measurement, it defines a class of its own distinct from other IRT models. The fundamental philosophical difference between IRT in general and the Rasch model in particular is the concept of how a measurement model and data are related. The Rasch model is a prescriptive model emphasizing the fit of the data to the model, while IRT models follow the statistical principle of the best possible description of data focusing on fit of the model to the data. This notion is reflected in the term Rasch Measurement Theory (see [8]).

In contrast to non-Rasch IRT models, the Rasch model also adheres to general principles of scientific measurement (cancellation conditions, [9]. It also addresses the scientific task of measurement, [10]) and can be seen as a realisation of additive conjoint measurement [11]).

3. What still remains in the dark

Compared to CTT, measurement based on the RM, brings light to the black box insofar as it models explicitly what we hypothesise to be origin of the observation from which we infer a measure.
Contrary to IRT, the RM specifies fundamental characteristics of measurement and prescribes those to be met by empirical data. Measures estimated under the RM, where data have been demonstrated to fit the RM, i.e. data meet the requirements the model embodies, certainly do have a much better foundation that measures rooted in CTT or IRT using models that disregard invariance.

While the application of the RM may bring more light into the link between the comparison of items and respondents, social measurement based on the RM might, and typically still does, fall short of physical measurement with respect to two aspects. First, the link between the conceptual theory of the variable to be measured and the item characteristics as accounted for in the model is essentially not covered by the model. The conceptual theory of the variable informs instrument development. Empirical response data may confirm the suitability of the instrument for being the basis for measurement. However, no direct, especially no quantitative link exists between the conceptual theory and the item parameters. The latter are calibrated in a data-driven way. Second, the person measures, while meaningfully interpretable with respect to item measures, lack an interpretation in terms of a tangible measurement unit. In fact, both issues are close intertwined. A quantitative theory that hypothesises a mechanism, a wheelwork as it were, that explains how items behave, would also lend meaning to what a difference between items means.

The unit of measurement is not only crucial for achieving and justifying measurement in a concrete instance. It is also very relevant in terms of consolidating different measurements and gaining insights that transcend one particular study. In the natural sciences, the importance of a clearly defined metric in measurement is a universally recognised matter of course. Metrology, the scientific study of measurement, is concerned with the establishment of common units of measurement [1]. Each instrument’s measures must be traceable to a standard unit, which remains stable across the entire continuum [12]. Furthermore, each measure, expressed in a defined metric, must involve a standard error signifying the uncertainty of the estimate [13]. These fundamental principles, entrenched in physical measurement, have slowly started to enter the realm of social measurement [1, 14, 15].

Currently, social measurement widely lacks explicit units of measurement. Apart from the fundamental deficiencies in terms of justifying measures in the first place, CCT yields respondent measures that are percentiles of a distribution assumed to be normal. Hence the unit of measurement is not only implicit but also intrinsically tied to a given population, for which a normal distribution has to hold, and a given context of measurement. Measurement based on the RM disentangles item and respondent characteristics and establishes a unit of measurement that is, in principle, independent of the respondents. However, the implicit unit suffers from two limitations. First, it is not interpretable in the same way as units in physical measurement as it lacks tangibility. Second, even though it does not depend on individual respondents within a given population, it may vary depending on contextual factors [16]. The latter is a problem that can be addressed empirically, even though its implications in terms of comparability of measures may be quite subtle. The former is an even more general problem in social measurement.

4. Conclusion: What needs to be done
The RM plays a crucial role in the improvement of measurement in the social sciences. Its characteristics ensure that fundamental and universal principles of measurement are imposed as prerequisites on empirical data. It paves the way towards achieving measurement that is on a par with measurement in the physical sciences. However, replacing CTT or IRT by the RM will not be enough. The conceptual theory of what it is to measure has to be improved to allow establishing a link between the conceptual theory and the model parameters [17, 18]. This will also contribute to efforts of developing explicit, tangible and interpretable units of measurement. Only then will the social sciences be in a position to fully counter Richard Feynman’s harsh assessment.
5. References

[1] Pendrill, L. R., and Fisher Jr, W. P. (2013). “Quantifying Human Response: Linking metrological and psychometric characterisations of Man as a Measurement Instrument”, Journal of Physics: Conference Series (Vol. 459, No. 1, p. 012057). IOP Publishing.

[2] Lord, F. M, and Novick, M. R. (Eds.) (1968), “Statistical theories of mental test scores”. Reading, MA: Addison-Wesley.

[3] Feynman R 1981 The Pleasure of Finding Things Out (BBC Horizon Interview). Available at: http://www.worldcat.org/wcpa/servlet/DCAResult?standardNo=0738201081&standardNoType=1&excerpt=true; [Sequence on the social sciences as an example of a pseudoscience can be found under https://www.youtube.com/watch?v=tWr39Q9vBgo, accessed March 1, 2018]

[4] Lord, F. M. (1980): Applications of item response theory to practical testing problems. Mahwah, NJ: Erlbaum.

[5] Pendrill, L. R., and Fisher, W.P. (2015). Counting and quantification: Comparing psychometric and metrological perspectives on visual perceptions of number. Measurement, 71, 46-55.

[6] Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests, Copenhagen: Danish Institute for Educational Research, expanded edition (1980) with foreword and afterword by B.D. Wright, The University of Chicago Press, Chicago, 1960.

[7] Rasch, G. (1967). On Specific Objectivity: an Attempt at Formalizing the Request for Generality and Validity of Scientific Statements. Danish Yearbook of Philosophy, 14, 58-93.

[8] Salzberger, T. (2009). Measurement in marketing research: An alternative framework. Cheltenham, UK: Edward Elgar.

[9] Karabatsos, G. (2001), “The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory”, Journal of Applied Measurement, Vol. 2, No. 4, 389–423.

[10] Michell, J. (1997). Quantitative Science and the Definition of Measurement in Psychology. British Journal of Psychology, 88(3), 355-383.

[11] Luce, R.D., and Tukey, J.W. (1964). Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement, Journal of Mathematical Psychology, 1(1), 1-27.

[12] Vim, I. S. O. (2004). “International vocabulary of basic and general terms in metrology (VIM)”, International Organization, 2004, 09-14.

[13] BIPM, I., IFCC, I., & ISO, I. IUPAP, and OIML (2008),“Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement,” Joint Committee for Guides in Metrology. Technical Report No. JCGM 100.

[14] Pendrill, L. R., Emardson, R., Berglund, B., Gröning, M., Höglund, A., Cancedda, A., Qunti, G, Cennìa, F., Rossi, G.B., Drnovšek, J., Gersak, T., Goodman, T., Harris, S., van der Heijden, G., Kallinen, K., and Ravaja, N. (2010), “Measurement with persons: a European network”, NCSLi Measure, 5(2), 42-54.

[15] Pendrill, L., Cano, S., Barbic, S., and Fisher Jr, W. P. (2017). “Patient-centred outcome metrology for healthcare decision-making”, in: Joint IMEKO TC1-TC7-TC13 Symposium: “Measurement Science Challenges in Natural and Social Sciences” Rio de Janeiro, Brazil, from July 31st to August 3rd 2017.

[16] Humphry, S. M., and Andrich, D. (2008). “Understanding the unit in the Rasch model”, Journal of applied measurement, 9(3), 249-264.

[17] Stenner, A. J., Fisher Jr, W. P., Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. Frontiers in psychology, 4, 1-14.

[18] Cano, S., Pendrill, L.R., Melin, J., Köbe, T., Fisher, W.P. (2018). Metrology for the Social Sciences: A case for Rasch Measurement Theory (not Rasch Analysis), IOMW 2018 New York.