JOURNAL OF APPLIED MEASUREMENT, *23*(3/4), 80-95
Copyright© 2023

FOCUS ARTICLE

The Pillars of Measurement Wisdom

George Engelhard, Jr.
The University of Georgia

The purpose of this study is to discuss the foundations of measurement in the human sciences. This discussion is framed by a consideration of the pillars of statistical wisdom proposed by Stigler (2016), and their relationships to key concepts in measurement theory. Stigler (2016) identified seven pillars of statistics: aggregation, likelihood, information, intercomparison, regression, design, and residuals. Each of these pillars has an interesting counterpart for measurement. There are several unique pillars for educational measurement that include a consideration of the power and consequences of using measures. Rasch measurement theory provides the guiding framework for considering the pillars of measurement.

Keywords: Foundations of statistics, foundations of measurement, history and philosophy of measurement, Rasch measurement theory

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

Acknowledgments: This manuscript is based on a keynote lecture given at the Pacific Rim Objective Measurement Society meeting in Hanoi, Vietnam (December 2022).

*When you cannot measure * your knowledge is * meager * and * unsatisfactory **

Lord Kelvin[1]

Each field of study is based on foundational pillars that guide research and practice. These pillars define the principles and aspects of the paradigms that undergird various programs of research. These foundational issues can be implicit or explicit, and it is important to periodically revisit and discuss these pillars.

Steven Stigler is a historian of statistics (1986, 1999, 2016). He identified seven pillars of statistical wisdom: aggregation, likelihood, information, intercomparison, regression, design, and residuals (Stigler, 2016). He based the notion of pillars of wisdom on a memoir by T. E. Lawrence who is also known as Lawrence of Arabia (Lawrence, 1926). Each of these statistical pillars can be used as a basis for reflecting on foundational issues in measurement. In particular, this study uses Rasch measurement theory as the basis for reflecting on these pillars of statistics, and their implications for measurement (Rasch 1960/1980).

The purpose of this study is to consider the basic pillars that support modern measurement. The following questions guide the structure of this study:

- What are the pillars of statistical wisdom?
- What are the connections between the pillars of statistical wisdom and the pillars of measurement wisdom?
- Are there additional pillars distinctive to educational measurement?

The major goal of this study is to reflect on the pillars of both statistics and measurement from a perspective informed by Rasch measurement theory. A secondary goal is to consider distinctive and useful pillars of measurement that were not identified by the original seven pillars of statistics.

Pillars of Wisdom

The history of science is the history of measurement. (Cattell, 1893, p. 316)

This study takes a broad view of trends and concepts that cut across statistics and measurement. It is inevitably historical and philosophical in its focus. The seven pillars are discussed separately in the following sections. Even though the pillars are discussed separately the pillars are strongly interconnected.

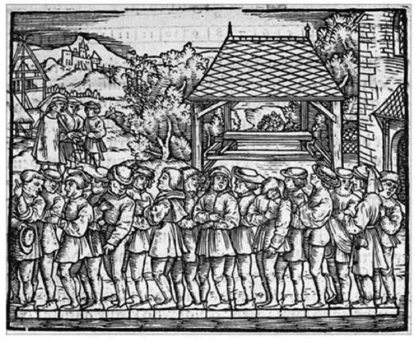
Aggregation

The object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data. (Fisher, 1922, p. 311)

Aggregation can be defined as an intentional and targeted summarization of data. The simple arithmetic mean is an example of data summarization. Stigler (2016) uses a classic example of data aggregation to empirically define the unit of a "foot" for measuring plots of land. Figure 1 shows an example of how aggregation might be used to define a standard

Figure 1

Definition of a Foot Based on Aggregation



Note. An illustration from the geometry book of Jacob Köbel (1460–1533), 1608 edition, Mathematical Association of America. (https://maa.org/press/periodicals/convergence/the-right-and-lawful-rood). In the public domain.

[1] Quote on the façade of Social Science Research Building at the University of Chicago (Merton et al., 1984).

unit for a “foot”. The standard unit can be based on an average of a foot size over the 16 persons in the example.

Many statistical methods embody the idea of aggregation and data summarization. For example, data can be summarized in frequency tables and contingency tables. Other statistical methods can be viewed as compact representations of a larger set of data. Correlation coefficients and least-squares regression are both examples of additional types of data aggregation. From a statistical perspective, the combination of observations to achieve data reduction and summary was viewed as problematic because some observations were being ignored.

From a measurement perspective, a major tool for data reduction is the concept of a latent variable. The idea of a latent variable has a long history in the human sciences. Classical test theory is a prime example with the conceptualization of a person’s test score as the sum of observed responses that can be decomposed into a true score and an error component:

Observed score = True score + Error score

The true score represents each person on a latent variable. Modern statistics and measurement theory are strongly grounded on the idea of a latent variable (Bollen, 2002).

Rasch measurement theory provides a framework for defining a line that represents the latent variable. Both persons and items can be located on this line based on the Rasch model. The generic view of a line used to represent a construct or latent variable is shown in Figure 2. A Wright map provides a key representation of the latent variable as a line with both person and item locations. Figure 3 shows an example of a Wright Map. The latent variable in Figure 3 is designed to represent learning stimulation in the home for children (Engelhard, 2013, p. 19).

In the next two sections, the application of the concepts of likelihood and information are used in conjunction with Rasch measurement theory to guide the creation of Wright Maps.

Figure 2

Generic View of a Line Used to Represent a Construct or Latent Variable

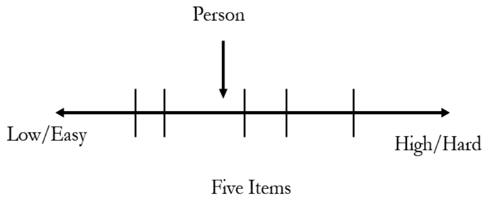
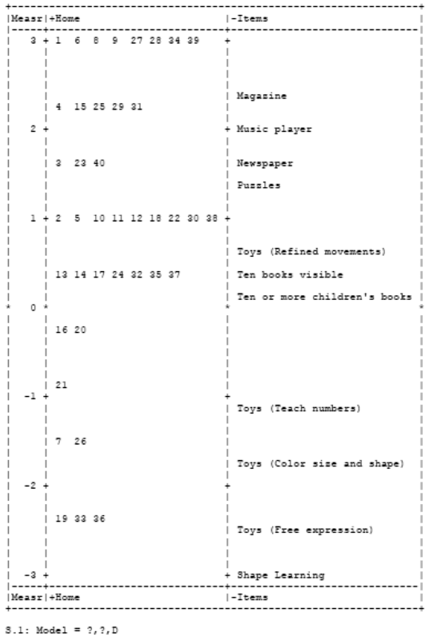


Figure 3

Wright Map for Learning Stimulation in the Home Environment



```
+-------------------------------------------------------------------+
|Measr|+Home                         |-Items                        |
|-----+------------------------------+------------------------------|
|   3 + 1  6  8  9  27 28 34 39      +                              |
|     |                              |                              |
|     |                              |                              |
|     |                              |                              |
|     |                              | Magazine                     |
|     | 4  15 25 29 31               |                              |
|     |                              |                              |
|   2 +                              + Music player                 |
|     |                              |                              |
|     |                              |                              |
|     | 3  23 40                     | Newspaper                    |
|     |                              |                              |
|     |                              | Puzzles                      |
|     |                              |                              |
|     |                              |                              |
|   1 + 2  5  10 11 12 18 22 30 38   +                              |
|     |                              |                              |
|     |                              |                              |
|     |                              | Toys (Refined movements)     |
|     |                              |                              |
|     | 13 14 17 24 32 35 37         | Ten books visible            |
|     |                              |                              |
|     |                              | Ten or more children's books |
*   0 *                              *                              *
|     |                              |                              |
|     | 16 20                        |                              |
|     |                              |                              |
|     |                              |                              |
|     |                              |                              |
|     | 21                           |                              |
|  -1 +                              +                              |
|     |                              | Toys (Teach numbers)         |
|     |                              |                              |
|     |                              |                              |
|     | 7  26                        |                              |
|     |                              |                              |
|     |                              | Toys (Color size and shape)  |
|     |                              |                              |
|  -2 +                              +                              |
|     |                              |                              |
|     |                              |                              |
|     | 19 33 36                     |                              |
|     |                              | Toys (Free expression)       |
|     |                              |                              |
|     |                              |                              |
|     |                              |                              |
|  -3 +                              + Shape Learning               |
|-----+------------------------------+------------------------------|
|Measr|+Home                         |-Items                        |
+-------------------------------------------------------------------+

S.1: Model = ?,?,D
```

Likelihood

An efficient statistic can in all cases be found by the Method of Maximum Likelihood; that is by choosing statistics so that the estimated population should be that for which the likelihood is greatest. (Fisher, 1970, p. 14)

Statistical methods utilize the concept of likelihood as a way of calibrating inferences with the use of probability. Likelihood methods provide approaches for developing a probability measuring stick for our inferences (Stigler, 2016). The calibration of this probability scale is an important contribution of statistics. In fact, some definitions of the field of statistics define statistical methodology as approaches for the quantification of uncertainty. Formal statistical inference is designed to answer specific research questions but also provides a measure of the reliability and uncertainty of the conclusions (Moore & McCabe, 1993).

A related view of likelihood is used in measurement. Logits are used to define the probability units in Rasch measurement theory. Logits, L_i, are log-odd units, and they can be defined for item i as follows

$$\mathrm{L}_i = \ln\left[\pi_i/\left(1-\pi_i\right)\right], \quad (1)$$

where π_i is the probability of correct responses to item i. Table 1 illustrates the connections between probability, odds, and the natural log of the odds (logits). For example, a probability of .50 corresponds to 1.00 and 0.00 for the odds and natural log odds respectively.

It is important at this point to formally introduce the dichotomous Rasch model. The Rasch model starts with the theory that a person’s response to a test item is a probabilistic function of the distance between the location of the person and item on the latent variable scale. The dichotomous Rasch model is an application of a simple logistic model to model this relationship. Rasch (1960/1980) developed a simple logistic model to represent the probability of a person responding correctly to a test item:

$$P\left(X_{ni}=1\right) = \pi_{ni} = \frac{\exp\left(\theta_n - \beta_i\right)}{1+\exp\left(\theta_n - \beta_i\right)}, (2)$$

where π_{ni} is the probability of person n (θ_n) responding correctly or positively to item i (β_i). It should also be noted that maximum likelihood estimation is commonly used to obtain estimates of the parameters in this model (Baker & Kim, 2004; Wright & Stone, 1979). This is discussed in more detail in the next section on information.

Table 1

Illustration of the Connections Between Probabilities, Odds, and Logits (log Odds)

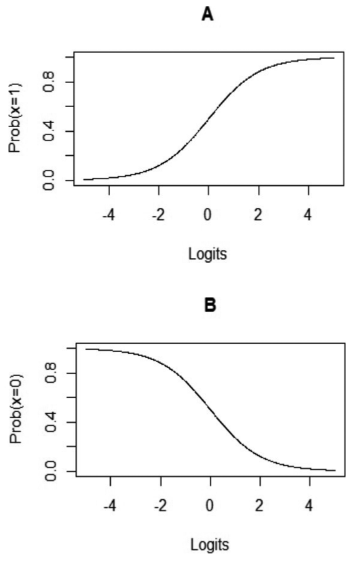
Prob	1–Prob	Odds	Logits
0.10	0.90	0.11	–2.20
0.20	0.80	0.25	–1.39
0.30	0.70	0.43	–0.85
0.40	0.60	0.67	–0.41
0.50	0.50	1.00	0.00
0.60	0.40	1.50	0.41
0.70	0.30	2.33	0.85
0.80	0.20	4.00	1.39
0.90	0.10	9.00	2.20

Figure 4 illustrates this equation in terms of item response functions. Figure 4 (Panel A) shows the relationship between the probability of getting a correct response on an item as a function of the latent variable on the x-axis. The probability of getting an incorrect answer is also shown in Figure 4 (Panel B).

The connection between statistics and measurement becomes evident in terms of the probability scale developed for a formal measurement model, such as the dichotomous Rasch model. Returning to the Wright Map in Figure 3, probabilistic inferences can be

Figure 4

Item Response Functions for Dichotomous Rasch Model (Item Location = 0)



drawn related to items and person locations. For example, the Home Environment Scale represents learning stimulation in the home. Magazines are relatively less likely to be observed, while toys tend to be easier to observe. The Wright Map also shows the locations of homes on the line. For example, Home 37 is located at .19 logits. This implies that it has approximately .50 probability of having "ten books visible." In terms of inferences about other objects in this home, the probability scale suggests that the likelihood of observing items above .19 logits is less than .50, while the likelihood of observing items below .19 logits is greater than .50. These values are based on the Rasch measurement model.

Information

The purpose of the statistical reduction of data is to obtain statistics which shall contain as much as possible, ideally the whole, of the relevant information contained in the sample. (Fisher, 1922, p. 366)

In general usage, information implies that something is known about a particular issue or topic. Statisticians are interested in developing methods for quantifying information, as well as the evaluating the rate of change in information based on different methods of estimation. Information provides an index of the degree of certainty in our inferences.

In statistics, Fisher (1922) proposed a specific and technical meaning for information. He defined information as the reciprocal of the precision with which a parameter can be estimated. Precision was defined by the variance of sampling distribution of the estimates.

As was the case for likelihood, information can be defined based on a formal measurement model. The Rasch model specifies information as follows

$$(I_{ni}) = \pi_{ni}\,(1 - \pi_{ni}), \qquad (3)$$

where π_{ni} is defined in Equation 2. The first two columns in Table 2 show person and item locations in logits on the latent variable. Column 3 (Table 2) is the difference in logits between these locations. Column 4 (Table 2) is the odds of succeeding on this item, while Column 5 (Table 2) is the probability of succeeding on this item based on the Rasch model. The final column is the information.

An information function can be used to provide a graphical representation of uncertainty. Figure 5 illustrates how information can be used to convey the precision of estimated person locations. Four people with a same sum score of 3 with different response patterns have different levels of precision as shown by the shape of the likelihood function. The item difficulties are −1.50, −.50, .50, and 1.50 logits. Figure 5 (Panel A) shows the most precision for defining person location, while Figure 5 (Panel D) has the most uncertainty. In measurement, information is also interpreted as an estimate of the uncertainty in a measure. The reciprocal of the square root of the information can be used to define standard errors of measurement in item response theory models (Baker & Kim, 2004).

Intercomparisons

Measurements are only useful for comparison. The context supplies a basis for the comparison—perhaps a baseline, a benchmark, or a set of measures for intercomparison. (Stigler, 2016, p. 64)

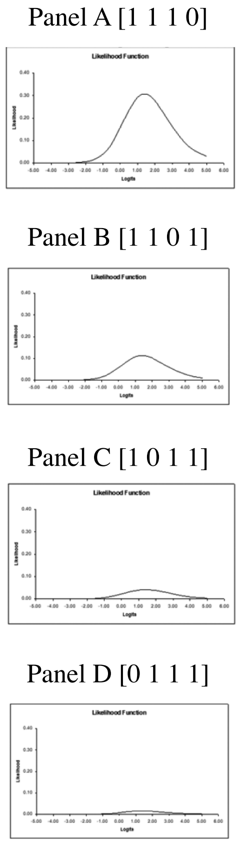
Table 2

Person and Item Locations in Logits With the Rasch Probabilities of a Correct Answer

Logits					
Person	**Item**	**Difference**	**Odds**	**Prob**	**Information**
5.00	0.00	5.00	148.41	0.99	0.01
4.00	0.00	4.00	54.60	0.98	0.02
3.00	0.00	3.00	20.09	0.95	0.05
2.00	0.00	2.00	7.39	0.88	0.10
1.00	0.00	1.00	2.72	0.73	0.20
0.00	0.00	0.00	1.00	0.50	0.25
0.00	1.00	−1.00	0.37	0.27	0.20
0.00	2.00	−2.00	0.14	0.12	0.10
0.00	3.00	−3.00	0.05	0.05	0.05
0.00	4.00	−4.00	0.02	0.02	0.02
0.00	5.00	−5.00	0.01	0.01	0.01

Figure 5

Likelihood Functions for Response Patterns With Sum Score of 3



Panel A [1 1 1 0]

Panel B [1 1 0 1]

Panel C [1 0 1 1]

Panel D [0 1 1 1]

Note. The item difficulties are −1.50, −.50, .50, and 1.50 logits.

The intercomparisons pillar is based on the idea that an index of uncertainty for our inferences can be constructed by using variation within an observed data set. One example of this pillar is the jackknife (Tukey, 1977). Tukey developed a method for estimating standard errors of estimates by successively omitting observations. Another example is bootstrapping developed by Efron (1982, 2003). Bootstrapping involves strategies for data resampling at random with replacement, and these bootstrap samples are then used to judge the variability of a statistic. As pointed out by Stigler (2016), "all of these methods involve intercomparison in estimating variability" (p. 101). This idea was controversial because an internal standard is being used based on the data set at hand without an external comparison group.

Intercomparisons play an important role in evaluating the quality of educational measurements. For the Rasch model, Rasch (1961) specified the importance of invariant comparisons in developing specifically objective measurements. His requirements are:

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared.
>
> Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion. (Rasch, 1961, pp. 331–332)

The first section of this quote deals with person-invariant calibration of items (stimuli). The goal of person-invariant item calibration is to locate items on a continuum, and to minimize the unwanted influences of person subgroups. The second section of Rasch's quote refers to item-invariant measurement of persons or individuals. The basic problem is to estimate a person's location on the same continuum or construct without undue dependence on the particular set of test items used or other persons.

Rasch measurement theory involves the application of these requirements. Rasch models reflect an ideal-type measurement model that meets these requirements. As shown in these requirements, comparison is a key idea for Rasch (Rasch, 1977). Andrich (2018) has argued that "Rasch's distinctive contribution to epistemology and social measurement [is] the centrality of invariant comparisons within a frame of reference" (p. 66).

Model-data fit is used to evaluate whether or not invariant comparisons have been achieved within a particular data set. Once a Rasch model is estimated, then item calibration and person measurement can be evaluated by comparing observed and expected data based on the models. The differences in these comparisons are the residuals, and they can be summarized across items and persons. The analyses of residuals is described later in the section on Residuals.

In the next section, linear models are used to define the expected values discussed in this section.

Regression

Fisher must have been the first to have that very broad vision of regression—or the linear model—which is one of the most fertile insights of modern statistics. (Savage, 1976, p. 451)

Regression is a method for studying the relationships between two or more variables. The response variable Y is the dependent variable, while X is used to designate the independent variables (covariates, predictor variables). Regression serves a similar purpose to aggregation because the data are represented by several summary statistics, such as regression coefficients.

The term regression is due to Galton (1886) who examined the relationship between mid- parent height (mean of heights of the parents) and their children's heights. Figure 6 shows his plot that supports the idea that the heights of children regress to the mean of the population. In other words, taller parents tend to have shorter children than expected, while shorter parents tend to have taller children than expected.

There is a long history of linear models in statistics that goes back to Gauss and Legendre who developed these models in astronomy (Stigler, 1981). For example, linear models were developed to represent the relationship between the positions of planets and stars based on the observations of astronomers. It was widely recognized that these observations varied based on potential measurement errors that varied randomly across astronomers. A detailed consideration of these error distributions led to the concept of a Gaussian or Normal distribution of errors, and the quest for methods to minimize the effects of these errors (McCullagh & Nelder, 1983).

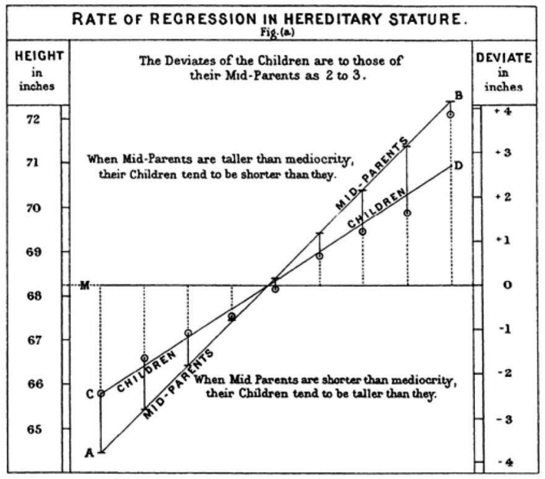
Reflecting this goal, the simple linear regression model is a linear model that predicts a dependent variable (Y) using values of an independent variable (X) with an error term (ε). The simple linear regression model can be written as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \qquad (4)$$

This model represents a straight-line plot with Y_i as a continuous dependent variable and X_i as the independent variable. The term $\beta_0 + \beta_1 X_i$ is defined as the mean response when $X = X_i$. The residuals ϵ_i are assumed to be independent and normally distributed with a mean of 0 and variance of σ^2. The parameters in the model are β_0 (intercept), β_1 (slope) and σ^2 (variance). Estimates of the regression coefficients, β_0 and β_1, are commonly obtained with either least squares or maximum likelihood estimators (Wasserman, 2010).

Figure 6

Regression Toward the Mean (Galton, 1886)



As mentioned in the opening quotation, regression or the linear model is one of the most fertile insights of modern statistics. Recent advances in statistics have shown how linear models can be generalized to analyze categorical variables (Tutz, 2012). This conceptual breakthrough led to the development of generalized linear models using the tools developed for linear models (Azen & Walker, 2010; McCullagh & Nelder, 1983). The key idea underlying generalized linear models is that a transformation of the categorical dependent variable (Y) is more likely to yield a linear relationship with an independent variable (X).

A generalized linear model for predicting a dichotomous outcome can be written as shown in Equation 5. The probability of a response in category 1 of the dichotomous outcome variable (Y = 1) using the exponential form can be expressed as

$$\Pr(Y=1|X) = \pi(X) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + (\exp(\beta_0 + \beta_1 X_i))}, \qquad (5)$$

where $\Pr(Y=1|X)$ is the conditional probability $\pi(X)$ of observing Y = 1 given X. The parameters in this model have a comparable meaning to those in the simple linear regression model presented earlier, with β_0 (intercept) and β_1 (slope).

This can be re-written as logits:

$$Logit = g(X) = \ln \frac{\pi(X)}{(1-\pi(X))}, \qquad (6)$$

and it can also be expressed as a linear model:

$$g(X) = \beta_0 + \beta_1 X_i. \qquad (7)$$

In measurement, it has been recognized that the Rasch model can be viewed as a generalized linear model. The Rasch model can be written as

$$P(X_{ni}=1) = \pi_{ni} = \frac{\exp(\theta_n - \beta_i)}{1 + \exp(\theta_n - \beta_i)}, (8)$$

where π_{ni} is the probability of person n (θ_n) responding correctly or positively to item i (β_i). Logits are defined as

$$Logit = \eta_{ni} = \ln\left(\frac{\pi_{ni}}{1-\pi_{ni}}\right) \qquad (9)$$

and

$$Logit = \eta_{ni} = \theta_n - \beta_i. \qquad (10)$$

The Rasch model can be written in the form of a GLMM as:

$$\eta_{ni} = \theta_n X_{i0} + \sum\nolimits_{k=1}^{K} \beta_{1i} X_{ik}, \qquad (11)$$

where η_{ni} is the logit for person n and item i. Equation 11 highlights that both persons and items have design matrices with X_{i0} as a constant vector equal to 1, and X_{ik} is a matrix with 1s on the diagonal for the dichotomous Rasch model. Based on Equation 11, the constant can be viewed as a person parameter (random effect) and β_{1i} as item parameters.

Once the Rasch model is viewed as a generalized linear model, then this opens up the door for numerous extensions to the Rasch model using generalized linear mixed models (GLMMs). GLLMs are very flexible, and include the random effects for various parameters in the model.

The logits (η_{ni}) can be modeled with person (Z) and items (X) covariates:

$$\eta_{ni} = \sum_{j=0}^{J} \theta_{nj} Z_{(n,i)j} + \sum_{k=0}^{K} \beta_k X_{(n,i)k}. \qquad (12)$$

De Boeck and Wilson (2004) described these extensions in terms of item and person predictors. They described Rasch models with no predictors as doubly descriptive (e.g., dichotomous Rasch model). Rasch models with item predictors (X) were referred to as item explanatory (e.g., linear logistic test model), and Rasch models with person predictors (Z) were referred to as person explanatory (linear logistic regression model). Finally, Rasch models with both item and person predictors were labeled as double explanatory.

In summary, linear models are a singular achievement of statistics. Generalized linear models extend this framework for analyzing categorical data. Rasch models can be estimated based on generalized linear mixed models. This perspective sets up an array of further extensions to Rasch measurement theory based on generalized linear mixed models.

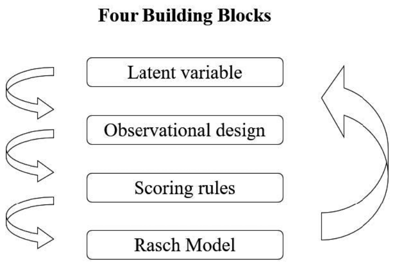
Design

One of the pillars of statistics is anchored on recognizing the importance of planning the collection of observations. The design of data collection plays an essential role in research. The importance of the design of experiments in particular is based in the ground breaking ideas of Fisher in his work on the design of experiments (Fisher, 1935). It is also important to recognize that the plan for data collection also plays a key role in observational studies. What to observe and how it is observed are crucial issues in all research.

In measurement, design plays an important role in the construction of scales that are used to define the latent variable. Figure 7 shows the main building blocks that can be used to guide the design of a scale (Engelhard & Wang, 2021; Wilson, 2023). The four building blocks are the latent variable, observational design, scoring rules, and Rasch model.

Figure 7

Four Building Blocks for Constructing Scales



The first building block of defining a scale starts with the initial imagery of the latent variable (Lazarsfeld, 1958). This is a unidimensional scale because "they coincide with the use of unidimensional language in social science theories—language that is intended to clarify the meaning of those theories" (McIver & Carmines, 1981, p. 86). The scale for measuring learning stimulation in the home reflects a unidimensional scale with the Wright Map the concrete instantiation of the latent variable.

The next building block (observational design) involves the creation of a set of observable indicators or items to represent the latent variable. Observational designs frequently include various item classifications and domains that guide the creation of specific items. A classic example is the creation of educational achievement tests using item classifications based on Bloom's Taxonomy (Bloom et al., 1956). This building block highlights the idea that the items can be viewed as small experiments designed to determine a person locations on the latent variable.

The third building block specifies the scoring rules used to specify how the person responses are coded. For our example, the responses to the 11 items are simply scored dichotomously (Yes = present, No = not present). A response of Yes indicates the item is present in the home. The more items present in the home the higher the level of learning stimulation.

The final building block is the Rasch model that defines the measurement model used to connect the observed responses to the theoretical or expected values based on a measurement theory (Engelhard & Wang, 2021). The Rasch model is used to link the observed responses to items and persons based on their locations on a latent variable scale. As pointed out earlier, Rasch (1960/1980) started with a simple idea that a person's response to an item depends on the difficulty of the item and the ability of the person. He selected a probabilistic model based on the logistic response function because of its desirable properties.

In summary, these four building blocks can be used to construct a scale that can be represented by a Wright Map when Rasch measurement theory is used. The basic questions underlying the building blocks are as follows:

- What is the latent variable being assessed?
- What is the plan for collecting structured observations on items in order to define the latent variable?
- How are observations scored to represent person locations on the latent variable?
- How are person and item responses mapped onto the latent variable?

Residuals

Complicated phenomena … may be simplified by subducting the effect of known causes … and thus leaving, as it were, a residual phenomenon to be explained. It is by this process, in fact, that science, in its present advanced state is chiefly promoted. (Herschel, 1831, p. 156)

The residual pillar plays an important role in guiding researchers in their explorations and comparisons of competing explanations in science. Stigler (1981) identified John Herschel as an earlier pioneer in recognizing the importance of residuals. His book titled *A Preliminary Discourse on the Study of Natural Philosophy* (Herschel, 1831) gave particular emphasis to what he called residual phenomena. Residuals play an important role in the identification of anomalies that have been recognized as an important aspect of scientific progress (Kuhn, 1970). In some ways, the focus on residuals might be controversial because attention is directed toward what is left out of our models and theories. However, progress in science frequently occurs in the gray areas that are unexplained by a particular model. Residual analyses can be a useful tool in exploring and comparing competing explanations in science.

Stigler (1981) uses this pillar to describe how statistics uses the concept of residuals to foster scientific logic by the use of statistical methods for model comparisons. Residuals play an essential role in examining model-data fit with diagnostic displays of the differences between expected and observed responses. According to Stigler (2016):

> Statistics has made residual analyses into a new and powerful scientific method that has changed the practice of science. The statistical interpretation of this idea, and the associated scientific models, has given it a new disciplinary force. The statistical approach is to describe the process that generates the data by a hypothetical model and examine the deviation of the data from that model either informally (for example by a graphical or tabular display) or by a formal statistical test, comparing the simpler model with a more complicated version (a comparison among two "nested" models, one being a special case of the other).The earliest instances involved small, focused nested models, where one theory is to be compared to a slightly more complicated version. (p. 173)

Residual analyses provide model-based diagnostics for summarizing and plotting residuals. For example, it is common practice to plot the residuals, and to see what the patterns emerge in the residuals.

In measurement, a variety of ways have been developed to examine model-data fit based on the foundational pillar of residual analyses. In the case of well-developed models, such as the Rasch model, a variety of methods can be used to evaluate model-data fit based on the analyses of residuals. As pointed out by Stigler (2016), "when we can limit attention to a few alternatives or to well-structured parametric models, we are comfortably at home" (p. 171).

Figure 8 illustrates how residuals are defined. The basic data structure in measurement consists of a person by item matrix with observations or responses as the cell entries. The residuals (R_{ni}) are defined as the difference between the observations (X_{ni}) and the expectations based on the Rasch measurement model (P_{ni}). This can be written as:

$$R_{ni} = X_{ni} - P_{ni} \,. \qquad (13)$$

These residuals can be summarized and plotted in a variety of ways in order to identify

Figure 8

Definition of Residuals in Measurement

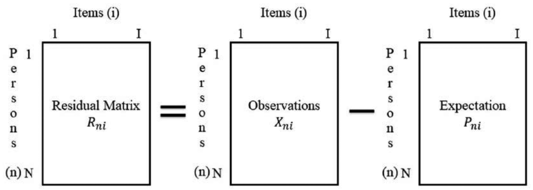
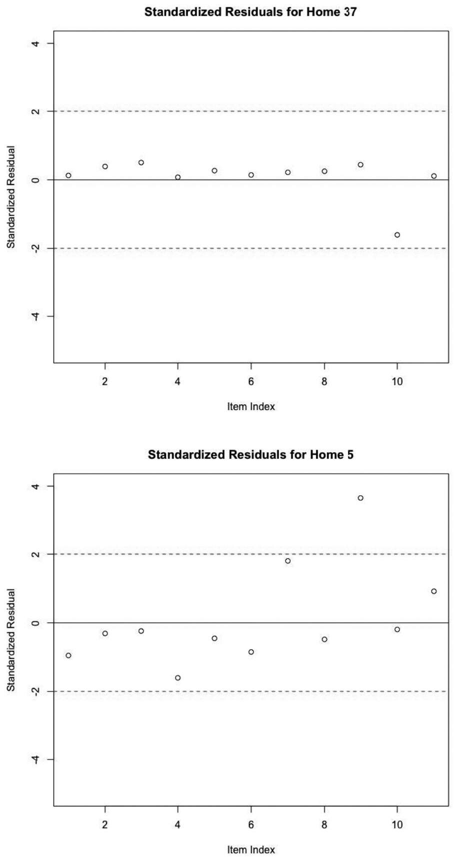


Figure 9

Scatterplots of Standardized Residuals for Person Fit Based on Home Data (Engelhard, 2013)



areas of misfit related to items and persons (Engelhard & Wang, 2021).

Figure 9 illustrates the use of standardized residuals for the measurement of learning stimulation in homes with pre-school children. The focus is on homes with notably low or high MSE fit statistics. Standardized residuals within the range of −2 to 2 indicate good fit, while values outside this range are unexpected based on the model. These outliers may merit further exploration. For Home 37 (Figure 9, Panel A), the standardized residuals were close to zero for all items except for Item 10. Home 5 (Figure 9, Panel B) on the other hand has relatively high misfit statistics, and a mix of both positive and negative standardized residuals, one of which exceeded 2.00.

Distinctive Pillars of Educational Measurement

In order to consider distinctive measurement pillars, educational measurement can be used as an illustrative area of measurement practice. Specifically, two distinctive pillars are discussed in this section. These are the power and consequences of measurement. Both of these pillars are influenced by the uses of educational assessments within a variety of contexts.

Power

Testing is a two-edged sword that can do incalculable good as well as great harm to the individual. (Bloom, 1970, p. 25)

Educational testing has a variety of functions with a focus on the role of assessment in improving educational processes broadly conceived. For example, assessments can be used to improve student learning by identifying what students know and can do, as well as what to do next. Assessments are also used to inform educational policies by evaluating teachers, schools, and broader educational entities. In some cases, educational testing is used to inform promotion and admission decisions. Assessments have a long usage in the measurement of language proficiency that can be used for immigration and admission decisions to universities. It is important to be aware of the power that measures exert over almost every aspect of our lives (Porter, 1995).

The evaluation of educational and psychological tests is guided to a large extent by the Test Standards (American Educational Research Association et al., 2014). The Test Standards feature three major sets of criteria (Validity, Reliability, and Fairness), but it can be argued that validity is the umbrella term that encompasses other types of evidence including reliability and fairness regarding the appropriate uses of test scores. Validity is defined as follows:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. (American Educational Research Association et al., 2014, p. 11)

As pointed out by Porter (1995), “The Latin root of validity means ‘power.’ Power must be exercised in a variety of way to make measurement and tallies valid” (p. 33). In addition to the technical aspects used for evaluating validity, it is important to consider the notion of validity as a type of power.

In multiple cases, tests become a key tool for guiding and enforcing issues related to power over individual and societal decisions. More than 50 years ago, Ben Bloom at the University of Chicago compared this power of testing with the power of atomic energy. He argued persuasively that measurement has the potential for positive benefits but that as with all technologies, it is important to consider the potential harm (Bloom, 1970).

Consequences of Measurement

The consequential aspect of construct validity includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use. (Messick, 1995, p. 746)

The use of tests also has consequences—some intended and others unintended. Measurement matters because it affects how we define the constructs we use to understand our world, how we create measures to represent these constructs, and ultimately how we evaluate the consequences of using these measures in a variety of contexts to inform decisions and policies. As pointed out by Messick (1995), it is important to consider both the intended and unintended consequences of measurement.

The sociologist Robert Merton introduced the concept of unanticipated consequences related to purposive social action (Merton, 1936). In the context of educational measurement, the focus is on formally organized social actions and their consequences. For example, it is expected that educational assessments will improve student learning, although some tests are primarily used for accountability with less clear connections to the improvement of student learning.

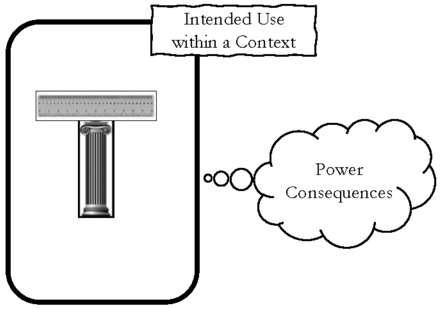
Merton (1936) identified two methodological pitfalls that are common to investigations of purposive social action. These pitfalls are casual imputation and identification of purpose. The first pitfall involves the issue of casual imputation. In his words, causal imputation is “the problem of ascertaining the extent to which ‘consequences’ may justifiably be attributed to certain actions” (p. 897). It is assumed that testing can improve student learning; however, it is remarkably difficult to document the connections between a high school graduation test and specific changes in educational processes that lead to changes in instruction and student behaviors.

The second pitfall is the error of the imputation of purpose. This pitfall relates to the difficulty in identifying the actual purposes of a given action. These pitfalls may lead to unanticipated consequences of purposive social action. It is clear that schooling is purposive social action and that testing has emerged as an integral part of the actions and activities of schooling. Both intended and unintended consequences may arise from the use of educational tests as mechanisms of social action. For example, accountability systems may have the unintended consequence of narrowing the curriculum when the focus is on teaching to the test. Another example is washback in language testing (Messick, 1996).

In summary, Figure 10 provides a representation of the two pillars of educational measurement: Power and Consequences. An important aspect of this perspective is the inclusion of the intended use with different educational contexts. Issues of power, as well as intended and unintended consequences, should be considered for all educational assessments. Serious attention and consideration should be directed to the potential unintended consequences of the purposeful use of testing to improve educational processes.

Figure 10

Pillars of Measurement Wisdom: Context, Power, and Consequences



Summary

Whatever exists, exists in some amount. To measure it is simply to know its varying amount. Man sees no less beauty in flowers now than before the day of quantitative botany. (Thorndike, 1921, p. 379)

The purpose of this study was to discuss the connections between statistics and measurement based on the seven pillars of statistical wisdom identified by Sigler (2016). In particular, this study considers three key questions:

- What are the pillars of statistical wisdom?
- What are the connections between the pillars of statistical wisdom and the pillars of measurement wisdom?
- Are there additional pillars distinctive to educational measurement?

Table 3 summarizes the seven pillars of wisdom by two fields of study: statistics and measurement. The pillars are as follows: Aggregation, Likelihood, Information, Intercomparisons, Regression, Design, and Residuals.

In response to the first question, Table 3 (Column 2) summarizes the major statistical interpretations of the pillars. The first pillar of aggregation refers primarily to methods of data summary. Likelihood refers to methods used for the calibration of a probability scale to evaluate inferences. The next pillar, information, provides a framework for the quantification of uncertainty related to our inferences. Intercomparisons provide a logical basis for using internal variation within a data set to specify a probability scale and quantify uncertainty. The pillar of regression is one of the key inventions of statistics that include widely used linear models for continuous and categorical data. Next, design provides a set of guidelines for data collection.

Table 3

Summary of Pillars of Wisdom as Viewed Within the Statistics and Measurement Fields

	Field of Study	
Pillars	Statistics	Measurement
Aggregation	Data summary	Latent variable
Likelihood	Calibration of a probability scale	Wright Map
Information	Quantification of uncertainty	Estimation of uncertainty of model parameters, such as standard errors of measurement
Intercomparisons	Using internal variation to specify probability scale	Invariance; differential item and person functioning
Regression	Linear Models	Rasch Models; explanatory item response models
Design	Design of Data Collection	Constructing Scales
Residuals	Logic of model comparison	Model-Data Fit

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

Finally, the last pillar (residuals) provides a logical framework for model comparisons by examining and summarizing the differences between expected and observed data.

Table 3 (Column 3) summarizes the major interpretation of the seven pillars for measurement (Question 2). There are close connections between the statistics and measurement pillars. In the field of statistics, the first pillar of aggregation can be viewed as data summary in terms of the creation of latent variable that underlies measurement. The likelihood process can be used to define a Wight Map. The Wright Map is a concrete representation of the probabilistic connections between the parameters in a measurement model and the observed data. Next, the information pillar can be used to quantify the uncertainty of model parameters, such as standard errors of measurement for person and item estimates. Next, intercomparisons provide a similar foundation that relates to the invariance of person and item locations on the underlying latent variable defined by the Wright Map. The regression pillar is also foundational in measurement. Item response models, such as the Rasch Model, are fundamentally generalized linear models for ordered categorical data. Next, the design pillar provides guidance for the construction of scales. The final pillar of residuals plays a key role in evaluating model-data fit in measurement.

The answer to the third question is that there are additional distinctive pillars of educational measurement. This study identified two additional pillars based on the idea that the use of measures leads to a consideration of power and consequences of testing (Engelhard & Behizadeh, 2017; Engelhard & Wind, 2013). Power stresses the use of measures to define and construct the key constructs that are used to structure the world around us. Consequences refer to the idea that measures are created to serve specific purposes and that the consequences may be both positive and negative. A key consideration is the identification of unintended consequences when our measures are used in practice.

In summary, this study highlights the strong connections between the pillars of statistical and measurement wisdom. The set of seven pillars of statistical wisdom is relevant for understanding measurement. It is important to recognize that there are important aspects of educational measurement that goes beyond the foundational aspects of statistics and measurement, such as a consideration of the power and consequences of measurement. It should be stressed that although the pillars are discussed separately, it is important to reflect on the cross-fertilization of the pillars for both fields of study. A key question remains: Are we any wiser now after considering these foundation pillars? The pillars lay the foundations of both fields, but the application of the pillars in various combinations to practice, such as educational measurement, highlights other essential pillars that merit attention. The pillars on their own do not necessarily constitute wisdom, but they are a good start.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. AERA.

Andrich, D. (2018). A Rasch measurement theory. In F. Guillemin, A. Leplege, S.

Briancon, E. Spitz, & J. Coste (Eds), *Perceived health and adaptation in chronic disease* (pp. 66–91). Routledge.

Azen, R., & Walker, C. M. (2010). *Categorical data analysis for the behavioral and social sciences*. Routledge.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. CRC press.

Bloom, B. S. (1970). Toward a theory of testing which includes measurement-evaluation-assessment. In M. C. Wittrock & D. Wiley (Eds.), *Evaluation of instruction: Issues and practices* (pp. 25–50). Holt, Rinehart and Winston.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, the classification of educational goals: Handbook I: The cognitive domain*. David McKay.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*(1), 605–634.

Cattell, J. M. (1893). Mental measurement. *Philosophical Review*, *2*, 316–332.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Society for Industrial and Applied Mathematics.

Efron, B. (2003). Second thoughts on the bootstrap. *Statistical Science*, *18*(2), 135–140. http://www.jstor.org/stable/3182843

Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Engelhard, G., Jr., & Behizadeh, N. (2017). Cogitations on power and consequences of measurement. *Measurement: Interdisciplinary Research and Perspectives*, *15*(1), 5–9.

Engelhard, G., Jr., & Wang, J. (2021). *Rasch models for solving measurement problems: Invariant measurement in the social sciences*. Sage Publications.

Engelhard, G., Jr., & Wind, S. A. (2013). Educational testing and schooling: Unanticipated consequences of purposive social action. *Measurement: Interdisciplinary Research and Perspectives*, *11*, 1–6.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(594–604), 309–368.

Fisher, R. A. (1935). *The design of experiments* (2nd ed.). Oliver and Boyd.

Fisher, R. A. (1970). *Statistical methods for research workers* (14th ed.). Hafner Press.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

Herschel, J. (1831). *A preliminary discourse on the study of natural philosophy*. Longman.

Köbel, J. (1460–1533). *Geometry book of Jacob Köbel* (1608 ed.) [Illustration]. Mathematical Association of America. https://maa.org/press/periodicals/convergence/the-right-and-lawful-rood

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). University of Chicago Press.

Lawrence, T. E. (1926). *The seven pillars of wisdom*. London.

Lazarsfeld, P. F. (1958). Evidence and inference in social research. In D. Lerner (Ed.), *Evidence and inference* (pp. 107–138). The Free Press.

McCullagh, P., & Nelder, J. A. (1983). *Generalized linear models*. Chapman Hall.

McIver, J., & Carmines, E. G. (1981). *Unidimensional scaling* (Vol. 24). Sage.

Merton, R. K. (1936). The unanticipated consequences of purposive social action. *American Sociological Review*, 894–904.

Merton, R. K., Sills, D. L., & Stigler, S. M. (1984). The Kelvin dictum and social science: An excursion into the history of an idea. *Journal of the History of the Behavioral Sciences*, *20*(4), 319–331.

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.

Messick, S. (1996). Validity and washback in language testing. *Language testing*, *13*(3), 241–256.

Moore, D. S., & McCabe, G. P. (1993). *Introduction to the practice of statistics* (2nd ed.) W. H. Freeman and Company.

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Rasch, G. (1961). On general laws and meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley Symposium on mathematical statistics and probability* (pp 321–333). University of California Press.

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

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

Savage, L. J. (1976). On rereading R. A. Fisher. *The Annals of Statistics*, *4*(3), 441–500.

Stigler, S. M. (1981). Gauss and the invention of least squares. *The Annals of Statistics*, *9*(3), 465–474.

Stigler, S. M. (1986). *The History of statistics: The measurement of uncertainty before 1900*. Harvard University Press.

Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Harvard University Press.

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

Thorndike, E. L. (1921). Measurement in education. *Teachers College Record*, *22*, 371–379.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Publishing Company.

Tutz, G. (2012). *Regression for categorical data*. Cambridge University Press.

Wasserman, L. (2010). *All of statistics: A concise course in statistical inference*. Springer.

Wilson, M. (2023). *Constructing measures: An item response modeling approach* (2nd ed.). Routledge.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. MESA Press.