JOURNAL OF APPLIED MEASUREMENT, *23*(3/4), 105-112

COMMENTARIES

How to Avoid a Parody of Measurement: Some Models are Wiser Than Others A Commentary on the Pillars of Measurement Wisdom by George Engelhard, Jr.

Thomas Salzberger
WU Wien University of Economics and Business

In discussing the pillars of measurement wisdom, it is probably appropriate to begin by considering what measurement currently means in the social sciences. Measurement of latent constructs in the social sciences is based on an immensely diverse range of approaches and measurement models. At first glance, this richness appears to be a virtue, allowing social researchers to be flexible and versatile. After all, one is free to choose linear (as in classical test theory) or non-linear link functions (as in Rasch measurement theory and item response theory) to model the relationship between the latent variable representing the construct and the manifest responses to stimuli, which are typically items of some sort. One may conceive of causality flowing from the latent variable to the observed responses or vice versa as it is done in index formation (Coltman et al, 2008). One can choose the number of parameters in the measurement model to be estimated (Rasch measurement model vs. general item response theory models; Embretsen & Reise, 2013).

But where does this apparent freedom come from? And is it ultimately to the benefit of the social sciences? The answer to the former question arguably is the reliance of social scientists on a definition of measurement Stanley S. Stevens proposed in 1946. According to this definition, measurement is accomplished by "the assignment of numerals to objects or events according to rules" (Stevens, 1946, p. 677). Stevens' definition ensured that the social sciences could claim to be "proper" quantitative sciences (Michell, 1999). This achievement comes at a hefty price, though. It is hard to imagine a definition that could be more liberal than Stevens'. The concept of measurement in the social sciences has thereby been detached from how our colleagues in the natural sciences define measurement based on Maxwell's systematic approach (McGrane, 2015). In the end, any procedure in the social sciences that

Requests for reprints should be sent to Thomas Salzberger, Wirtschaftsuniversität Wien Vienna, Austria; email: thomas.salzberger@wu.ac.at

generates numbers may maintain that it provides some sort of measurement. The arbitrariness of what measurement can be in the social sciences and the divergence of the concept in the social sciences and the natural sciences raise serious doubts that Stevens' legacy could benefit the social sciences in the long run. If one subscribes to scientific realism and the idea that every proposed latent construct carries with it an ontological claim (Borsboom, 2005) that must be empirically corroborated, Stevens' definition is at odds with the appropriate scientific rigor of quantitative research.

In Stevens' defense, Stevens would probably be very skeptical about many measurements in the social sciences, particularly the scale levels claimed. In addition, the above-mentioned freedom of what constitutes measurement is, in practice, not as unlimited as it might seem. Proper social measurement must be based on a solid conceptualization underlying the stimuli used to elicit responses indicative of a latent variable. After all, measurement is as much a qualitative as it is a quantitative undertaking. The quantitative element is central to any general discussion of measurement, as it is of a general nature and does not address the specific content of a construct. However, a statistical measurement model is essentially an empty shell, a blueprint for how to relate observable outcomes to measurements of latent variables from a purely formal point of view. Thus, selecting an appropriate measurement model that is fit for purpose is a necessary condition for social measurement but needs to be accompanied by a substantive theory of the construct to be assessed. Ideally, the measurement theory and the substantive theory are linked with one informing the other. In any case, in terms of measurement models, Stevens' definition places very little, if any, restrictions on their choice. In the social sciences, measurement is, as it were, a fundamentally "unregulated" matter. As a result, individual researchers vary widely in what they consider appropriate when measuring and, accordingly, which measurement model they prefer.

In a formal sense, measurement models are statistical models that come with different assumptions and have different properties. While some assumptions are empirically testable and should therefore rather be referred to as requirements, others are simply considered given and not addressed empirically introducing an element of speculation. Particularly for researchers who are not experts in measurement theory per se, it is important to be cognizant of the properties of the model they use. While these properties are statistical in nature, the question arises as to whether they are compatible with psychometric properties the researchers deem important or at any rate what the statistical properties imply psychometrically. In other words, we need to thoroughly examine the relationship between given statistical properties and the desired or required psychometric properties of a measurement model. Statistical models are not right or wrong per se. They can describe given data better or worse. However, a good description of the data in the sense of fit between a measurement model and the data is a necessary but not a sufficient requirement for measurement—provided one is prepared to define psychometric properties of measurement that must be met to constitute measurement. Stevens' definition is of no help in this regard. Psychometric requirements must be derived from a conceptualization of measurement that goes beyond Stevens' account of measurement. Once these psychometric properties have been determined, a statistical model accounting for these properties can be chosen. A fit between the statistical model and the data then also implies a fit between the psychometric requirements and the data. The statistical model becomes a prescriptive model and data misfitting the model implies a failure to measure the intended construct (Fisher, 2010). In contrast, in the absence of concrete psychometric requirements, any statistical model can be used; and the model fitting the data has descriptive power but does not allow for substantiating measurement. Certainly, one could invoke Stevens' definition and "safe" the claim that something has been measured. It would not live up to the concept of scientific measurement, though. It would rather be a parody of measurement where any number-generating procedure would be measurement.

The issue can be looked at from two different perspectives. First, one could fundamentally rethink measurement in the social sciences, identify psychometric requirements, and select a statistical model as the appropriate measurement model that specifies the required properties that need to be found in the data. Principals of measurement subscribed to in the natural sciences lend themselves as a starting point. Recently, a stream of research emerged that aims to realign measurement in the social sciences with measurement in the natural science by referring to metrological principles of measurement (Mari & Wilson, 2014). Concepts such as specific objectivity (Rasch, 1977) are also key in this regard. In practice, this approach often does not fall on fertile ground as researchers are not prepared to give up longstanding beliefs that they have held sometimes for decades in favor of new principles they do not deem necessary. Such a shift would be radical indeed, and paradigmatic resistance is to be expected (Andrich, 2002).

Second, one could start at the level of concrete statistical models used in social measurement, explicate their properties, and highlight the psychometric implications. This approach could be thought-provoking and contribute to a better understanding of measurement in the social sciences. Paradigmatic resistance could still be an issue as the relevance of psychometric implications might be questioned.

George Engelhard's (2022) "The Pillars of Measurement Wisdom" essentially tries to incorporate both perspectives. The pillars of measurement wisdom are modeled on Stigler's (2016) seven pillars of statistical wisdom. Thus, Engelhard investigates whether statistical principles can be meaningfully transformed into measurement principles. As such, statistical properties are, in a sense, the starting point. Another research question is whether there are additional pillars relevant to educational measurement. On the other hand, Engelhard discusses the derived pillars of measurement wisdom in the context of Rasch measurement theory (Andrich, 2017), a framework for social measurement that is driven by fundamental principles of measurement (such as invariance or specific objectivity; Andrich, 2017; Rasch, 1977) from which the Rasch measurement model can be derived.

Engelhard (2022) convincingly argues that the seven pillars of statistical wisdom retain their meaningfulness and relevance when transferred to the realm of measurement. To start with, ***aggregation*** certainly lies at the heart of social measurement aiming at estimating a measure of a latent, hence unobservable, variable based on a series of observed responses. Thus, a response pattern is summarized by one number. Whereas in statistics, aggregation might arguably be more relevant in terms of sample description (aggregation across subjects), in measurement aggregation predominantly applies to summarizing data within a subject. Whether classical test theory (CTT) is a prime example of the concept of a latent variable is debatable. At least in its most archaic form, in which an observed score is defined to be composed of a true score and an error score, CTT does in fact not account for a latent variable. Any sum across multiple data points can be decomposed into the two components, whether there is a latent variable involved or not. In practice, researchers very likely believe that a latent variable is, in the end, causing the true score. But these considerations, or beliefs, are external to CTT. They are not addressed in the model. Without an explicit account of a latent variable, CTT simply summarizes a response pattern by its sum. Today, applications of CTT typically refer to the congeneric model (Traub, 1997), which corresponds to factor analysis. This model does explicitly account for a latent variable represented by a factor score, and aggregation may then mean to estimate the factor scores. Nevertheless, the linear relationship between manifest item responses and the latent variable score essentially implies that item scores are considered measures. As

such, CTT is concerned with the behavior of presumed measures but is a questionable tool when it comes to modeling measurement and its accomplishment in the first place. What is more, CTT is predominantly concerned with aggregation across subjects (variances, covariances, correlations). These considerations show that aggregation on its own is not a sufficient criterion of measurement. While the simple score is all what it takes in CTT (in its most basic form, but typically also when referring to latent variable versions of CTT, when sum scores are computed instead of factor score estimates), in the Rasch model, the sum score is an input to measurement, not the endpoint.

Likelihood is certainly at the core of inferential statistics. Since measurement is about inferring a measure from observations, it is evident that likelihood also matters in social measurement. All measurements come with uncertainty, and the concept of probability informs the estimation of uncertainty. But as with aggregation, some models are wiser than others. In CTT, uncertainty may be estimated based on the standard error of measurement, which, counterintuitively, depends on sample characteristics (standard deviation) and the sample-dependent property of reliability, in other words on aggregations across subjects. In the Rasch model, the standard error is based on the properties of the items contained in the measurement instrument, which are sample-independent. In one way or another, the concept of likelihood of data is probably applicable to all measurement models. However, the meaningfulness in terms of requirements of measurement varies widely. The interpretability of measurements based on predicted probabilities of responses to concrete items given a measure of the latent variable in the Rasch model is a good example of how likelihood can be meaningfully applied in measurement.

The concept of ***information*** is crucial since it is instrumental in the estimation of uncertainty in the Rasch model as well as in non-Rasch item response theory (IRT) models. Some observations yield more information for the estimation of a latent variable measure than others. In the Rasch model, the amount of information depends only on the distance between the item and the person estimate, from which the probability of a positive response follows. More information implies less uncertainty. However, information does not depend on the actual response. Engelhard (2022) uses the example of a total score of 3 across 4 dichotomous items (ordered in terms of their difficulty). The Guttman pattern of [1, 1, 1, 0] does indeed have the highest likelihood given a total score of 3, whereas the anti-Guttman pattern [0, 1, 1, 1] has the lowest likelihood. However, the conclusion that measurement uncertainty is highest for the Guttman pattern and lowest for the anti-Guttman pattern is somehow misleading as in all four cases, the total score of 3 is a sufficient statistic for the estimation of the person measure. In principle, there is no information in the specific pattern other than the sum score. Since information only considers the response probabilities depending on the relative distance of the person and the four items (Wright, 1990), the standard error as the inverse of the square root of information is the same in all four cases. Having said that, the likelihood of the observed response pattern is the basis for person fit (Smith, 1986) and, as such, provides, in a sense, "information" about whether or not measurement has been accomplished. An unlikely response pattern might trigger serious doubts, or "uncertainty," about successful measurement for that person. However, this "uncertainty" would be a qualitative conclusion and cannot easily be converted into a range of uncertainty around a measurement estimate. Incorporating the likelihood of the response pattern into an estimate of uncertainty would require a redefinition of the standard error. For example, one might argue that the pattern [0, 1, 1, 1], in an educational setting at least, suggests that failing on the easiest item is a sign of carelessness and that a pattern of [1, 1, 1, 1] would be equally plausible for that person. Then the lower boundary for the uncertainty interval might be based on the pattern [0, 1, 1, 1], while the upper boundary could be based on [1, 1, 1, 1]. In this case, one would not call measurement for that person into question, though, but account for a plausible explanation for the unexpected response pattern. In other cases, perhaps for [0, 1, 0, 1], one would not trust measurement for that subject altogether.

The pillar of **intercomparison** certainly lies at the heart of measurement. A measurement of a property in one subject rarely is informative on its own. There is typically some sort of comparison standard involved. In many applications, a subject is compared to another subject, or multiple subjects are compared to one another including group mean comparisons. For such comparisons, it is crucial that the meaning of measurement remains the same from one subject to another. From this, it follows that the properties of the items used to trigger responses must be stable and independent of the subjects used to estimate them. Vice versa, the estimate of a subject's measure must be independent of the concrete items it is based on provided the items form a unidimensional scale. In physical measurement, everyone would subscribe to these requirements and distrust measurements for which they are violated. In social measurement, Rasch's concept of specific objectivity (Rasch, 1977) formalizes this requirement, which must be empirically demonstrated and is only valid for an empirically determined frame of reference. Besides the requirements of comparisons across different subjects and different items, Engelhard (2022) also discusses resampling methods (such as jackknife and bootstrapping) that allow for determining uncertainty ranges around parameter estimates. While the element of intercomparison is somehow evident (different samples ought to yield an estimate for the same parameter), these methods may be more strongly related to likelihood and perhaps information.

One aspect that could be added in the context of intercomparison, is the issue of a unit of measurement. After all, any measurement expresses, or ought to express, a comparison with a unit. Social measurement persistently struggles with units of measurement. In CTT, researchers typically refer to the labels of the response scale (what is the item mean score of a subject) as factor scores essentially express percentiles in a given population and lend no real meaning to a measurement other than comparisons across subjects. In the Rasch model, the unit of measurement is typically implicit (Briggs, 2019; Humphry & Andrich, 2008) and related to differences in the response probabilities to different items (Ludlow & Haley, 1995). A given difference in two subjects implies a constant log-odds ratio for the two subjects for all items. In IRT models that feature a discrimination parameter, the unit of measurement becomes blurry. The response probability does no longer exclusively depend on a property of the person and the item (i.e., on their locations) as is the case in the Rasch model, but also on the value of the discrimination parameter, which is a property of the sample. As a Rasch measure for a subject can also be interpreted in terms of item properties via response probabilities, the intercomparison of persons and items plays an important role when it comes to measurement interpretation. New developments in the area of construct specification equations (Adroher & Tennant, 2019; Fisher & Cano, 2023; Melin et al., 2021; Stenner et al., 2022) aiming at exposing the causal factors that explain item parameters are a promising avenue to a more tangible measurement unit. What is more, these approaches link a substantive theory of the construct to the formal measurement theory and thereby provide much stronger support for successful measurement.

Considerations of the measurement unit lead us to the pillar of **regression**. Engelhard (2022) explains how the Rasch model can convert non-linear phenomena, such as item responses, to a linear form. The Rasch model as a general linear model or generalized linear mixed model can be the basis for further extensions of the basic Rasch model. **Design**

is certainly an important pillar in statistics as much as it is in measurement. Data are never entirely objective independent of any context factor. Rather data are always collected in a particular manner and with a particular purpose in mind. Engelhard discusses design issues in a very general way starting with the definition of the latent variable, then the observational design, the scoring rules, and the Rasch model. What these considerations highlight is the dependence of measurement and its meaning on design. Every measurement scale is developed for a particular data collection design and with a specific purpose in mind. It is of utmost importance to be cognizant of these basic conditions and apply the scale and interpret measurements accordingly. Of course, it is conceivable to expand the applicability of a scale, to implement a new data collection design, or to use the measures in a different context. In doing so, one would explore new territory and should therefore operate in a scale development mode (where a comprehensive psychometric analysis is required) rather than a scale application mode (where one essentially trusts previously established properties). There is a close connection to intercomparison, where the principle of specific objectivity allowing for intercomparison also refers to a well-designed frame of reference.

The concept of **residuals** concludes the application of the seven pillars of statistical wisdom to measurement. Engelhard (2022) emphasizes that residuals are an important source of information and help us determine whether all requirements of measurement are sufficiently met. They are the basis for fit statistics, assessments of unidimensionality, local independence, or differential item functioning. It is in the residuals that reveal the limitations and inadequacies of our measurements. Another aspect of residuals that could have been added is the intrinsic relevance of residuals for measurement. The Rasch model predicts the probability of a particular response to an item by a given subject. A manifest response necessarily represents an "extreme" outcome. In a dichotomous item, we can only observe a negative response (scored 0) or a positive response (scored 1). In contrast, the model predicts a probability, which is a score between 0 and 1. Therefore, there will always be a residual as the difference between observed and expected responses in the Rasch model. In the deterministic Guttman model, there would be no residuals (provided there are no Guttman errors). Then, however, it would not be possible to estimate quantitative differences between items. We could only order items and persons. Thus, the concept of residuals is crucial and indispensable for quantitative measurement. Using the Rasch model, perfect measurement would not imply the absence of residuals but the lack of any information in the residuals. Then we would reach the boundaries of measurement.

Finally, Engelhard (2022) expands the pillars of measurement with the two additional concepts of power and consequences, which constitute distinctive pillars of educational measurement. **Power** is related to the function of measurement, that is, its purpose and its role in a broader educational policy. For example, educational measurements can be carried out to identify students who require further support. Identifying deficits, therefore, has positive consequences for those affected. On the other hand, measurements can help ensure that only the most suitable candidates are selected for admission to higher education. This could indeed be beneficial for the society as a whole. For individuals failing to reach the required threshold, it will have adverse consequences, though. Such conflicting interests of individuals and/or the society cannot be resolved objectively. Clarifying to what extent educational measurements are socially desirable does not lie at the heart of psychometrics. After all, the scope and purpose of educational measurements are political decisions. But ensuring the quality of measurements that are supposed to inform decision-making is a core task of psychometrics. Engelhard discusses power in the context of validity, and it is indeed validity, which lends measurement legitimacy. In principle, all pillars of measurement wisdom are relevant for validity. Intercomparison is probably particularly relevant as fairness and comparability of measurements (across genders, age groups, and ethnic groups to name a few) is key when decisions are based on educational measurement. Thus, psychometrics has to be aware of the power of measurement and what is at stake for individuals, and consider the context of use in the scale development and validation process. While Engelhard discusses power in the context of implications of measurement to the individual, **consequences** deal with the effects of testing regimes on educational practice. Engelhard mentions the narrowing of the curriculum as an unintended consequence of educational assessment. Again, the purpose of measurement must be kept in mind when defining constructs, designing scales, and carrying out measurements. The consequences may not be a core element of the measurement per se, but they must be considered to avoid unintended harmful effects. However, what counts as harmful largely remains a political choice.

Engelhard's (2022) attempt to transfer the pillars of statistical wisdom to the realm of measurement provides a frame of reference for a better understanding of different measurement models, their characteristics, their virtues, and their limitations. While some pillars represent descriptive criteria, intercomparison stands out as a prescriptive concept highlighting the importance of invariance as a key property of measurement. In any case, the pillars are closely related to each other and must be considered in their entirety.

The two additional pillars of power and consequences relate measurement to the societal context, in which measurement serves a particular function. They certainly play an important role in educational assessment. However, their relevance certainly extends to many, if not all, branches of the social sciences. In health, measurements can be used, for example, to identify patients who need treatment. On the other hand, the same measurements could be used to single out those likely to benefit most from treatment while excluding others. Even though such difficult decisions must be made in health economics, responsible psychometrics must not be blind to the consequences of measurement. In market research, measurements among consumers can be used to identify customer needs in order to best meet those needs in product design. But measurements may also be used to solely inform strategies to maximize company profits.

The pillars of measurement wisdom are like a flashlight highlighting the properties of the measurement models. It remains to be seen whether the proposed framework will enlighten researchers who are still championing measurement models that do not meet these requirements. After all, you have to be willing to pick up the flashlight and classify the things you see accordingly. The measurement requirements are not immediately apparent just from looking at the pillars. With these requirements in mind, looking at measurement models from the perspective of the pillars of measurement wisdom arguably does show that some models are wiser than others. In the next step, the scheme of pillars of measurement wisdom could benefit significantly from incorporating the requirements of measurement. In the end, the common goal of researchers in the social sciences must be to prevent social measurements from becoming a parody, which is a significant risk, whatever the field of research. Power and consequences as additional pillars show how much can be at stake for the individual and, ultimately, for society as a whole.

References

Adroher, N. D., & Tennant, A. (2019). Supporting construct validity of the evaluation of daily activity questionnaire using linear logistic test models. *Quality of Life Research*, *28*, 1627–1639.

Andrich, D. (2002). Understanding resistance to the data-model relationship in Rasch's paradigm: A reflection for the next generation. *Journal of Applied Measurement*, *3*(3), 325–359.

Andrich, D. (2018). Advances in social measurement: A Rasch measurement theory. In F. Guillemin, A. Leplège, S. Briançon, E. Spitz, & J. Coste (Eds.), *Perceived health and adaptation in chronic disease* (pp. 66–91). Routledge.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Briggs, D. C. (2019). Interpreting and visualizing the unit of measurement in the Rasch model. *Measurement*, *146*, 961–971.

Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research*, *61*(12), 1250–1262.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Engelhard, G., Jr. (2022). The pillars of measurement wisdom. *Journal of Applied Measurement*, *23*(3/4), 80–95.

Fisher, W. P., Jr. (2010). Statistics and measurement: Clarifying the differences. *Rasch Measurement Transactions*, *23*(4), 1229–1230.

Fisher, W. P., Jr., & Cano, S. J. (Eds.). (2023). *Person-Centered outcome metrology, principles and applications for high stakes decision making*. Springer Nature.

Humphry, S. M., & Andrich, D. (2008). Understanding the unit in the Rasch model. *Journal of Applied Measurement*, *9*(3), 249–264.

Ludlow, L. H., & Haley, S. M. (1995). Rasch model logits: Interpretation, use, and transformation. *Educational and Psychological Measurement*, *55*(6), 967–975.

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for metrologists. *Measurement*, *51*, 315–327.

McGrane, J. A. (2015). Stevens' forgotten crossroads: The divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Frontiers in Psychology*, *6*, 431.

Melin, J., Cano, S. J., Flöel, A., Göschel, L., & Pendrill, L. R. (2021). Construct specification equations: 'Recipes' for certified reference materials in cognitive measurement. *Measurement: Sensors*, *18*, Article 100290.

Melin, J., Cano, S., & Pendrill, L. (2021). The role of entropy in construct specification equations (CSE) to improve the validity of memory tests. *Entropy*, *23*(2), 212.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge University Press.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, *14*(1), 58–94.

Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement*, *46*(2), 359–372.

Stenner, A. J., Fisher, W. P., Jr., Stone, M. H., & Burdick, D. (2023). Causal Rasch models. In W. P. Fisher, Jr. & P. J. Massengill (Eds.), *Explanatory models, unit standards, and personalized learning in educational measurement: Selected papers by A. Jackson Stenner* (pp. 223–250). Springer Nature Singapore.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Stigler, S. M. (2016). *The seven pillars of statistical wisdom*. Harvard University Press.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement*, *16*(4), 8–14.

Wright, B. (1990). What is Information? *Rasch Measurement Transactions*, *4*(2), 109.