THE ROLE OF MATHEMATICAL MODELS IN MEASUREMENT: A PERSPECTIVE FROM PSYCHOMETRICS

Mark Wilson

University of California—Berkeley, Berkeley, USA

Abstract − The paper uses the functional concept of a measuring system, as developed by Mari [1], to explicate the logic of several measurement approaches used in psychometrics, and thus enable a comparison with measurement approaches used by other fields such as engineering and physics. This characterization contrasts with the stereotype of measurement in the social and behavioral sciences, which is seen (from without) as typically following the representational viewpoint. The paper surveys Guttman Scaling, Classical Test Theory, Rasch Scaling and Construct Modeling, as examples of measurement approaches in the area of psychometrics, and explicates the underlying standard reference set that is one of the essential features of Mari's formalization, and shows how these differ among the four approaches. The importance of these differences, and the consequences for measurement using those approaches are also explicated and discussed.

Keywords measuring system, psychometrics

1. INTRODUCTION

The representational viewpoint [2] is widely seen as dominating the formal modeling of measurement in psychology and the behavioral sciences in general, and in psychometrics in particular. This is quite understandable, as the philosophical works of the authors of Ref. [2] stand head and shoulders above the works of any others in this area. Not that there are not significant works by others, but that the scope and comprehensiveness of [2] is generally seen as being without equal.

However, this dominance in formal modeling has little or no correspondence with the reality of most actual measures that are constructed in these domains. The sad state of philosophizing in the area of psychometrics is that the philosophical grounding provided by these giants of the field is "More honor'd in the breach than the observance" [3]. In fact, it is very difficult to find examples of applications of the representational approach beyond the works of the authors of [2]. One reaction to this has been for some authors to amend the tenets of the representational approach to incorporate a probabilistic element (e.g., [4], [5]). Another reaction has been to seek alternative philosophical bases for measurement, such as "scientific realism" (e.g., [6], [7]). The debate about this is still in its early stages, with several presentations given and planned at psychometric conferences, and only little of it yet having reached publication (though see [8] for some background to this debate).

As a contribution to this debate, this paper utilizes the functional concept of a *measuring system*, as developed by Mari [1], to explicate the logic of several measurement approaches used in psychometrics, and thus establish grounds for the comparison of these with measurement approaches used by other fields.

2. BRIEF BACKGROUND

The operation of a measuring system (MS), as described by Mari [1], is summarized by Figure 1.

$$\begin{tikzcd}
\text{Standard Reference Set } (\boldsymbol{A}) \arrow[d, dash] & \\
\text{Measuring System (MS)} \arrow[r] \arrow[d, dash] & \text{Measurement Result (in } \Theta) \\
\text{Measured thing } (\boldsymbol{u_p}) &
\end{tikzcd}$$

Fig. 1. Summary of the operation of a measuring system.

In this Figure, the evaluation of the "thing" is accomplished by the following 3-step procedure.

(1) Establish a *calibration*: create a "standard reference set" (of things) $\boldsymbol{A}$, by associating them with a set of symbols Θ. Generally these symbols will be the elements of a mathematical structure such as a set of integers or the real numbers along with their usual arithmetic relations: $m_I(a) = \theta$.

$$\begin{tikzcd}
\boldsymbol{A} \arrow[d] & \\
\text{MS} \arrow[r] & \Theta
\end{tikzcd}$$

Fig. 2. Establishing a calibration.

(2) Carry out *data acquisition*: through empirical interaction, select an element of $\boldsymbol{A}$ that corresponds to the thing $\boldsymbol{u_p}$ (for person p) using the "thing selection function $\chi(\ )$, i.e., $\chi(\boldsymbol{u_p}) = a$.

$$\begin{tikzcd} \chi(\boldsymbol{u_p}) = a \\ \text{MS} \arrow[u] \\ \boldsymbol{u_p} \arrow[u] \end{tikzcd}$$

Fig. 3. Data acquisition.

(3) Make the *data presentation* (i/e/. re-use calibration): select the symbol to measure by compounding 1 and 2: $m(\boldsymbol{u_p}) = m_I(\chi(\boldsymbol{u_p})) = \theta$.

$$\begin{tikzcd} \text{a} \arrow[d] & \\ \text{MS} \arrow[r] & \theta \end{tikzcd}$$

Fig. 4. Data Presentation.

What makes this a measuring system is that this mapping m is a homomorphism between the empirical relational structure of things, ($\boldsymbol{U}$, $\boldsymbol{R_U}$), and the symbolic relational structure of symbols ($\boldsymbol{\Theta}$, $\boldsymbol{R_\Theta}$): that is $r_U(\boldsymbol{u_p})$ whenever $m(r_U)(\mathrm{m}(\boldsymbol{u_p}))$. In Mari's approach, evaluations are measurements if and only if this latter is the case. In the following we will see how this looks for some measurement approaches that have been used in psychometrics.

3. GUTTMAN SCALING.

For all the examples of measurement approaches in this paper, the basic situation under consideration is that we have an instrument composed of a set of items, $\boldsymbol{I} = \{I_1, \ldots I_I\}$ for which prior substantive theory indicates that a person's response to those items (indicated by the response vector $\boldsymbol{u_p}$ for person p) is an indicator of the construct to be measured, Θ (i.e., the measurand). Without loss of generality, I will assume that the responses to the items are dichotomous, that is 0 or 1.

In Guttman Scaling [9] the essential idea is that, on the basis of substantive theory and practical knowledge about items, one can order, in terms of expected responses, a set of items from "easiest" to "hardest" (or "least positive" to "most positive," etc., depending on the context). Then the ordered set of possible response vectors is assumed to constitute a standard reference set, directly indicating the value of Θ in terms of the rank of the "highest" item for which the response is 1:

$$m(\boldsymbol{u}_p) = guttmanscore(\boldsymbol{u}_p),$$

which is the rank of $\chi(\boldsymbol{u}_p)$ minus 1 if it is defined, and undefined otherwise. As there is a finite number of items, I, the usual symbol set is the integers 1 to $I+1$. The standard reference set can be written as $\boldsymbol{A} = \{\boldsymbol{a}_0, \boldsymbol{a}_1, \ldots \boldsymbol{a}_I\}$ where

$$\boldsymbol{a}_0 = \{(0, 0 \ldots \quad 0)\}$$
$$\boldsymbol{a}_1 = \{(1, 0 \ldots \quad 0)\}$$
$$\boldsymbol{a}_2 = \{(1, 1, 0 \ldots\ 0)\}$$
.
.
.
$$\boldsymbol{a}_I = \{(1, 1, \ldots 1, 1)\}.$$

Note the response vectors in $\boldsymbol{A}$ are called "Guttman true scale-types." There is a frustrating incompleteness in this approach, as the there are many possible response vectors that are not scale-types, response such as (1, 0, 1, 0 …0), etc. Hence, one aspect of the creation of the instrument $\boldsymbol{I}$ is the selection of items that are suitable for use in a Guttman scale—intuitively, one would seek items that are increasing in "difficulty" as one went from the first to the last item, and where the increments in difficulty were as large as possible (although this will become more difficult to achieve if the number of item, I, is large). Guttman recommended the use of a quantitative indicator, the *coefficient of reproducibility* [9], which is the ratio of the observed number of true scale-type response vectors to the total number of responses, to gauge the suitability of the set of items for use in a Guttman scale, both in a relative sense (i.e., sets with larger coefficients are better), and in an absolute sense (e.g., accept item sets with coefficient values greater than .85).

The problem of what to do with persons with non scale-types has led to the low usage of this approach in most areas of application. According to Kofsky [10, pp. 202-203],

> … the scalogram model may not be the most accurate picture of development, since it is based on the assumption that an individual can be placed on a continuum at a point that discriminates the *exact* [emphasis added] skills he has mastered from those he has never been able to perform. ... A better way of describing individual growth sequences might employ probability statements about the likelihood of mastering one task once another has been or is in the process of being mastered.

The next approach can be seen as a step towards dealing with this problem, although historically it much predates Guttman Scaling.

4. CLASSICAL TEST THEORY

In classical test theory (CTT) [11–14], the problem of what to do about non scale-types is finessed by simply ignoring it, and response vectors are given a symbol (usually called a "score") that is equal to what the Guttman score would have been if the 1s and 0s had been ordered according to a scale-type, or equivalently, the sum score of the response vector: $m(\boldsymbol{u_p})$ = *sumscore* $(\boldsymbol{u}_p) = \Sigma\, u_{pi}$, where u_{pi} is the i^{th} response in the vector $\boldsymbol{u}_p$. The symbol set $\boldsymbol{\Theta}$ will then be the integers from 0 to the maximum score I. The standard reference set can be written as $\boldsymbol{A} = \{\boldsymbol{a}_0, \boldsymbol{a}_1, \ldots \boldsymbol{a}_I\}$ where

$\boldsymbol{a}_0 = \{(0, 0 \ldots \quad 0)\}$
$\boldsymbol{a}_1 = \{(1, 0 \ldots \quad 0), (0, 1 \ldots \quad 0), \ldots (0, 0 \ldots \quad 0, 1)\}$
$\boldsymbol{a}_2 = \{(1, 1, 0 \ldots 0), (1, 0, 1, 0 \ldots 0), \ldots (0, 0 \ldots 0, 1, 1)\}$
.
.
.
$\boldsymbol{a}_I = \{(1, 1, \ldots 1, 1)\}$.

Note that this standard set now accounts for every possible response vector under the assumptions. Just as for Guttman scaling, the item set $\boldsymbol{I}$ is typically a smaller set that results from some item selection based on empirical data. However, unlike Guttman scaling the criteria are not based on substantive theory about the interpretation of the items. Instead the standard criteria are based on statistical considerations having to do with various aspects of uncertainty. These are grouped together under the term "item analysis" [15]—typical criteria are:

(a) reliability of the item set, *reliability*($\boldsymbol{I}$),
(b) discriminations of the items, discrimination(I_i),
(c) etc.

Uncertainty in the measure is estimated in terms of the *standard error of measurement* (*sem*) for the scores,

$$sem = \mathrm{S}\sqrt{1-r},$$

where S is the standard deviation of the scores, and r is the reliability [15].

Using this concept, each measure should be more accurately expressed as a binary: (θ, *sem*). Effectively, this moves the symbol set $\boldsymbol{\Theta}$ beyond the set of integers in the interval [0,I], as it was above, to encompass the segment of the real number line, [0-*d*, *I*+*d*], where the value of *d* is dependent on the level of uncertainty one wishes to express. The most common representation of CTT is X_p=T_p+E_p where X_p, is the observed score for person *p* (or, *sumscore*($\boldsymbol{u}_p$) above), E_p is the error (or *sem* above) and T_p is the "true score" (i.e., theoretical average of person *p*'s observed score over a large (infinite) number of observations).

One alternative to the sum score symbols that is often used is the percentile. This is simply the value of the cumulative distribution function for the sum score in a chosen reference sample of persons, expressed as a percentage. This conceals the differences between instruments that have different numbers of items, and can be used as a basis for "equating" same. For this approach, the symbol set (ignoring the *sem* issue) is the real numbers between 0 and 100. All of the remaining points above hold, however.

Sometimes, when a decision is to be made on the basis of the measures whether a person is "above" some point (or, equivalently, "below"). This requires the setting of a *cut-score*. If that is the sole purpose of the instrument, this can be seen as equivalent to establishing a new, coarser, standard reference set $\boldsymbol{A'}$, where (assuming the cut-off is *k*)

$\boldsymbol{a'}_0 = \{\boldsymbol{a}_0, \boldsymbol{a}_1 \ldots \quad \boldsymbol{a}_k\}$
$\boldsymbol{a'}_1 = \{\boldsymbol{a}_{k+1}, \boldsymbol{a}_{k+2} \ldots \boldsymbol{a}_I\}$.

The cut-score will usually be set using a procedure that invokes substantive knowledge from among professionals in areas related to the construct and the typical applications.

5. RASCH SCALING.

In Rasch scaling [16, 17] the CTT approach is amended and extended to (a) formalize the relationship between the person and the item (i.e., rather than the instrument as a whole) using a mathematical model, (b) adopt a metric (specifically the log of the odds) that frees the scale from a dependence on the (largely) incidental aspect of the number of items, and (c), as a result of (a), bring the item and the person parameters onto the same metric, allowing a wide range of possibilities for the development of analogical and figurative aids to interpretation. The mathematical relationship is given by

$$\Pr(u_{pi} = 1 | \theta_p, \delta_i) = \pi_{pi1} = \exp(\theta_p - \delta_i)/\gamma_{pi},$$

where θ_p is the person symbol (often called the "location", or the "ability" depending on the context), δ_i is an item parameter (often termed the "difficulty"), and γ_{pi} is a norming value equal to $[1 + \exp(\theta_p - \delta_i)]$. The connection to the log-odds is immediately seen as:

$$\log(\pi_{pi1}/\pi_{pi0}) = \theta_p - \delta_i$$

(where π_{pi0} has an obvious definition). The probability of a response vector is given by applying the local independence assumption:

$$\Pr(\boldsymbol{u}_p | \theta_p, \delta) = \prod_{i=1}^{I} \Pr(u_{pi} | \theta_p, \delta_i),$$

where δ is the vector of item parameters. In psychometric modeling, many other functions (termed "item response models") besides the simple logistic function are used (see e.g., [18]), but the Rasch model stands out due to (a) its simplicity, and (b) its unique properties, such as "specific objectivity" which confers particular strengths on item sets that are found to be amenable to Rasch modeling (i.e., "fit" the model). Further comment on other item response models is found below. As for CTT, a symbol is found for each possible response vector, hence the standard reference set is the same as for CTT (see above). Unlike the case for CTT, the symbols are not automatically assigned via a simple explicit function, but must be statistically estimated [18], and may take values anywhere on the real number line (-∞, ∞)[1].

[1] The minimum and maximum cases, (0, 0 … 0) and (1, 1, … 1) receive these symbols,

Just as for Guttman Scaling and CTT, the item set $\boldsymbol{I}$ is typically a smaller set that results from some item selection based on empirical data. As for CTT the criteria are based on statistical considerations having to do with various aspects of uncertainty. These are grouped together under the term "item fit analysis" [17, 18], and generally capitalize on the probability of response vectors to accumulate the likelihood of observing the responses to a given item, given its difficulty, and the person parameters for the observed persons (more traditional item analysis procedures are also commonly used). Other considerations are also involved, such as the match between the observed and expected distribution of persons on the logit scale, and the match of item locations to that distribution[2]. Uncertainty in the measure is expressed in terms of the standard error for the person estimates, which is found as a by-product of the estimation procedures, and differs from the for CTT in that it is conditional on the location itself: $s(\theta)$. As for CTT, each person's measure should be more accurately expressed as a binary: $(\theta, s(\theta))$.

In a step back from CTT towards Guttman Scaling, these symbols are not accorded equal standing—each response vector has a probability of being observed (conditional on the estimated parameters δ). Thus, some response vectors will have smaller probabilities than others—this would be the case, for example for the response vector (0, 0 … 0, 1) in the case where the items are ordered in terms of item difficulty. This fact has been used as the basis for developing several "person fit statistics" [18, 19], which are then used to decide that some persons with low-probability response vectors should then not be assigned symbols (estimates). This results in a reduced standard set $\boldsymbol{A'}$, although the number of symbols (estimates) remains the same. Thus the standard reference set can be written as $\boldsymbol{A'} = \{\boldsymbol{a'}_0, \boldsymbol{a'}_1, \ldots \boldsymbol{a'}_I\}$ where

$$\boldsymbol{a'}_0 \subset \boldsymbol{a}_0 = \{(0, 0 \ldots \quad 0)\}$$
$$\boldsymbol{a'}_1 \subset \boldsymbol{a}_1 = \{(1, 0 \ldots \quad 0), (0, 1 \ldots \quad 0), \ldots (0, 0 \ldots \quad 0, 1)\}$$
$$\boldsymbol{a'}_2 \subset \boldsymbol{a}_2 = \{(1, 1, 0 \ldots \ 0), (1, 0, 1, 0 \ldots 0), \ldots (0, 0 \ldots 0, 1, 1)\}$$
$$\vdots$$
$$\boldsymbol{a'}_I \subset \boldsymbol{a}_I = \{(1, 1, \ldots 1, 1)\},$$

and the rules for reducing the sets $\boldsymbol{a}_0$ through $\boldsymbol{a}_I$ are determined by the specific person fit procedures chosen.

When these person fit rules are applied, Rasch Scaling represents an interesting, and potentially powerful, compromise between the strictness of the Guttman Scale adherence to the pre-eminence of the substantive theory (via the item ordering implied by the substantive interpretation), and CTT's flexibility in accepting all response vectors. Thus, Rasch Scaling has been seen as a way to reconcile the perspectives of CTT and Guttman Scaling [20].

6. CONSTRUCT MODELING

Construct Modeling [20, 21] builds upon the reconciliation of CTT and Guttman Scaling, as represented by Rasch Scaling. It takes as its technical side the ground-work of Rasch Scaling, and moves one step further along the path towards adhering to the substantive theory. In this case, it is assumed that the substantive theory takes a particularly simple form: The construct consists of a simple linear succession of discrete segments of a continuum, from a lowest level to a highest level, and when these are laid out in a figure, it is termed a "construct map" —see Figure 5. A concrete example of this is shown in Fig. 6, which was developed in the context of a test of students' knowledge about buoyancy. In this case, the Rasch Scaling described above serves as a starting place for establishing the standard set. All of the steps above for Rasch Scaling are followed. Simultaneously with that, a second set of steps is followed that takes into account the construct map, and additional substantive information concerning each item, to wit, a substantive link from each item to a (single) level of the construct map.

```
| level K
| level K-1
|
| level K-2
|
|
:
| level 2
|
|
|
| level 1
```

Fig. 5. A graphical representation of a construct map with K levels.

-∞ and ∞, respectively. (In practical terms, these symbols are not very useful, and practitioners typically either refrain from giving persons with those response vectors symbols, or they assign finite values to them, following certain procedures.)

[2] Note that the graphical device showing both persons and items on the same scale is sometimes referred to as a "Wright map."

What the student knows about Why Things Sink and Float
Knows how relative density affects floating and sinking in different liquids.
Knows how density affects floating and sinking in water.
Knows how the relationship of mass to volume affects floating and sinking.
Knows how volume affects floating and sinking when mass is held constant.
Knows how mass affects floating and sinking when volume is held constant.
Has productive misconceptions about why things sink or float.
Has fundamental misconceptions about why things sink or float.
Does not appear to understand any aspect of why things sink or float.

Fig. 6. A buoyancy construct map.

When deciding on the item set, an additional criterion that is used is this link from an item to the construct map—for each item it needs to be judged whether the estimated item location is well-matched to the hypothesized order in the construct map. Of course, accumulated empirical information that it does not match could lead to a revision of the construct map and/or the hypothesized link, as well as modification/deletion of the item.

Once an item set is established, "banding" or "standard setting" takes place—this is the equivalent of Mari's calibration: the placement of the values into segments of the logit scale (the "Wright Map"). The upper and lower limits of these bands are determined by the locations of the lower and upper limits of the locations of the items linked to each level of the construct map. Most often, these item locations do not result immediately in a "clean" segmentation of the logit scale—hence a judgmental process [22] is required to determine reasonable locations for the band edges[3]. This process may result in further decisions regarding the suitability of certain items, and may also involve information about the persons, when that is available. Denote these limits by $\Theta = (\Theta_1, \Theta_2, \ldots \Theta_K)$, the K boundary values between the K+1 segments. Then we see that:

$$\begin{aligned}&\text{Band 1 is } (-\infty, \Theta_1],\\ &\text{Band 2 is } (\Theta_1, \Theta_2],\\ &\quad\vdots\\ &\text{Band } k \text{ is } (\Theta_{k-1}, \Theta_k],\\ &\quad\vdots\\ &\text{Band } K{+}1 \text{ is } (\Theta_K, +\infty).\end{aligned}$$

Or, equivalently, the standard reference set is $\boldsymbol{A'}$, where (as above, assuming the items are ordered by their difficulty)

$$\begin{aligned}\boldsymbol{a'}_1 &= \{ \boldsymbol{a}_0, \boldsymbol{a}_1 \ldots \quad \boldsymbol{a}_{*1}\},\\ \boldsymbol{a'}_2 &= \{ \boldsymbol{a}_{*1+1}, \boldsymbol{a}_{*1+2} \ldots \boldsymbol{a}_{*2}\},\\ &\vdots\\ \boldsymbol{a'}_k &= \{ \boldsymbol{a}_{*k+1}, \boldsymbol{a}_{*k+2} \ldots \boldsymbol{a}_{*(k+1)}\},\\ &\vdots\\ \boldsymbol{a'}_K &= \{ \boldsymbol{a}_{*K+1}, \boldsymbol{a}_{*K+2} \ldots \boldsymbol{a}_I\},\end{aligned}$$

and {*1, *2,…*k,…*K, I} represents the number (order) of the item at the upper limit of the items in each of the respective Bands above (i.e., the highest response vector in Band K is symbol *k, etc.). These values are determined by finding the upper limit of the location of the items linked to each level of the construct map.

These bands become a basis for criterion-referenced interpretations of the measurements, enhancing and deepening the interpretations available to those who must apply the measurements. At the same time, the existence of the underlying Rasch scale means that (a) technical aspects of the measures are available, such as standard errors etc., and (b) technical advantages of item response scales are still available, such as flexibility in item choice, ability to link forms through items, and the possibility of computerized adaptive item administration. This is one reason to choose the construct modeling approach compared to the alternative of latent class modeling [24], which might be seen as potentially appropriate, given what is shown in Figure 5.

7. DISCUSSION AND CONCLUSION.

Mari [1, p. 80] characterizes measurement as needing to attain both *objectivity* and *intersubjectivity*. In his words:

> objectivity implies that the MS is able to discriminate the measurand from the various influence quantities so that the acquisition component of the MS is sensitive only to the measurand;
>
> intersubjectivity implies that the MS is able to refer the measurand to

[3] The label for this different depending on when the set $\boldsymbol{A}$ is developed: if it is developed before the scaling (although it may be adapted after), then the process is termed "construct modeling" [21]; if it is developed after the scaling, then the resulting scale is termed a "described variable" [22].

> the primary standard, so that all the measurements expressed in terms of that standard are comparable with each other.

From the point of view adopted here, we need to specify how these properties would appear in the context of Construct Modeling.

First, consider objectivity. If we think of "the various influence quantities" as being embodied by the possibility of using different sets of items as the instrument, then this amounts to independence from the specific set of items used. This is indeed what Rasch's "specific objectivity" is concerned with (i.e., if the set of items fits the Rasch model, then it does not matter which items are used to measure the person). Hence, when using a Rasch scale as the basis for Construct Mapping, (and where the items do indeed fit a Rasch model) it would remain to check whether the banding was relatively robust to choices of the items as representatives.

Second, consider intersubjectivity. Unlike the case for CTT, Construct Modeling inherits the Rasch Scaling (and Guttman Scaling) characteristic of disallowing some response vectors. However, these are not represented in the standard reference set, hence this is not a formal problem. In practical terms, this amounts to a situation where the measurement system does not give some people measures. The best procedure in this case is to seek a re-administration of the measurement process for that individual, with perhaps the possibility of gathering extra information to check on the conditions under which these misfit response vectors tend to be found.

In the text above, it was noted that there are many item response models available beyond the Rasch model. Where these are being used, some, but not all, of the development above in the Construct Modeling section can be developed. In particular, banding is not readily possible, as the concept of the "location" of an item on the logit scale does not have a straightforward interpretation. Also, the specific objectivity possible under the Rasch model is not attainable for other models [16]. Thus, for other item response model approaches, the development here seems difficult.

The generic type of construct (measurand) that is used to motivate the development of Construct Modeling (i.e., a simple linear succession of discrete segments of a continuum) may seem quite restrictive on first glance. However, most published measures in the social sciences are in fact of just this type, or simpler (i.e., they have no segments, just a continuum). That said, where there are more complex constructs under consideration, many of them represent quite simple extensions of the generic construct discussed above. For example, where there are multiple linear continua (i.e., a "multi-dimensional" construct), then the scaling can be accomplished using multi-dimensional versions of the Rasch model [25, 26], and each dimension can be treated then as a separate case for banding. Where there are polytomous items and/or multiple substantive categories within a particular polytomous score [27], the banding procedure can be generalized to deal with the situation [22]. Where the latent class is posited to be an ordered latent class rather than a latent continuum, the methods described above can be applied, with the proviso that one should check for the most appropriate model using fit procedures [28]. Where a more complex construct is under consideration, such as a "learning progression" [29] (which posits level-based links among different dimensions), there are also methods analogous to those described above, although these are still under development [30]. Of course, there are more complex constructs yet, but the list above contains a very large proportion of the extant types.

This paper has used the functional concept of a measuring system to explicate the logic of several measurement approaches used in psychometrics, and thus enable a comparison with measurement approaches used by other fields such as engineering and physics. It surveyed Guttman Scaling, Classical Test Theory, Rasch Scaling and Construct Modeling, as examples of measurement approaches in the area of psychometrics, and explicated the underlying standard reference set that is one of the essential features of Mari's formalization [1], and showed how these differ among the four approaches. The importance of these differences, and the consequences for measurement using those approaches, hinge on the capacity to identify theoretically tractable substantive properties capable of supporting both objectivity and intersubjectivity. Connecting psychometric approaches to measurement with Mari's formalization of the functional concept of a measuring system opens up new opportunities for productive dialogue between the natural and social sciences.

REFERENCES

[1] L. Mari, Beyond the representational viewpoint: a new formalization of measurement, Measurement 27 (2000) 71-84.

[2] D.H. Krantz, R.D. Luce, R. D., P. Suppes, A. Tversky, Foundations of measurement. Volume 1: Additive and polynomial representations. New York, Academic Press (1971).

[3] A. Thompson, N. Taylor, Hamlet, Arden, London (2006).

[4] R. Perline, B. Wright, H. Wainer, H., The Rasch model as additive conjoint measurement, Applied Psychological Measurement, 3 (1979) 237–256.

[5] G. Karabatsos, The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory, Journal of Applied Measurement, 2(4) (2001) 389-423.

[6] J. Michell, An introduction to the logic of psychological measurement. Lawrence Erlbaum Associates, Mahwah, New Jersey (1990).

[7] D. Borsboom, G. Mellenbergh, H. Van, The theoretical status of latent variables. Psychological Review, 110 (2) (2003) 203-.

[8] R. Lissitz, The concept of validity: Revisions, new directions, and applications, Information Age Publishing, Charlotte, North Carolina (2009).

[9] L. Guttman, A basis for scaling qualitative data, American Sociological Review, 9 (1944) 139-150.

[10] E. Kofsky, A scalogram study of classificatory development, Child Development, 37 (1966) 191-204.

[11] F. Edgeworth, The statistics of examinations, Journal of the Royal Statistical Society, 51 (1888) 599-635.

[12] F. Edgeworth, Correlated averages, Philosophical Magazine*,* 5th Series, *34* (1892) 190-204.

[13] C. Spearman, The proof and measurement of association between two things, American Journal of Psychology, 15 (1904) 72-101.

[14] C. Spearman, Demonstration of formulae for true measurement of correlation, American Journal of Psychology, 18 (1907) 161-169.

[15] J. Nunnally, I. Bernstein, Psychometric Theory 3rd edition, McGraw-Hill, Columbus. Ohio (1994).

[16] G. Rasch, (1960/1980). Probabilistic Models for Some Intelligence and Attainment Tests*,* University of Chicago Press, Chicago (1980).

[17] B. Wright, M. Stone, Best Test Design, MESA Press, Chicago (1979).

[18] W. van der Linden, R. Hambleton, Handbook of Item Response Theory, Springer, New York (1997).

[19] R. Meijer, K. Sijtsma, Methodology review: Evaluating person fit, Applied Psychological Measurement, 25 (2001) 107–135.

[20] B. Wright, G. Masters, *Rating Scale Analysis*. MESA Press, Chicago (1981).

[21] M. Wilson, Constructing Measures: An Item Response Modeling Approach, Erlbaum, Mahwah, New Jersey (2005).

[22] M. Wilson, K. Draney, A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), Measurement and multivariate analysis, pp 325-332, Springer-Verlag, Tokyo (2002).

[23] Organisation for Economic Co-operation and Development, PISA 2000 Technical Report, OECD, Paris (2002).

[24] M. Croon, Latent class analysis with ordered latent classes, British Journal of Mathematical & Statistical Psychology, 43(2) (1990) 171-192.

[25] R. Adams, M. Wilson, W. Wang, The multidimensional random coefficients multinomial logit model, Applied Psychological Measurement, 21(1) (1997) 1-23.

[26] M. Wu, R. Adams, M. Wilson, S. Haldane, (2008). ACERConQuest 2.0 [computer program], ACER, Hawthorn, Australia (2008).

[27] M. Wilson, R. Adams, Marginal maximum likelihood estimation for the ordered partition model, Journal of Educational Statistics, 18(1) (1993) 69-90.

[28] D. Torres Irribarra, R. Diakow, Model selection for tenable assessment: Selecting a latent variable model by testing the assumed latent structure, Paper presented at the 17th International Meeting of the Psychometric Society, Hong Kong SAR, (2011).

[29] A. Alonzo, A. Gotwals, (Eds.), Learning Progressions in Science, Sense Publishers, Rotterdam, The Netherlands (2011).

[30] M. Wilson, Responding to a challenge that learning progressions pose to measurement practice: hypothesized links between dimensions of the outcome progression. In A.C. Alonzo & A. W. Gotwals, (Eds.), Learning Progressions in Science, Sense Publishers, Rotterdam, The Netherlands (2011).

Author: Prof. Mark Wilson, Graduate School of Education, UC Berkeley, Berkeley, CA 94720, USA, +1-510-642-7966, MarkW@berkeley.edu.