JOURNAL OF APPLIED MEASUREMENT, *23*(3/4), 117-120
Copyright© 2023

REJOINDER

Response to Commentaries on Pillars of Measurement Wisdom

George Engelhard, Jr.
The University of Georgia

A major purpose of my focus article was to foster discussion regarding the foundations of measurement using the pillars of statistical wisdom as a starting point. It was exciting to read the responses of the commentators to my invitation to consider pillars of measurement of wisdom. Their reactions and cogitations are exactly what I had hoped for in writing this piece.

Why is a discussion of the foundations of our field important? The foundations of our field provide a useful starting point for students and scholars entering our measurement community. Identification of the foundations of our field encourages all of us to view our work with a different lens. The goal is not necessarily to set out a definitive set of pillars of measurement wisdom but to foster discussion and reflection. As pointed out by Stigler (2016), "the seven pillars are the principle support for statistical wisdom; they do not by themselves constitute wisdom" (p. 195). With these points in mind, I offer brief comments on each commentary.

De Ayala (2022)

De Ayala raised a key question in his commentary: Do the pillars of educational measurement apply to non-educational measurement as well? His short answer is 'yes', and I agree with this response! He suggested adding two pillars of measurement wisdom: invariance and psychometric validity.

I placed invariance within the intercomparisons pillar in order to maintain a link to the pillars of statistical wisdom. However, I concur with De Ayala's suggestion that invariance warrants being considered a separate pillar of measurement wisdom. My interpretation of the intercomparisons pillar for measurement focused on the use of internal variation to specify a probability scale. I also related intercomparisons to invariance. Table 3 in my article included differential item and person functioning under the pillar of comparisons, but I inadvertently left them out of the text. In retrospect, I strongly agree with several of the commentators who identified invariance as worthy of a separate pillar.

De Ayala notes that "invariance allows

Requests for reprints should be sent to George Engelhard, Jr., The University of Georgia, 311 Connecticut Ave NE, Atlanta, GA 30307, USA; email: gengelh@uga.edu

for intercomparisons. However, these intercomparisons can occur when one compares individuals and/or items within a given data set without reference to exterior criteria" (2022). He goes further in saying "invariance also allows these comparisons to transcend a given data set from an external frame of reference" (2022). This second point provides the basis for considering structural models that seek invariant relationships between latent variables. Invariance can be evaluated based on internal comparisons within a data set, including differential item and person functioning, but it is also important to examine external comparisons, including a quest for invariant relationships between variables.

The distinction between internal and external invariance is clearly a topic for further discussion in our field (Asún et al., 2017).

De Ayala suggested psychometric validity as a second pillar of measurement with power and consequences subsumed under this pillar. He points out that a unique feature of psychometric validity is its focus on the consequences of measurements. In addition to considering psychometric validity as a pillar, the current *Test Standards* identify reliability and fairness as foundations of assessment (American Educational Research Association et al., 1999). Myford (2022) also suggested considering validity, along with reliability and fairness as foundational pillars.

Myford (2022)

Myford suggests that it is always important to be able to offer explanations of our work that our moms would understand—at this stage in my life, I also strive for explanations that my young grandsons will understand!

She believes that the house of modern measurement wisdom is supported by more than seven pillars. For example, she draws attention to the last chapter in Stigler's book where he muses about whether seven pillars are sufficient for communicating "the central intellectual core of statistical reasoning" (Stigler, 2016, p. 3). Myford questions if the pillars that I identified provide a sufficient foundation for supporting the house of modern measurement wisdom. Her answer is that additional pillars are needed. In particular, she suggests that validity, reliability, and fairness all deserve to stand on their own as separate pillars. This would follow closely the *Test Standards* that refer to validity, reliability, and fairness as the three foundations of measurement. The *Test Standards* also include separate sections on testing applications, such as psychological testing and assessment, workplace testing and credentialing, and educational testing and assessment. These sections focus on salient and perhaps distinctive pillars that may arise within different application areas.

The *Test Standards* are highly influential in our field, it seems quite reasonable to consider validity, reliability, and fairness separately as additional pillars supporting the house of modern measurement wisdom. De Ayala also made the case for including psychometric validity as a measurement pillar.

Salzberger (2022)

Salzberger points out the importance of having a solid conceptualization of the items used to define the latent variable. He argues that social measurement should be accompanied by a substantive theory of the construct to be assessed. In fact, this connection between measurement theory and the substantive theory is not addressed directly in any of the pillars of wisdom. Without both theories, Salzberger worries that this may lead to a parody of measurement that can be created where any number-generating procedure would be considered measurement.

Salzberger also indicated that the pillar of intercomparisons needs modification. Specifically, he mentions the importance of the issue of having a unit of measurement. He also points out that social measurement persistently struggles with defining meaningful units of measurement.

Power is related to the intended purposes of measurement within a broader policy environment, and Salzberger argues that conflicting interests of individuals and society cannot be resolved objectively. He points out that the scope and purpose of educational measurements are political decisions while ensuring the quality of measurements that are supposed to inform decision-making is a core task of psychometrics. The consequences may not be a core element of the measurement per se, but they must be considered to avoid unintended harmful effects.

Salzberger uses the image of a flashlight to highlight the properties of the measurement models. He notes that some measurement models may be wiser than others. I agree that the common goal of researchers in the social sciences must be to prevent social measurements from becoming a parody. The issues raised by Salzberger warrant careful attention in our field.

Liou (2022)

Liou suggests that data information (mutual information or relative entropy) can serve as a bridge to integrate the pillars of statistics with those of measurement. She uses differential item functioning as an example of this idea. In her commentary, she introduces the use of mutual information as an approach for the orthogonal decomposition of information related to cross-classified categorical data. I was not familiar with this approach by Liou and her colleagues (Liou et al., 2023). It appears to be a very promising way to examine not only differential item functioning but also other issues in measurement.

She also reminds us that educational measurement involves at least two dimensions. The first dimension is related to the widely recognized psychometric properties that focus on variation between persons. The statistical pillars clearly reflect the value and applicability of this first dimension to measurement-related issues. The second dimension focuses on changes within each person due to an educational intervention. Liou challenges us to consider maximization of gains achieved between pre- and post-assessments rather than simply maximizing between-person variation. This edumetric perspective merits more attention (Carver, 1974). I heartily agree with her point that researchers should develop measurements that focus on supporting gains in student achievement.

Summary

The purpose of my presidential address to the Pacific Rim Objective Measurement Society was to consider the basic pillars that support modern measurement. My goal was to discuss the pillars of statistical wisdom suggested by Stigler and then to consider how these pillars may connect to foundational issues in measurement with particular attention to Rasch measurement theory. Finally, I chose educational measurement as a place to start in considering additional pillars because I have worked extensively on measurement problems in educational contexts.

I centered my discussion of measurement pillars around Rasch measurement theory with a focus on educational measurement. Rasch measurement theory is a particular "flashlight," and it would be informative to consider how other measurement traditions relate to the foundations of our fields. What are the pillars of measurement wisdom from the perspective of classical test theory, other item response models (e.g., 2PL and 3PL), and measurement models related to the factor analysis and structural equation modeling?

All of the commentators have given us much to consider about the foundations of our field of measurement. As expected, there are some disagreements about the specific pillars that are relevant for measurement in the human sciences. I am excited that an important conversation has been started about pillars of measurement wisdom. I look forward to continuing to engage in further discussions regarding the foundations and additional pillars of measurement.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. AERA.

Asún, R. A., Rdz-Navarro, K., & Alvarado, J. M. (2017). The sirens' call in psychometrics: The invariance of IRT models. *Theory & Psychology*, *27*(3), 389–406.

Carver, R. P. (1974). Two dimensions of tests: Psychometric and edumetric. *American Psychologist*, *29*(7), 512–518.

De Ayala, R. J. (2022). Are there pillars of measurement? *Journal of Applied Measurement*, *23*(3/4), 96–102.

Liou, M. (2022). Commentary: Measurement and data information. *Journal of Applied Measurement*, *23*(3/4), 113–116.

Liou, J.-W., Liou, M., & Cheng, P. E. (2023). Modeling categorical variables by mutual information decomposition. *Entropy*, *25*(5), 750.

Myford, C. M. (2022). Discussant remarks on the pillars of measurement wisdom. *Journal of Applied Measurement*, *23*(3/4), 103–104.

Salzberger, T. (2022). 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. *Journal of Applied Measurement*, *23*(3/4), 105–112.

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