JOURNAL OF APPLIED MEASUREMENT, *23*(3/4), 96-102
Copyright© 2023

## COMMENTARIES

# Are There Pillars of Measurement?

R. J. De Ayala
*University of Nebraska – Lincoln*

Professor Engelhard's *Pillars of Measurement* (2022) demonstrates how Stigler's (2016) *The Seven Pillars of Statistical Wisdom* relates to measurement and introduces two pillars specific to measurement. We adopt *The Seven Pillars of Statistical Wisdom* framework by using a historical perspective to understand and comment on the pillars of measurement. (In the following we bold and italicize Stigler's pillars and italicize Engelhard's.)

Psychological measurement has long been influenced by the work of statisticians (e.g., Allan Birnbaum, Francis Edgeworth, and Georg Rasch). Therefore, it is not surprising that Engelhard invokes the statistical pillars in his quest to establish measurement pillars. The use of statistical techniques in measurement has a long history. For instance, in 1888, Edgeworth applied the theory of errors in his work "The Statistics of Examinations." In his "letter," he reiterated that observations of measurable quantities (e.g., times, distances) were "blurred by a fringe of error and margin of uncertainty" (p. 600). We say "reiterated" because the use of the mean to address error-laden observations had previously been done in other fields. For instance, although it may have certainly occurred earlier and in different fields, the first recorded instance of using the average of a set of measurements to address the errors in measurement appears to be by astronomers in the latter 1500s. Galileo noted that the errors in measurement were symmetrically distributed about their "true" value (Read, 1985, p 348). Thus, it was reasonable for astronomers interested in determining a characteristic of a celestial body (e.g., its location) to use the mean of their observations as the object's distance. By accumulating (***aggregation***) these erroneous (independent) observations ($X$s) and calculating the mean of the $X$s astronomers were able to approximate the "true" celestial body's characteristic (e.g., its location).

In the mid-1700s Simpson using a probabilistic perspective and assuming independence of $X$s (Simpson's Supposition 1) showed that using the mean led to an overall reduction in error and thereby provided a justification for the astronomers' approach. As such, the use of the mean was to be preferred to other approaches such as a carefully selected value. Therefore, despite errors in observed measurements one could use the mean to reduce the impact of these errors to obtain a better estimate of, for instance, a planet's location.

As an example, let us suppose we have $n$ independent distance observations of a planet's location ($X_1$, $X_2$, …, $X_n$). The mean of these values ($\bar{X}$) is our estimate of the planet's location. Stated another way, we can fit a null (means) model to these data so our "best" prediction of the value for an observation ($X'$) is

$$X' = \beta_0, \tag{1}$$

where $\beta_0$ is a constant equal to $\mu_X$ and estimated by $\bar{X}$. The parameter $\mu_X$ is the "true" location based on the fallible $X$s over an infinite number of observations, that is, $\mu_X = E(X)$.

Following Galton's work on ***regression***, we can separate "permanent and transient effects" (Stigler, 2016, p. 131). Thus, we can separate the errors in the $X$s from the "true" value by representing $Xs'$ fallibility as the deviation of an observed location ($X$) from its mean: $\epsilon = X - \mu_X$ (i.e., the distance the observation is from the mean). The simplest statistical (linear) model that relates $\epsilon$, $X$, and $\mu_X$ is

$$X = \beta_0 + \epsilon = \mu_X + \epsilon = E(X) + \epsilon. \tag{2}$$

In Equation 2, we have a sole parameter, $\beta_0\left(\mu_X\right)$, the mean of an infinite number of error-laden independent observations. Not only does $\epsilon$ reflect our uncertainty about the "true" value, but it also inversely reflects the lack of information we have about the characteristic of interest. Additionally, we can summarize the total variability in our $n$ $X$s and summarize the errors to reflect the total variability not accounted for by $\mu_X$.

Returning to our planetary location example, it is important to note we can never know the planet's true location relative to Earth at any given point in time because planets are not smooth, perfectly spherical, have elliptical orbits, parallax, a finite number of measurements, and so forth. Rather, conventions/standards are designed that define a framework within which the mean location is taken as a reasonable approximation of truth. It is the utility of the measured average location that is paramount. In short, the planet's exact location is unobservable, but one may consider the average to be the planet's "true" location within some small margin of error. Stated another way, we conceive of a latent location variable that has no error while recognizing that our observations contain error and thereby so does our estimate. Additionally, we are moving beyond the data at hand to infer the parameter's location from an estimator $\bar{X}$.

Simpson's work on the mean focused on errors. Similarly, Edgeworth's work in testing focused on (random) errors. As part of his treatise, he argued that the distribution of these errors in testing could be expected to form a unimodal symmetric distribution shaped like a *gensd'armes' hat*. That is, some deviations from the "true" value would be smaller and far more common than others, whereas others would be larger and rarer. Unlike our planet location example where $X$ is not a composite, in Edgeworth's presentation the observations were the examinee's average summed scores across multiple examiner ratings across multiple assessment tasks. Nevertheless, one may utilize the model above (Equation 2) for examination data. Therefore, the model may be rewritten as the true score model

$$X \equiv E(X) + \epsilon = \mathrm{T} + \epsilon, \tag{3}$$

where $X$ is a respondent's observed score across assessment tasks (typically a sum score across item responses; an ***aggregation***), $\epsilon$ is a random error defined as $\epsilon \equiv X - \mathrm{T}$(***residual***), and T is the respondent's expected observed score across an infinite number of independent assessments, that is, $\mathrm{T} \equiv E(X)$, the respondent's true score. Thus, the $X$ we observe is conceptualized to be one of an infinite number of possible $X$s that could have been observed for a respondent on a particular assessment with T as the mean of this theoretical distribution of $X$s. Analogous to not being able to determine a planet's true location, we can never know the true value of T. Because T is unobservable the error of measurement $\epsilon$ is also unobservable. However, it is important to

Requests for reprints should be sent to R. J. De Ayala, Dept of Educational Psychology, TEAC 114, University of Nebraska, Lincoln, NE 68588, USA; email: rdeayala2@unl.edu

note that the parameters in Equation 3 are *not* defined in terms of the real world (Lord, 1980) nor is the model falsifiable. Given $\mathrm{T} \equiv E(X)$ the mean error for this theoretical distribution of *X*s must be 0 and therefore across a group of examinees administered the same assessment the ***regression*** of $\epsilon$ on T yields a coefficient of 0 (Lord, 1980). (The ***regression*** pillar subsumes the correlational statistic that is at the heart of many psychometric classical concepts and techniques [e.g., reliability, validity coefficients, factor analysis].)

As easily demonstrated (see Wright, 1968), the *X*s obtained on an assessment can be raised or lowered by manipulating the assessment. For example, in a proficiency context if we make the assessment comparatively easy relative to our examinees' proficiencies the *X*s will increase. As a result, the corresponding Ts will also increase. Therefore, *X*s and Ts are not independent of the instrument used in their determination.

The implicit assumption in Equation 3 is that the *J* components that comprise the *X*s measure a common characteristic and the $\epsilon$s are unique. These components may differ from one another with respect to their means as well as how well they relate the *X*s to the common (single) characteristic—the factor $\theta$. Standardizing the *X*s we can rewrite the model as

$$z_j = \beta_{0j} + \beta_1\theta + \epsilon_j, \tag{4}$$

where $\beta_{oj}$ is the mean of component $j$, $\beta_1$ indicates the linear relationship between $z$ and $\theta$, $\theta$ is the respondent's measure on the factor (i.e., a factor score), $\epsilon_j$ is error (i.e., the respondent's measure of the unique property of component *j*), and $\mathrm{T} \equiv \beta_1\theta$. In this form, each of the components may vary in terms of their means. However, the use of $\beta_1$ rather than $\beta_{1j}$ indicates that all components have the same relationship with $\theta$ (essentially tau-equivalent). Equation 4 is a simple linear ***regression*** model. As such, it is an example of the generalized linear model using the identity link function. In a factor analytic context, Equation 4 is referred to as the common factor model. In this case, the regression coefficient(s) would be called factor loading(s) and would capture how well a component differentiates among respondents with respect to the common factor $\theta$. If the components represent items, then a loading represents a component's discrimination capability. (Recall we have assumed a constant loading in Equation 4; i.e., constant discrimination.) Moreover, because a factor score is a weighted linear composite then $\theta_i$ represents an ***aggregate*** and is, as Engelhard states, a form of data reduction.

Looking at the systematic portion of Equation 4 (i.e., $\beta_0 + \beta_1\theta$), considering the components to be items, dichotomous responses, setting $\beta_1$ to the constant 1 (for convenience), then the mean response across respondents for item *j* is its difficulty ($\beta_0 = -\delta_j$) and we have

$$z_j = \beta_0 + \beta_1\theta = -\delta_j + \theta. \tag{5}$$

Equation 5 has a parameter reflecting an individual's location on the common factor (e.g., mathematics ability) and another reflecting item *j*'s location (e.g., difficulty, $\delta_j$). To establish a metric that would allow one to make ***intercomparisons*** assume, as Rasch (1960/1980, pp. 74–75) did, that there are two individuals one of whom has twice the ability of the other, $\theta_2 = 2\theta_1$, and there are two items such that one item is twice as difficult as the other, $\delta_2 = 2\delta_1$. If the second person correctly answers the second item and the first person correctly answers the first item, then the probabilities of a correct response should be the same. Moreover, this means that the probability of the response is a function of the ratio $\theta/\delta$ and not on the values of $\theta$ and $\delta$ separately. It is this characteristic ratio that allows for item-free measurement of $\theta$. This is the essence of the invariance concept (Thorndike, 1904, 1912; Thurstone, 1926; also see Thurstone, 1925) that underlies IRT.

Engelhard appears to consider invariance to fall within Stigler's ***intercomparisons*** pillar. Stigler defines this pillar as "the idea that statistical comparisons may be made strictly in terms of the interior variation in the data, without reference to or reliance upon exterior criteria" (Stigler, 2016, p. 87). As an example, he refers to the use of standard errors. It is apparent that invariance allows for intercomparisons. However, ***intercomparisons*** occur when one compares individuals and/or items within a given data set without reference to exterior criteria. Invariance allows these comparisons to transcend a given data set without an external frame of reference. Although invariance may be considered to be a form of ***intercomparisons***, invariance is a psychometric fundamental principle that has "... supported our field in different ways in the past and promise[s] to do so in to the indefinite future" (Stigler, 2016, p. 2). In this regard, we consider invariance to be a measurement pillar in its own right.

According to Rasch (1960/1980, pp. 74–75) the simplest function that he was aware that increases from 0 to 1 as $\theta/\delta$ increases is the logistic function. Thus, by substitution of Equation 5 into the logistic function we have the Rasch model for dichotomous responses:

$$p\left(x_j = 1|\theta,\, \delta_j\right) = \frac{e^{Zj}}{1+e^{Zj}} = \frac{e^{\left(\theta-\delta_j\right)}}{1+e^{\left(\theta-\delta_j\right)}}, \tag{6}$$

where $x_j$ is the response by a person to item *j*. Because the Rasch model is a logistic regression model it is a generalized linear model with the logit link. Additional variables may be added to the model that, under the right circumstances, could be explanatory variables. Equation 6 is an example of one member of the Rasch family of models. One may consider the Rasch family of models to be an attempt to establish a standard by which all measurements are obtained not unlike an astronomer's creation of the astronomical unit, light year, or parsec for measuring distances. This idea of creating a measurement standard underlies other measurement approaches as well (e.g., Guttman Scalogram, Coombs Unfolding) and appears to be in Engelhard's *power* pillar.

Because Equation 6 is a ***regression*** model it can be used to make model-based predictions that are compared to the observed data through ***residual***-based indices. This residual analysis is used to diagnose the degree of correspondence (or lack thereof) between the model and the observed data. This diagnostic analysis might lead one to the realization that one does not have sufficient model-data correspondence (i.e., fit) to proceed. At first glance, this may appear to be a disadvantage of Equation 6. However, this is not the case. In fact, it is better to be able to identify a lack of congruence between the model and the data, than it is to continue with a model that is not falsifiable. Furthermore, residual analysis allows for model comparisons, such as a comparison of Equation 6 with a variant, such as the linear logistic test model.

The estimates of the parameters $\theta$ and $\delta$ are those values that are most likely to yield the observed responses. Thus, the ***likelihood*** pillar underlies almost all estimation procedures for both $\theta$ and $\delta$ as well as their corresponding estimation errors. This pillar is also reflected in the confidence intervals corresponding to the point estimates of $\theta$ and $\delta$.

Whether or not the $\theta$ estimate is a measure of the construct of interest (validity) is affected by Stigler's ***design*** pillar. This pillar includes "... the planning of observation generally, and the implications for analysis of decisions and actions taken in planning" (Stigler, 2016, p. 149). In the context of psychometrics, this pillar is represented in the process by which instruments are created and refined to sample behaviors. This process is represented in Engelhard's design blocks and reflects part of the validation process.

Although Engelhard uses educational measurement "... as an illustrative area of measurement practice" (2022, p. 90) and presents his pillars in a section entitled 'Distinctive Pillars of Educational Measurement' the focus is on, as the manuscript's title states ("The Pillars of Measurement Wisdom") distinctive measurement pillars. Stated another way and given our focus, are Engelhard's pillars of *power* and *consequences of measurement* found in non-educational measurement as well? The short answer is 'yes.'

Engelhard, quoting Porter, states "The Latin root of validity means 'power' … it is important to consider the notion of validity as a type of power" and "tests become a key tool for guiding and enforcing issues related to power over individual and societal decisions" (2022, p. 91). There is no denying the veracity of the latter quote in educational measurement. The current conception of validity recognizes this.

Broadly stated, one may consider validity to be how well measurement-based inferences are supported by the measurements. The "how well" may be captured by the correlation between the measurements with some criterion/criteria, what "may be properly inferred from a test score" (Brennan, 2006, p. 2), or by "… by the appropriateness, meaningfulness, and usefulness of the specific inferences made from the test scores" (American Educational Research Association [AERA] et al., 1974; cited in Brennan, 2006, p. 2). These definitions of "how well" reflect the evolution of validity over the past 100 years. Consequently, they subsume the predictive, content, and construct categories of validity (see Kane, 2006); the first two categories involve the ***regression*** pillar. We can see the concept has evolved from being somewhat solely statistically-oriented to a somewhat argument-based approach. That is, given an individual's responses on an instrument we seek to claim that the measure (e.g., θ) reflects the individual's position on the construct of interest (θ) regardless of whether this construct is, for example, their level of depression, generalized anxiety, or mathematics ability. In support of this claim our argument rests on obtaining evidence from test content, response processes, the instrument's internal structure, and so on (AERA et al., 1999). Unlike earlier conceptions of validity, more recent perspectives emphasize that validity is a property of the interpretations and uses of the measurements and not of the instrument itself. One might argue that the current perspective on validity is a response to public pressure for transparency and accountability given the power that measurements have to influence the decision-making process. Thus, validity is, in part, focused on the credibility of the measurement process.

The *power* pillar (i.e., "Power stresses the use of measures to define and construct the key constructs that are used to structure the world around us," 2022, p. 93) appears to involve the philosophical argument concerning objective truth versus subjective truth. That is, this pillar appears to invoke the idea that our measures are used to create and support the existence of behavioral constructs. Historically, there have been measurement approaches that adopted a perspective that if the approach worked (within some degree of tolerance) then it was possible to measure the intended construct, otherwise not. In other words, simply because we conceive of a construct does not mean we are able to measure it. In contrast, an alternative perspective is we conceive of a construct and it is a given that an instrument can be designed to measure it. In either case, the validation process for our measures informs us whether we can legitimately argue that our measurements have utility (e.g., providing "structure [to] the world around us"). In short, it appears that the *power* pillar falls within the current conceptualization of validity. This is not meant to discount the importance of the *power* pillar conception but simply to place it within a larger context.

Engelhard defines the *consequences* pillar as "… the idea that measures are created to serve specific purposes, and that the consequences may be both positive and negative." The issue of measurements' social consequences can be found in Messick (1989) and Kane (2006). In psychological measurement, the measurements reflect both the individuals as well as the context within which the individuals provide their responses. Recognition of this led Messick (1989) to advocate considering the consequences of the measurements as part of the conceptualization of validity and the validation process. Messick's (1995) consequential validity (an aspect of construct validity) reflects "value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice" (p. 745). In this respect, the *consequences* pillar appears to fall within consequential validity aspect of construct validity. What is unique to (psychometric) validity is its focus on the consequences of the measurements. As mentioned above, Stigler (2016) considers a pillar to have supported the field in the past and to have the promise to do so into the indefinite future. Psychometric validity meets these criteria. Thus, one of the pillars of measurement is psychometric validity.

We endorse Professor Engelhard's objective of identifying pillars of measurement. As such, it is important to ensure these pillars are not tied solely to educational testing, but that they underlie all psychological measurement regardless of whether the measurement is educationally focused, used for health-related purposes in the health sciences, in an industrial-organizational setting, and so on. Clearly, some of the pillars of measurement are the same as the pillars in statistics. As mentioned above, we consider invariance to be a pillar of measurement. Although invariance could be seen to be nothing more than ***intercomparisons***, the idea of freeing the measurements from the instrument used makes invariance uniquely different than ***intercomparisons***. Invariance may also be considered as evidence in the validation process. For example, in differential item functioning it is the absence of invariance that indicates that members of one group (e.g., females) are being unfairly disadvantaged compare to another group (e.g., males). Thus, the instrument's measurement may be determined to be biased and thereby adversely affect the validation process. Engelhard presents two pillars, *power* and the *consequences* of measurement, he considers to be distinct from the statistical pillars. We agree that *power* and the *consequences of measurement* are distinct from the statistical pillars. What is less clear to us is why these should be not considered to fall within the current conceptualization of validity. We believe the second pillar suggested above (i.e., validity) encompasses the *power* and *consequences* pillars.

## References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp 1–16). American Council on Education/Praeger Publishers.

Edgeworth, F. Y. (1888, September). The statistics of examinations. *Journal of the Royal Statistical Society*, *51*(3), 599–635.

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

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp 17–64). American Council on Education/Praeger Publisher.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education; Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press. (Original work published 1960)

Read, C. B. (1985). Normal distribution. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 6, pp. 347–359). Wiley.

Simpson, T. H. (1755). On the advantage of taking the mean of a number of observations in practical astronomy. *Philosophical Transactions*, *49*, 82–93.

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

Thorndike, E. L. (1904, 1912). *An introduction to the theory of mental and social measurements*. Teachers College, Columbia University.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*, 433–451.

Thurstone, L. L. (1926). The scoring of individual performance. *Journal of Educational Psychology*, *17*, 446–457.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Educational Testing Service.