preprint of a paper published in:
Measurement, 51, 315-327, 2014, doi:10.1016/j.measurement.2014.02.014

An introduction to the Rasch measurement approach for metrologists

Luca Mari[a#], Mark Wilson[b]
[a] School of Industrial Engineering, Università Cattaneo – LIUC,
C.so Matteotti, 22, 21053 Castellanza (VA), Italy
lmari@liuc.it +39 0331 5721
[b] Graduate School of Education, University of California, Berkeley
4415 Tolman Hall, Berkeley, CA 94720, USA
MarkW@berkeley.edu

Abstract: In the interests of fostering an inter-disciplinary dialogue, increasing collaboration between "hard" and "soft" measurement scientists, and learning from one another, the paper develops an analytical discussion of common elements between metrology and psychometrics. A simple example of physical measurement is introduced according to the conceptualization and terminology of the *International Vocabulary of Metrology* (VIM), and then its structural analogy to a test using Guttman items is shown. On this ground the example is generalized so to include a probabilistic component, and this leads to the basic Rasch model. Some notes on the delicate issue of measurement validity conclude the paper, whose aim, in the long run, is a measurement-related shared concept system, and a terminology understandable in both physical and social sciences.

Keywords: metrology; psychometrics; Rasch models; measurement validity

1. Introduction

Measurement science is a good context in which to consider, once again, the asymmetric relations between the natural and social sciences. The impressive effectiveness of their methods and instruments seems a sufficient reason for physicists, chemists, engineers, etc. to follow their path and be largely uninterested in developments in the measurement of non-physical properties. On the other hand, even though in many aspects emancipated from "physics envy", it is not unusual for social sciences to take physical measurement as a reference, and possibly a target point, given "their propensity to imitate as closely as possible the procedures of the brilliantly successful physical sciences" [1].

The fact that sensors implementing physical effects, the core components of physical measurement instrumentation, cannot be exploited for non-physical properties has discouraged passive imitation, and this has (at least in part) led to the development of different theories, methods, and instruments in the area of social measurement. The two disciplines grew up along parallel routes, sometimes approaching each other – a significant example is Finkelstein's endeavor to import representational theories in physical measurement – but also sometimes with clashes, as in the well known case of the committee activated by the British Association for the Advancement of Science in the 1930s (extensively discussed in [2]; a more concise analysis is in [3]), which produced, among other effects, the predominance of operationalism in psychological measurement for much of the 20th century, as well as Stevens' theory of scale types [4].

In the mentioned asymmetric situation, it may be interesting to continue exploring the contributions that social measurement has to offer to physical measurement. An excellent context in which to pursue this goal is so-called *Rasch measurement*, an approach to measurement developed within the social sciences that posits that the mathematical model of measurement is such that:

1. the result of the experimental stage of measurement, i.e., the indication, is given in probabilistic, instead of deterministic, terms;
2. measurands can be meaningfully compared by their ratio.

It is surely not a new subject and of which several introductory texts exist. However, these texts, and the scientific papers describing this approach to social measurement, are, in general, not always easily readable by physical measurement researchers and practitioners. There are at least two reasons of this difficulty:

One of the authors is a member of the Joint Committee on Guides in Metrology (JCGM) Working Group 2 (VIM). The opinion expressed in this paper does not necessarily represent the view of this Working Group.
A preliminary version of this paper was the basis for a keynote lecture jointly presented by the authors at the 2013 Joint IMEKO TC1 - TC7 - TC13 Symposium, 4-6 September 2013, Genova, Italy. Several enhancements resulted from the lively discussions at the Symposium.

- the emphasis in these texts on measurability specified in terms of algebraic conditions, whereas physical measurement is a moving target on this matter (for example ordinal measurement is routinely accepted nowadays);[1]
- some critical differences in the presentation of basic concepts and the related terms, so that for example "latent trait" (or "latent variable") and "manifest observation" are sometimes used in social measurement for "measurand" and "indication" respectively

Even the expression "Rasch measurement" sounds peculiar in metrology, where names are typically given to measurement principles (e.g., Peltier effect) and measuring instruments (e.g., Bourdon pressure gauge), and "*x* measurement" is reserved to *x* = given quantity, as in "force measurement". Apart from historical reasons, a possible justification of the expression "Rasch measurement" is that it can be thought of as referring to a combination of a *measurement method* and some assumptions on the underlying *measurement principle*. For this reason we will adopt here the more appropriate "Rasch measurement models", or the simpler "Rasch models". But are Rasch models actually *measurement* models? In the last part of the paper this delicate question is considered.

While we do not necessarily expect that Rasch measurement models would be immediately useful for physical measurement, a common, well founded understanding on them might foster more fruitful relationships between physical and social measurement, towards a desirable shared concept system and related terminology. This is the underlying purpose of the present paper, which introduces the basics of Rasch models by systematically interpreting them in the conceptual and lexical framework of the *International Vocabulary of Metrology*, third edition (VIM) [7], a freely accessible document that may be consulted in parallel to this paper (the first occurrence of terms taken from the VIM is in italics, so to ease the search of the corresponding definitions in the VIM).

The paper can be read as an interdisciplinary exploration of the concept of (mathematical) *measurement model* –"mathematical relation among all quantities known to be involved in a measurement" according to the VIM – particularly when specialized as a *measurement function*, i.e., the function that formalizes the (inverse) behavior of the *sensor* at the core of the *measuring instrument*, and that produces *measured quantity values* when applied to *indication values* and possibly values of other quantities such as *corrections* and *influence quantities*. The fact that this is a purely structural characterization makes it applicable in principle to both physical and social instruments: Rasch models are indeed measurement models in this sense, where typically *indications* are outcomes of tests (e.g., in the form of number of correct answers) and *measurands* are properties such as attitudes, abilities, ... of individuals. A whole family of models is termed after Rasch, all sharing this basic structure. In Section 4 the simplest of them will be presented. A by-product of the paper is then to show that a significant case of measurement in social sciences can be effectively spelled out in metrological terms. An admittedly simple, and somehow artificial, example of physical measurement will guide us to recognize the analogies between physical transducers and tests, as they can be understood as measuring instruments of Rasch models and psychometrics in general (to emphasize such analogies the symbols will be maintained to be consistent in the two cases, thus departing from the accustomed symbols in Rasch models). The conclusions drawn from this comparison will be devoted to the validation of measurand definitions / models, an issue that physical and social measurement usually approach with different strategies.

Our hope is that from what follows natural scientists and engineers may learn something of Rasch models, as a specifically relevant case of social measurement, and social scientists may re-interpret something of their knowledge of measurement in the light of the current physical measurement models.

---

1 An interesting example concerns the possible requirement that the scale is continuous, or at least its elements are dense (i.e., isomorphic to rational numbers), that in Holder's axioms is expressed as "For every magnitude there exists one that is less." (and note that Holder himself presents his aim in this way: "I intend only to propose a simple system of axioms from which the properties of the *ordinary continuum of magnitudes* can be derived." – emphasis added) [5]. This implies that counting cannot be a type of measuring and that discrete properties are not measurable. According to Joel Michell [2], "we have no good reason to suppose that measurable quantities are not continuous." This is in stark contrast to the views of many scientists. Consider, for example, the following quotation from Richard Cox, working from a different tradition: "Reflection suggests, indeed, that the only perfectly precise measurement is counting and that the only quantities defined perfectly are those defined in terms of whole numbers." [6].

2. Example 1: Hookean springs and Boolean springs

With the aim of measuring a given force f, a spring can be exploited as an *indicating measuring instrument*, and specifically as a sensor, which is supposed to behave according to a transduction function (sometimes also called "observation function") specified by Hooke's law:

$$x = \frac{f}{k} \tag{1}$$

i.e., the measurement principle is that a force f applied to a spring of elastic constant k generates an elongation x in the spring (we will omit *measurement units* from now on, but of course N is the unit of f, m is the unit of x, and Nm^{-1} is the unit of k).

The measurement principle is then that the measurand f is transduced by the spring of elastic constant k and generates an indication x: it is the usual input-output characterization of a device behavior, the measurand being the cause, and the indication the effect of the transduction process.

The relation between indication values and measured quantity values, together with *instrumental measurement uncertainty*, is provided by the *calibration* of the spring, that in this example (as long as uncertainty is neglected) consists in establishing a value for the constant k. As long as eq. (1) is maintained, this can be obtained by applying to the spring a single *measurement standard* which realizes a force f of known value, reading the corresponding value for the obtained elongation x, and inferring the value of the elastic constant k from the pair ⟨value of f, value of x⟩.

Once the instrument has been calibrated, and therefore a value for k is given, a *measurement* is performed by applying the measurand f to the spring, getting an indication x and finally exploiting the inverted version of the law:

$$f = kx$$

i.e., applying a measurement function:

measurand = measurement_function(indication, other_quantities)

so as to obtain a *measurement result* (in the simplest case a single measured quantity value, but more generally a more complex entity such as a measured quantity value and a *standard measurement uncertainty* or a whole probability distribution over the set of *quantity values*). Calibration, from which a measurement model is defined, is aimed at guaranteeing the *metrological traceability* of measurement results, and thus in particular their independence from any specific measuring instrument (whereas, of course, indications depend on both the measurand and the measuring instrument).

Even such a simple measurement is then both an experimental and a formal process, structurally based on the following inference:

1. if the measurement model is correct, and
2. the instrument is correctly calibrated, and
3. an indication value has been experimentally obtained, and
4. the measurement function is applied to the indication value
5. then a value for the measurand is obtained

(once again: we are neglecting measurement uncertainty, which generally should be considered in particular to take into account the effect of the influence quantities, such as the spring temperature, on the indication).

Accomplishing a measurement according to this inferential process assumes the premises of the process (and therefore in the present example in particular the scale type of the involved quantities and the linearity of their functional relation) as taken for granted. Were the *validation* of the measurement model or the instrument calibration the target, some other, independent process could be performed and its results properly analyzed.

The calibration process critically depends on the hypotheses assumed regarding the transduction function by which the spring behavior is modeled. For a linear, zero-fixed function, as in the case of transducers modeled as behaving according to Hooke's law, a single calibration point is sufficient to establish the slope of the straight line. Were only a generic causal dependence of x on f assumed, then according to a black box model of the behavior, the calibration should be performed by exploiting multiple measurement standards, each of them realizing a different value of force, so to obtain multiple pairs ⟨value of f_i, value of x_i⟩, to be suitably interpolated to produce the measurement model, in either analytical or numerical form. A decisive point here

is that performing measurement by means of a transducer does not require the transduction function to be analytically known. Of course, if the transduction function is known in its analytical form and is invertible in the measurement range, then the measurement function is the inverse of the transduction function:

indication = transduction_function(measurand, other_quantities).

But in general the only critical measurement-related requirement on the transduction process seems to be *its causality*, i.e., the assumption that the indication depends on the measurand, and therefore conveys information on it. This guarantees that at least in principle measured quantity values can be computed from indication values through the measurement model / function (the same concept is spelled out in different ways, e.g., "models are used to convert observations into measurements" [8]).
Suppose now that, instead of a spring whose behavior is described by Hooke's law, a modified spring is available, again characterized by a constant k related to its stiffness, operating according to the following transduction function:

$$\begin{cases} \text{if } \dfrac{f}{k} < 1, \text{ then the spring stays in its rest position, } x = 0 \\ \text{if } \dfrac{f}{k} \geq 1, \text{ then the spring elongates to a fixed position, } x = 1 \end{cases} \qquad (2)$$

Let us call such a transducer a "Boolean spring", whereas a "Hookean spring" will be the term for any transducer behaving according to eq. (1).
While the behavior of a Hookean spring is mathematically modeled as a continuous, linear function, eq. (2) defines a function whose range is discrete, and in fact binary. A second major difference between eq. (1) and eq. (2) is related to the *dimension* of the parameter k: in the case of Hookean springs, $\dim f / k = \dim x = \mathsf{L}$, and therefore $\dim k = \mathsf{MT}^{-2}$. On the other hand eq. (2) assumes that $\dim f / k = 1$ (i.e., is a *quantity of dimension one* – sometimes the term "dimensionless quantity" is used in this case), so that $\dim k = \dim f$ for Boolean springs. The fact that now the parameter k is dimensionally homogeneous to a force has the important consequence that it can be interpreted as a "threshold force", such that the spring elongates only if the applied force is greater than the threshold. This supposition will be crucial for what follows, as it allows the comparison of the involved quantities not only through ratios ($f / k > 1$) but also through differences ($f - k > 0$) and orderings ($f > k$), and therefore makes it possible to place values of the measurand and the parameter of the measuring instrument in the same scale, as in Fig.1.

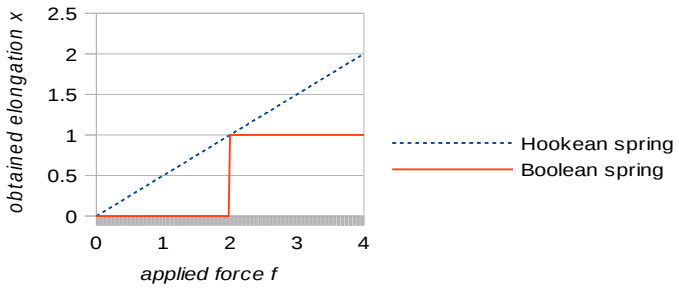

Figure 1 – The transduction function of a Hookean spring and a Boolean spring, both with $k = 2$ (see eq. (1) and (2)).

Calibrating a Boolean spring ideally requires applying a continuously increasing force whose values are known and registering the value f' of the force that makes the spring elongate, so that $k = f'$. Having applied a measurand f to a calibrated Boolean spring, let us suppose that the indication value $x = 1$ has been obtained. The only conclusion that can be drawn in this case is that $f / k \geq 1$, and therefore that $f \geq k$. Hence, despite the underlying algebraically rich scale of the quantity subject to measurement, the Boolean spring operates as a pass-fail classifier, and the measurand is actually dealt with as an *ordinal quantity*. While the scale type of the measurand determines the richest possible scale type of the measurement result, the characteristics of the measurement may lead – as in this case – to a weaker scale type (the distinction between property types and property evaluation types is extensively presented and discussed in [9]).
A critical issue is about the instrument *resolution*, defined by the VIM as "smallest change in a quantity

being measured that causes a perceptible change in the corresponding indication". The resolution of a Boolean spring is the inverse function of the distance $|f–k|$: for forces far enough from the threshold k, the resolution is practically null. With the aim of increasing the instrument resolution, and then refining the measurement result, let us suppose that an array of N calibrated Boolean springs is available, each of them with a different constant k_i, and sequenced so that $k_i < k_{i+1}$ (sequencing is an immediate by-product of calibration). The *measurement procedure* specifies now that the measurand f has to be applied to the Boolean springs in sequence until the j-th spring is determined such that:

- f elongates all springs i, $i<j$, i.e., the indication value $x_i = 1$ is obtained, so that $f \geq k_i$;
- f does not elongate the j-th spring, i.e., the indication value $x_j = 0$ is obtained, so that $f < k_j$

(hence if $j = 1$, i.e., no springs are elongated, $f < k_1$, and if $j = N+1$, i.e., all springs are elongated, $f \geq k_N$). [2]

In the simplest case of a sequence of N=2 Boolean springs, with constants k_1 and k_2, $k_1 < k_2$, three cases can then arise:

(α) $x_1 = 0$, i.e., the applied force does not elongate any Boolean spring: $f < k_1$;

(β) $x_1 = 1$ and $x_2 = 0$, i.e., the applied force elongates the first Boolean spring but not the second one: $k_1 \leq f < k_2$;

(γ) $x_2 = 1$, i.e., the applied force elongates both the Boolean springs: $f \geq k_2$.

Let us call such a scale "B[k_1k_2]", i.e., a Boolean scale of order [k_1k_2], so that the three cases might lead to attribute quantity values such as 0 B[k_1k_2], 1 B[k_1k_2], and 2 B[k_1k_2] respectively, to be read "0 in scale B[k_1k_2]" and so on. Indication values could be then reported by counting the number of elongated Boolean springs to give a "raw score" x', i.e., the indication for the array of Boolean springs:

(α) if $x_1 = 0$ (and therefore $x_2 = 0$) then $x' = 0$;

(β) if $x_1 = 1$ and $x_2 = 0$ then $x' = 1$;

(γ) if $x_2 = 1$ (and therefore $x_1 = 1$) then $x' = 2$

(in the above, the maximum raw score is taken as implicit, but it might be more complete to say that $x' = 0$ should be interpreted as "0 out of 2" or "0/2" and so on).

This is an interesting situation in which the available underlying model on the measurand, i.e., classical mechanics, interprets it as a ratio-scale quantity and the instrument indication can be thought of as an absolute-scale quantity (as any cardinality, number of elongated Boolean springs in this case), and at the same time the measurand is actually measured in a weaker scale (we are referring to the classification of scale types proposed in [4]. Some of the strong, sometimes vehement, objections that it received are founded on the peculiar lexicon in the presentation, in particular relating to the choice of calling "permissible" or "admissible" the scale-invariant transformations / statistics ("Stevens' attempt to legislate acceptable uses of statistical methods is better forgotten" [10]). What properly remains is indeed the (correct) algebraic concept of *scale invariance*). Indeed, even if the ratios k_j / k_i were known by calibration, so that for example $k_2 / k_1 = 2$, from $f_a = 1$ B[k_1k_2] and $f_b = 2$ B[k_1k_2] the correct conclusion would be $f_b > f_a$, whereas $f_b = 2f_a$ would be generally unjustified: the measurement result $f_a = 0$ B[k_1k_2], obtained from the indication value $x_1 = 0$, is to be interpreted as $f_a < k_1$, and so on. As in the case of a single Boolean spring, the information conveyed by a measurement in the B[k_1k_2] scale is purely ordinal. Note, however, that the underlying ratio-scale quantity could be approximated to an arbitrary degree by inserting more Boolean springs into the measurement system, so to obtain intervals small enough to be suitably represented, e.g., by their middle points, and therefore by rational numbers.

3. Example 2: A Guttman scale for attitude

Now consider the issue of how to measure an attitude f of persons a, b, ..., f_a, f_b, ... for short, where a test is designed for this purpose, including a set of N items with dichotomous responses, i.e., each item producing an indication x, either $x = 1$ or $x = 0$ (positive and negative response respectively). Indications are hypothesized depending on both the attitude of the person and the difficulty k of the test item.

As shown in Fig.2, the black box modeling highlights the (high level) isomorphism of this example with the previous one.

2 While the analogies with the behavior of a quantizer are manifest, an array of Boolean spring is a transducer, *not* a quantizer. Indeed, no hypotheses are assumed on the structure of the measurand, which might even be known as an already discrete quantity.

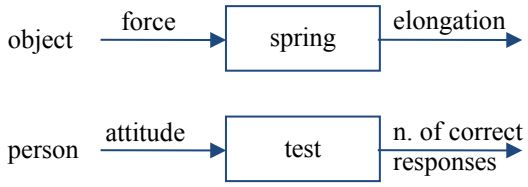

Figure 2 – Black box models for examples 1 and 2.

The dependence of attitudes (of persons) and difficulties (of test items) is sometimes presented in terms of their direct comparison, a peculiar position given that in principle they are properties of different *kind*. On the other hand, item difficulty can be interpreted as the minimum attitude required to respond positively to the item itself, and we will consider it as such here.

Let us suppose that a basic model for f and the test is accepted such that:

(1) a non-empty set of persons a, b, ... is given, such that each of them has, to a higher or lower degree, the attitude f; the attitude f is assumed to be at least ordinal, i.e., some independent knowledge (as obtained by item calibration and validity studies) is available justifying the invariance of the relation $f_a < f_b$, meaning that the person b has a greater / better attitude f than the person a; the dependence on time is implicit here: at a different time it might well be that $f_a > f_b$;

(2) a test including one or more items is given, such that each item has an f-related difficulty k; being in its turn of the kind of an attitude, the difficulty k is also assumed to be at least ordinal, i.e., some independent knowledge is available justifying the invariance of the relation $k_i < k_j$, meaning that the item j is more difficult than the item i with respect to the attitude f, i.e., the item j requires a greater / better attitude f to be responded positively than the item i;

(3) for a person a and an item i, f_a and k_i can be compared, and the item transduction function:

$$\begin{cases} \text{if } f_a < k_i, \text{ then the response is (expected to be) negative, } x_i = 0 \\ \text{if } f_a \geq k_i, \text{ then the response is (expected to be) positive, } x_i = 1 \end{cases} \qquad (3)$$

is assumed (note that the indication should be denoted $x_{a,i}$: the subscript a will be omitted for simplicity).

The test, whose items have been calibrated and then sequenced according to their difficulty (also in this case calibration consists in establishing a value for the constant k), i.e., $k_i < k_{i+1}$, can be then exploited as a measuring instrument for f, on the basis of the measurement procedure specifying that the person a whose attitude f_a is the measurand has to respond to the test items in sequence until the j-th item is determined such that:

- a responds positively to all items i, $i < j$, i.e., the indication value $x_i = 1$ is obtained, so that $f_a \geq k_i$;
- a does not respond positively to the j-th item, i.e., the indication value $x_j = 0$ is obtained, so that $f_a < k_j$

(hence, as above, if $j = 1$, i.e., a responds negatively to all items, $f_a < k_1$, and if $j = N+1$, i.e., a responds positively to all items, $f_a \geq k_N$).

Exactly as in the case of an array of Boolean springs, in the simplest case of a test of two items, with constants k_1 and k_2, $k_1 < k_2$, three cases can arise:

(α) $x_1 = 0$, i.e., a does not respond positively to any item: $f_a < k_1$;

(β) $x_1 = 1$ and $x_2 = 0$, i.e., a responds positively to the first item but not to the second one: $k_1 \leq f_a < k_2$;

(γ) $x_2 = 1$, i.e., a responds positively to both items: $f_a \geq k_2$.

Let us call such a scale "G[$k_1 k_2$]", i.e., a Guttman scale of order [$k_1 k_2$]. In a Guttman scale items are indeed arranged in an order so that an individual who agrees with a particular item also agrees with items of lower rank-order [11]. As considered before for the arrays of Boolean springs, the information conveyed by a measurement in the G[$k_1 k_2$] scale is purely ordinal.

The analogies of this example and the previous one about arrays of Boolean springs are manifest, as Table 1 shows.[3]

3 A delicate point is about the role of the person a in the test. The measurand is here a property of a, which is then the

Table 1 – Structural analogies in the measurement of force by means of arrays of Boolean springs and of attitude by means of tests of ordered dichotomous items.

	measurement of force	*measurement of attitude*
object under measurement:	an object *a* that can exert a force	a person *a* that can exhibit an attitude
measurand:	force f_a applied by *a*	attitude f_a of *a*
measuring instrument:	array of Boolean springs	test as a sequence of ordered items with dichotomous responses
measurement principle:	Boolean springs elongate if the applied force is sufficiently great	test items are responded positively if the attitude of the person is sufficiently great
instrument parameters to be calibrated:	elastic constants k_i of Boolean springs	difficulty constants k_i of test items
calibration outcome:	attribution of value to the elastic constants of Boolean springs	attribution of value to the difficulty constants of test items
measurement procedure:	application of a force to the Boolean springs in sequence	responses of test items in sequence
indication ("raw score"):	sequence/number of Boolean spring elongations	sequence/number of item responses
measured quantity value:	value in a scale such as B[k_1k_2]	value in a scale such as G[k_1k_2]

This makes the discourse about arrays of Boolean springs and tests of ordered dichotomous items interchangeable, and both understandable in the same concept system and with the same lexicon. Rasch models builds upon these bases.

4. Extending the examples: Rasch models

Two crucial assumptions need to be added to this measurement setup to make it a Rasch model:
1. the result of the transduction is given in probabilistic, instead of deterministic, terms and the probability of positive response increases as the person attitude increases and decreases as the test item difficulty increases;
2. attitudes (and difficulties) can be meaningfully compared by their ratio.

Let us argue about these assumptions and develop their consequences.

Transduction, as performed by a Boolean spring or a test item, has been modeled above by means of a deterministic function: if the measurand exceeds the instrument parameter a 'pass' indication is obtained, and a 'fail' indication otherwise. A deterministic transduction function, and thus a deterministic measurement model, might be considered not appropriate in some cases, and in particular for the measurement of attitudes. Hence, the instrument output could be modeled as a probability distribution, instead of a singleton, thus interpreting:

- the presence of an underlying unobserved variable, i.e., an influence quantity, whose variability determines the probability distribution on indications;
- the non-deterministic dependence of the indication on the measurand;[4]
- the hypothesis that the measurand is itself stochastic (this, which recalls some interpretations from quantum physics, is discussed in the context of psychometrics by [13]).

Accordingly, the probability of positive response given the attitude f and the item difficulty k, $P(x=1|f,k)$, denoted P_1 for short, is introduced so that $P_0=1-P_1$ is the probability of negative response, $P_0=P(x=0|f,k)$ in

object under measurement (the VIM does not define the concept, and sometimes uses the expression "phenomenon, body, or substance"), and *not* (part of) the measuring instrument. In other contexts, where the aim the acquisition of information on a different object, the person could instead be (part of) the measuring instrument, as in the case questionnaires are used to collect opinions on the perceived quality of products or services.

4 Non-deterministic transduction models are unusual in (classical) physical measurement, where transduction is based on deterministic physical effects. On the other hand, probabilistic models of measurement have also been proposed in metrology (e.g., by [12]), and however are clearly a generalization of the deterministic ones. In this sense in psychometrics Rasch models generalize Guttman ones.

the same conditions. The reference here is to the attitude f_a of a given person a and to the difficulty k_i of a given item i, so that the correct notation is, e.g., $P(x_{a,i}=1|f_a,k_i)$. With this understanding the subscripts will be omitted for simplicity.
Under the supposition that the item difficulty k is given, as resulting from calibration, and the indication x is observed, $P(x=1|f,k)$ is in fact the likelihood of the attitude f. Qualitatively, an increase of the difficulty k for a given attitude f is supposed to decrease the probability of positive response.
The operational definition of item difficulty k as minimum attitude required to respond positively to the item, given in the deterministic case, has to be revised here, and the following definition is adopted: k is the attitude of a person whose probability to respond positively is 0.5, i.e.:

$$\text{if } f = k, \text{ then } P_1 = 0.5 \tag{4}$$

(it should be noted that the measurement procedure introduced so far does not permit one to assess equalities, so that this might be intended as a limit condition). Moreover:

$$\begin{array}{l} \text{if } f < k, \text{ then } P_1 < 0.5 \\ \text{if } f \geq k, \text{ then } P_1 \geq 0.5 \end{array} \tag{5}$$

a probabilistic generalization of eq. (2).
The second assumption in Rasch models is that attitudes, and then item difficulties, can be compared by their ratio, so that the empirical meaningfulness of attributing values in ratio scales to attitudes is assumed. The previous condition can be then rewritten as:

$$\begin{array}{l} \text{if } \dfrac{f}{k} < 1, \text{ then } P_1 < 0.5 \\ \text{if } \dfrac{f}{k} \geq 1, \text{ then } P_1 \geq 0.5 \end{array} \tag{6}$$

which, on the other hand, still does not give a specific value to P_1 as a function of the two variables f and k, and thus their ratio f/k. By noting that P_1 increases if f increases given k, and P_0 increases if k increases given f, in Rasch models the ratio f/k is equated to the odds $P_1/P_0 = P_1/(1-P_1)$:

$$\frac{f}{k} = \frac{P_1}{1 - P_1} \tag{7}$$

Hence:

$$P_1 = (1 - P_1)\frac{f}{k} = \frac{f}{k} - P_1\frac{f}{k} \text{ then } P_1\left(1 + \frac{f}{k}\right) = \frac{f}{k}$$

and finally:

$$P_1 = \frac{f/k}{1 + f/k} \tag{8}$$

which may interpreted by stating that P_1 is proportional to f/k, with $1+f/k$ representing the proportionality factor. The functional dependence of P_1 on f and k is graphically represented in Fig.3.

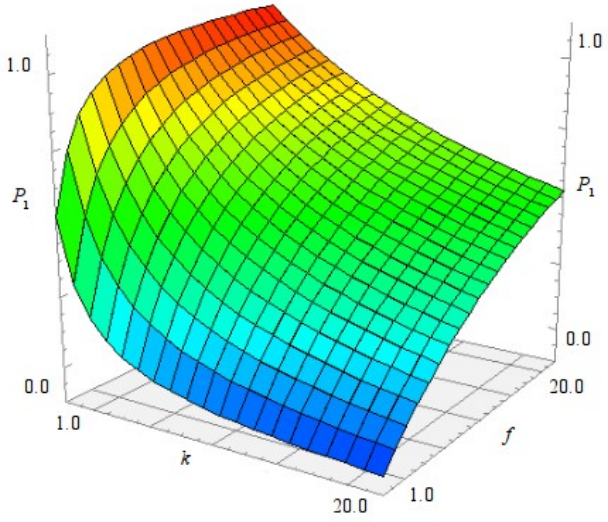

Figure 3 – The probability P_1 of positive response as a function of a person attitude f and item difficulty k (see eq. (8)).

The canonical form of the (dichotomous) Rasch model, here formalizing the transduction function of the test item of difficulty k when submitted to a person of attitude f, is obtained by substituting the quantities f and k with their logarithmic counterparts, $\varphi = \ln(f)$ and $\kappa = \ln(k)$, so that:[5]

$$P_1 = \frac{\exp(\varphi)/\exp(\kappa)}{1+\exp(\varphi)/\exp(\kappa)} = \frac{\exp(\varphi - \kappa)}{1+\exp(\varphi - \kappa)} \tag{9}$$

Note that, although this is the standard form that is given to represent the Rasch model (although it sometimes appears in a logit or antilog expression instead) in the social sciences literature (see e.g., [8] or [13]), it is somewhat unsatisfactory to metrologists, as it is not an explicit algebraic formula for the measurand in terms of the indication. When eq. (9) is solved (i.e., estimated) in a statistical algorithm, and the estimation is successful, there does indeed result a one-to-one mapping from the indicator (i.e., the "raw score") to the measurand φ, but this is usually simply an empirical table, and not expressed in a more elegant fashion.

The functional dependence of P_1 on φ, for $\kappa = 0$, is graphically represented in Fig.4.

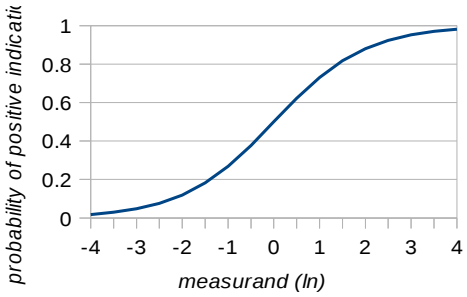

Figure 4 – The transduction function of a Rasch model as a function of $\ln(f) = \varphi$, with $\ln(k) = \kappa = 0$.
Note that the condition in eq. (4) is satisfied: $P_1 = 0.5$ when $\varphi = \kappa$, i.e., $\varphi = 0$.

The chart may be more generally interpreted by representing in the x-axis the values $\varphi - \kappa$, i.e., $\ln(f/k)$, the unit in this case being called the *logit* (i.e., a contraction of "log-odds" unit). By definition, the logit of a

5 What is gained by the application of logarithms in transforming ratios into differences is lost in scale type. As mentioned, f and k are assumed to be evaluated in a *ratio* scale, and therefore f/k is invariant under multiplication, i.e., $f/k = (cf)/(ck)$ for $c \neq 0$. In general $\ln(f)/\ln(k) \neq \ln(cf)/\ln(ck)$, i.e., φ and κ are not invariant in ratio scale. On the other hand, $[\ln(f_1) - \ln(f_2)]/[\ln(k_1) - \ln(k_2)] = [\ln(cf_1) - \ln(cf_2)]/[\ln(ck_1) - \ln(ck_2)]$ for $c \neq 0$. Hence φ and κ are invariant in *interval* scale.

probability P is $\ln(P/(1-P))$. In this case, given the assumption that $P_1/P_0=f/k$, the value $\ln(f/k)=\ln(f)-\ln(k)=\varphi-\kappa$ is "in logits". The conceptually immaterial differences of a multiplicative factor and the base of the logarithm do not hide the analogy with the way ratios of physical quantities are represented in decibels.
Hence the chart represents the probability that the measuring instrument (the test item / the Boolean spring) produces a positive indication value ($x=1$) as function of the measurand value (the person attitude / the applied force) compared to the instrument parameter (the item difficulty / the spring stiffness) with units in logits (note that 1 logit corresponds to $\varphi-\kappa=1$, i.e., $\ln(f/k)=1$, and therefore $f/k=\exp(1)\approx 2.71$). Once a reference ("zero") attitude / difficulty has been fixed, person attitudes and item difficulties can be evaluated and placed on the x-axis ("a person whose attitude is i logits"; "an item whose difficulty is j logits").
Despite its simplicity, this model has several interesting consequences. For example, let P_a be the probability P_1 for the person a, $P_a=\exp(\varphi_a-\kappa)/G_a$ where $G_a=1+\exp(\varphi_a-\kappa)$, and let P_{ab} be the probability that a succeeds and b fails, i.e. (under the assumption of statistical independence), $P_{ab}=P_a(1-P_b)$. Easy algebraic transformations lead to $P_{ab}=\exp(\varphi_a-\kappa)/G_aG_b$, so that $P_{ab}/P_{ba}=\exp(\varphi_a-\varphi_b)$ and finally $\varphi_a-\varphi_b=\log(P_{ab})-\log(P_{ba})$, where P_{ab} might be estimated by the relative frequency of the items in which a succeeded and b failed, and vice versa for P_{ba}. Hence, according to this model the (logarithmic, i.e., in logits) distance in attitude between two persons is estimated by the logarithmic distance of such relative frequencies. From an estimation point of view, this expression is also very useful, as it implies the *separability* of estimation of the person parameters from the item parameters (and vice versa).

The information conveyed by a single Boolean transducer (a single item test / a single Boolean spring array) is meager. To overcome this issue, a more complex measuring instrument can be adopted as discussed above, made of multiple transducers each of them characterized by a parameter k_i, as shown in Fig.5.

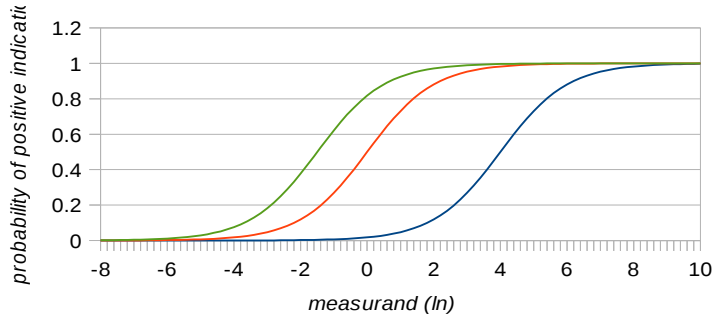

Figure 5 – The transduction functions of three items / Boolean springs with $\kappa=-1.5$, $\kappa=0$, and $\kappa=4$.

Under the hypothesis that the transducer behaviors are independent of each other, i.e., the indication x_i obtained from the i-th transducer (depends on the measurand but) does not depend on the indication x_j obtained from any other transducer, $j\neq i$, the probabilities P_{1i} can be meaningfully added, and the sum can be interpreted as the total score, i.e., the indication, produced by the test. Note that, in the case of attitude measurement this implies that having answered item i does not change the probability distribution of correctly answering the j-th item.

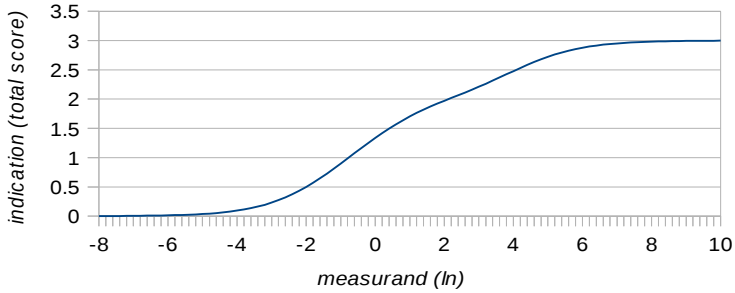

Figure 6 – The transduction function of the measuring instrument as in Fig.5.

It is supposed that the generated transduction function is invertible at least in the points $x=0$, $x=1$, $x=2$, and $x=3$, corresponding to the indication values that can be obtained as total scores of the test. The inverse of the transduction function in these points is the instrument calibration function, and its graphical representation in the diagram (indication × measurand) is the *calibration curve*, to which an indication value is applied to get a value for the measurand, i.e., a measured quantity value, in logits or the (dimensionless) unit assumed for attitudes.

5. The problem of validation

The analogy between arrays of Boolean springs and tests of ordered dichotomous items can be now interpreted in terms of measurement, which in both cases (in all cases, actually) has the structure of an inferential process. Let us rewrite what has been considered above for the arrays of Boolean springs:

1. if the measurement model is correct, and
2. the instrument is correctly calibrated, and
3. an indication value has been experimentally obtained, and
4. the measurement function is applied to the indication value
5. then a value for the measurand is obtained.

We have shown that exactly the same can be said of tests, intended as measuring instruments of attitudes.

It has been noted above that measurement assumes certain premises required for the inference above, particularly the scale types of the involved quantities and the structure of their functional relation: thus, these are expected to be validated independently of the measurement. This theory-ladenness of measurement has been emphasized at least since [14], and has become explicit in the very concept of measurement model as defined in the VIM. The idea that traditionally measurement was intended as a process able to convey "pure data" has been perhaps retrospectively overemphasized. Even in a book such as [15], which gave rise to the operationalism, one can read sentences as "it is of course the merest truism that all our experimental knowledge and our understanding of nature is impossible and non-existent apart from our own mental processes". In the case of springs used as transducers, dynamometers, and many physical measuring instruments, the underlying theory is well established, it being the theories of physics themselves, so that this principled dependence is not operatively problematic for the establishment of a measurement context. Indeed, both the measurand (force) and the indication (length) are embedded in a rich network of relations with other physical quantities, which can be exploited to positively answer the question "is the quantity measured by means of this spring actually a force?".

The situation for the measurement of attitude, assumed here as exemplary of many non-physical properties, is fundamentally different, since such a network of well-established theories is not available in this case[6] and the only possible way out might seem to be the operationist strategy: "this test apparently measures something; let us call it an attitude.". On the other hand, the claim of measurability is minimally grounded here on the acknowledgment that an underlying, informal model of the property is an established common understanding within the intellectual community from which the test developers derive their standing, and the development of measuring instruments of it (such as through the Construct Modeling approach described below) can be seen as a positive step towards a more formal model of the property. Without a previous, at least informal knowledge on the attitude under consideration, the measurement of the unknown property would be a blind process of tentative discovery. The concept of 'underlying model' of a measurand is explored in [17]. Hence the crucial step of the inferential process is here the first one: *how to guarantee that the measurement model is correct?* And this brings up a preliminary issue: *what is required for a model to be a correct measurement model?* The very concept 'measurement model' has been so recently formulated as such that a clear-cut answer to these questions is still to be found. As already mentioned, a minimal requirement seems to be the hypothesis that the indication causally (but perhaps probabilistically) depends on the measurand, and therefore the indication conveys information on the measurand itself.

In the social sciences, this problem – that is almost never spelled out in the measurement of physical quantities – is addressed through the concept of *internal construct validity*. Here a *bootstrap strategy* is typically used: by means of the analysis of measurement results (i.e., their correspondence with what is

6 According to Ludwik Finkelstein [16], "Weakly defined measurement has one or more of the following features: (i) it is based on an ill-defined concept of the quality, (ii) there is significant uncertainty in the empirical relational system that it represents, (iii) the symbolic relational system has limited relations defined on it, (iv) there is no adequate theory relating the measurement to other measurements in the same domain.".

expected based on the underlying model), the measurement model itself is validated and, if required, progressively refined. This applies in particular to the assumption that attitudes are measurable in an interval scale: under this hypothesis, Rasch models give a procedure to get measurement results from the raw data provided by the answers given to the tests. The correspondence of such results to what is accepted about attitude in terms of the informal model mentioned above is a confirmation that the hypothesis was correct: the instrument does measure an attitude, and useful information on the measurand has been obtained.

In psychometrics this bootstrap strategy has several characteristic versions. The most widespread one is associated in practice with the so-called Classical Test Theory (CTT): it is the traditional "blueprint" approach [18], [19] which includes the following activities:

(a) the definition of the measurand;

(b) the instrument setup, as a specification table for the items that will be used to gather data about the measurand – this is traditionally referred to as the "blueprint" – typically involving, say in the case of an achievement test, a matrix of skills by content;

(c) the instrument calibration, as a set of rules for scoring the responses to the items, which may range from very simple rules for a multiple choice item (e.g., the correct option is scored 1 and the rest are scored 0) to very complex rules requiring human judgments, and a training for the raters and a reconciliation program for inconsistent ratings;

(d) the validation stage as such, consisting in checking the empirical relationships between the results obtained by means of the instrument with the values of other, independently acquired variables that are hypothesized to have relations to the measurand.

In addition, several other validation strategies have been proposed – these have been changing in their nature and emphasis over the past 100 years or so of the history of measurement in the social sciences – such as those based on content analysis of the items, evidence concerning the response (i.e., transduction) process, and evidence concerning the consequences of utilizing the measurements results [20].

An alternative approach, termed *Construct Modeling*, starts off by assuming that the model of the measurand can be represented by discrete values, or "levels", in a linearly ordered sequence, each value possibly intended as corresponding to a segment of a continuum, from a lowest to a highest level. In this context the measurand model is typically called its *substantive theory*, i.e., a theoretical background to the measurand itself that provides one with some sort of a structure to compare with empirical evidence, particularly in the validation process. In social sciences such substantive theories may be in a form other than equations, such as a verbally-defined relationship, and even a diagrammatic one. Consider, moreover, that the term "construct" is used in psychometrics to refer to the property object of the model, i.e., the measurand. When these levels are laid out in a diagram, as in Fig.7, it is termed a "construct map".

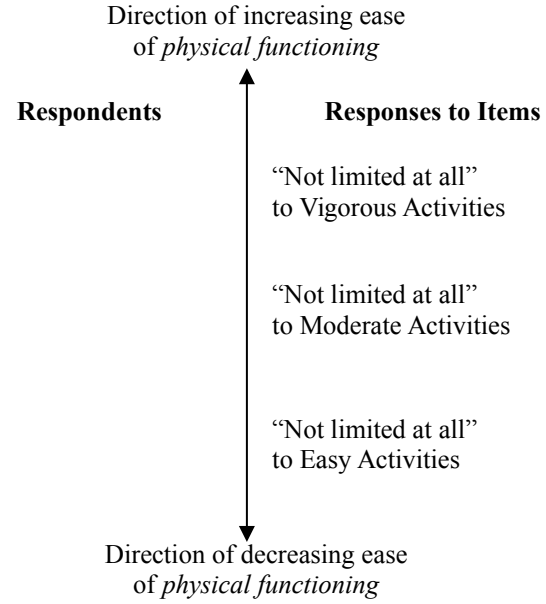

Figure 7 – A sketch of the construct map for the Physical Functioning subscale (PF-10) of the SF-36 Health Survey.

The main difference with the CTT is that, in activity (a), the blueprint is specified as a construct map, consisting of successive levels, and, in activity (d), the validation stage consists principally of the checking the consistency of the empirically estimated levels (as recorded, for instance, as in the "Wright map" described below) with the hypothesized ordering shown in the construct map.

In order to make the argument here more accessible, we will describe a concrete example that was developed in the context of a person's judgment of his typical physical performance: the Physical Functioning subscale (PF-10) [21] of the SF-36 health survey [22]. The SF-36 instrument is used to assess generic health status, and the PF-10 subscale assesses the physical functioning aspect of that. The items of the PF-10 consist of descriptions of various types of physical activities to which the respondent may respond that they are typically either "Not limited at all", coded as 1, or "Somehow limited", coded as 0.[7] The actual items in this instrument are given in Table 2. An initial construct map for the PF-10 is shown in Fig. 7, where the sequence of increasing ease of physical functioning can be noted, as indicated by the order of the item responses. This sequence ranges from very much more strenuous activities, such as those represented by the label "Vigorous Activities", down to activities that take little physical effort for most people. The full item set for the PF-10 is shown in Table 2, along with the abbreviations that are used in the Figures and the text for each item. In developing the items, the levels of the construct map can be linked to the items as follows: Easy: 9, 10; Moderate: 2 to 8; Vigorous: 1.

Note that the definition of the levels in this example is given solely in terms of the ordering of the items. However, in general, it is also possible to display the order using levels of the persons responding to the items (the "Respondents"), and that is why in Fig.7 there is a column on the left hand side also. In this example, the instrument developers did not use that possibility, and hence the left column is empty.

Table 2 – Items in the Physical Functioning subscale (PF-10).

Item n.	Item label	Item description
1	VigAct	Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports
2	ModAct	Moderate activities, such as moving a table, pushing a vacuum cleaner, or playing golf
3	Lift	Lifting or carrying groceries
4	SevStair	Climbing several flights of stairs
5	OneStair	Climbing one flight of stairs
6	Bend	Bending kneeling, or stooping
7	WalkMile	Walking more than a mile
8	WalkBlks	Walking several blocks
9	WalkOne	Walking one block
10	Bath	Bathing or dressing yourself

A large data set of patients' responses to these items has been collected [23], and a Wright map has been calibrated using the Rasch model to estimate the item difficulties from those data, dichotomized as described. The results are displayed in Fig.8 using a *Wright map*, with the respondents' locations and items' locations presented on either side of a vertical line.

There are a number of features in the Wright map that are worth pointing out. The dashed vertical line in the center represents the measurand in logits, relating the measurand itself to the probability of response, i.e., indication. The raw score units are also presented to the right of the logits (e.g., 0 logit corresponds to having a moderate level of physical functioning, similar to a raw PF-10 score of 5). On the left hand side of the central line, under "Respondents," the locations of the respondents on the logits scale are indicated by "X"s. These form a histogram showing the shape of the respondent distribution, a fairly flat, which is surely not a Gaussian one, indicating that respondents had a wide range success in overcoming barriers toward physical activity (given that these are hospital data, we expect to see a somewhat skewed distribution with more respondents having lower levels of physical functioning, i.e., more respondents with negative logits). On the right hand side of the central line, under "Item responses", the locations of the items are shown. Each item location indicates the amount of physical functioning a generic person must have if there is a 0.5 probability of that person giving a positive response to these items. Comparing this Wright map to the construct map in

7 As developed, there were actually three categories of response to the PF-10, "Limited a lot", "Limited a little", and "Not limited at all". Just two categories will be considered here, where the first two categories are collapsed together, thus making the data dichotomous.

Fig.7, we can notice several ways in which they differ. First, this map is not just a sketch of the idea of the construct but an empirical map, based on respondents' self-reports. A histogram of the responses is shown on the left hand side of the map. What is unusual for a histogram is that the spaces between the bars of the histogram are not evenly spaced. That is because the locations of the bars are the estimated locations of the respondents, which can vary in a continuous way. Although each bar corresponds to a particular score, the estimation procedure can result in person estimates located at any point on the continuum – they are not located at integer values as are raw scores. The units for that continuous scale are shown on the far left hand side, in the column headed "Logit". The respondents range from those that are "less limited" at the top, to those that are "more limited" at the bottom. Each location, i.e., each histogram bar, corresponds to a score on the instrument, ranging from 0 to 9 (no one scored 10, so it is not shown).

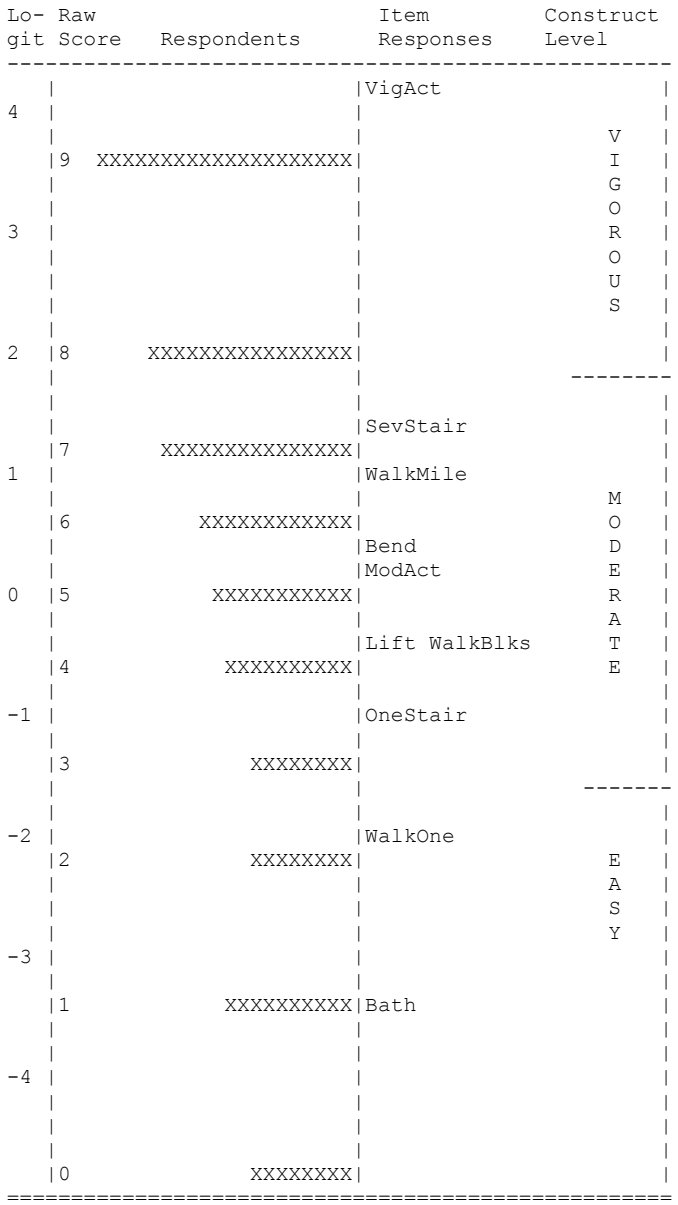

Figure 8 – A Wright map for the dichotomized PF-10 instrument (each X is approx. 18 cases; each row is 0.20 logits).

The right hand side of the map in Fig.8 shows the estimated item locations, corresponding to the values $\kappa = \ln(k)$ in Eq. (9). Notice, for instance, that the respondents with a score of 6 are at almost the same point on the map as the "Bend" item. This means that they have approximately a 0.5 probability of responding "Not limited at all" to that item. Noting that "SevStair" is about 1 logit above this location, we can see that

for the respondents with a score of 6 the probability of getting the same response to that item is approximately 0.27. And since "WalkBlks" is about 1 logit below this location, we can say that the probability of respondents with a score of 6 giving the more positive response to that item is approximately 0.73 (these probabilities can easily be worked out using a calculator to implement eq. (9)).

In order to examine the evidence for construct validity, and specifically internal structure validity [20], we examine how well the results reflected in the Wright map correspond with the theory of the variable embodied in the construct map. Comparing the right hand side of Fig. 8 to the right hand side of the construct map in Fig. 7, we can see that "Vigorous activities" are indeed the most difficult, and that easy activities like "Bathing" and "Walk one block" are located at the easy end of the Wright map. The items hypothesized to lie between these two extremes do indeed so. Thus, the results of the Rasch analysis, as expressed in the Wright map, show that the empirical data support the hypothesized grouping of the items into the levels, and also the relative order of the levels.

Note that the pattern expected in Fig.8 is a quite simple one, and hence the example may seem almost trite: in fact, in order to make the concept as straightforward as possible, we have indeed chosen a simple example, and we are only illustrating the final round of data collection and analysis. Most often the results do not come out anything like as cleanly as they have in Fig.8: more typically, the results from the initial data collection for a new set of items will not reveal anything like the hypothesized order, and the instrument developers then must return to their assumptions and examine each one, from the specifics of individual items, to the design plan for items, to the specification of levels, and even, sometimes, all the way back to the definition of the measurand. And this whole iteration may be required several times. This sort of analysis is given a more thorough examination in [24] and [25].

As was noted, the checking of internal structure validity consists minimally in the comparison of the empirical evidence contained in the Wright map with the hypotheses embedded in the construct map. Hence, in addition, one should also check on the technical evidence to support the specific statistical model that has been used. In this case, that amounts principally to the checking of the fit of the data to the Rasch model. For the data set used here, a reasonable fit was found [24], although there was some evidence that the VigAct item was not working quite like the others (this perhaps relates to the fact that this item is far less specific in its context than most other items).

6. Discussion and conclusions

The Rasch model relates the indications obtained by the experimental application of a test to a property intended to be measured, and formalizes this relation through equations (7), (8), and (9), that could admit of an empirical validation. This would typically be accomplished by assuming that the probability of positive response P_1 is approximated by the relative frequency of positive responses in a repeatable experiment. Since metrologists are accustomed to consider that their measuring instruments implement transduction effects known by some physical law, we suspect that a typical approach to Rasch models by metrologists would be: if these models work, i.e., if they are validated, then we have discovered an invariant, that can be exploited for future measurements. Indeed, in the metrology terminology, (repeatability and) validation leads to (the hypothesis of) reproducibility (of course, reproducibility implies repeatability, not vice versa).

Now, the usual practice in measurement in the contemporary social sciences is not so much direct observations of the form of the equation for P_1, but rather the checking of model fit, as mentioned above. There are many different methods of testing fit that have been developed in the psychometric and also the broader statistical literature – we will not try and summarize even a few of them here. But the core element is that the aim is to find out how "large" are the residuals, the difference between the responses that are expected given the estimated parameters, and the actual observed values. Hence, the residuals will be aggregated in various ways that either (a) focus on specific aspects of fit, or (b) attempt to take a global perspective. This step looks like merely a technical one, but it is indeed an equivalent of the "if they work" step above. To see this occurring, we have to go back a bit in psychometrics history, and look at, for example, results reported by [26], shown in Fig.9, where the empirical cumulative probability distribution functions (cdfs) are shown for student success on items, using their age as a proxy for their ability, i.e., the measurand. In Fig.4, the x-axis is the measurand, and that is what one would prefer in Fig.9 also. However, Thurstone did not have the statistical machinery behind Fig.4 available to him in 1925, so he used a variable that he was very confident was a strong covariate, i.e., student age. These curves do indeed have a similar shape to the Rasch curve shown in Fig.4 (in fact, they even show, predominantly, the prototypical Rasch

feature that they do not intersect). Of course, similar curves can be obtained by other equations, in fact nearly any standard cdf would do, the Gaussian cdf being also commonly used, although there has been strong interest in the logistic function due to its relative simplicity, and also its special features such as the separability of estimation of φ and κ [27]. A convergent trend has also been from the psychophysics literature (see [28] for a summary account).

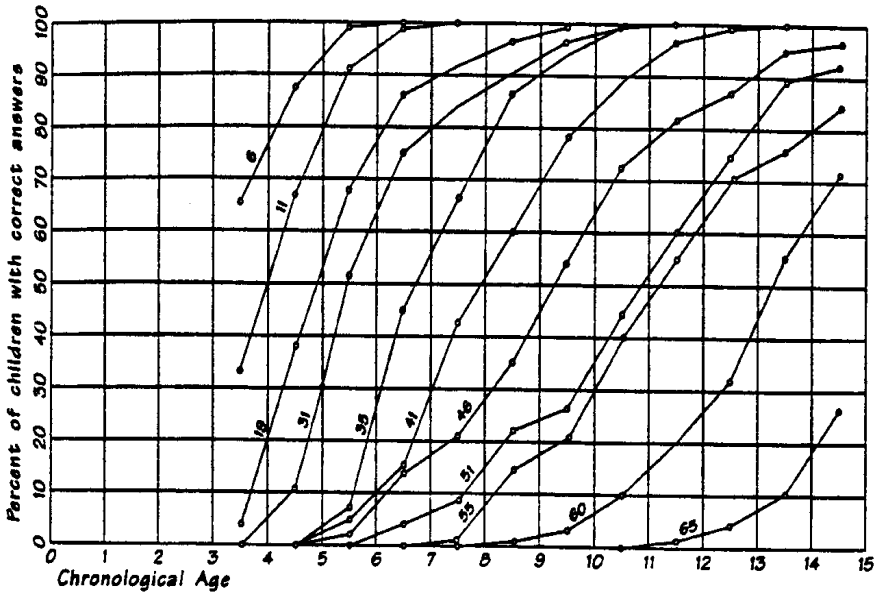

Figure 9 – Empirical cumulative probability distribution functions for student success on items, using their age as a proxy for their ability.

The search for invariants is one of the experimental bases upon which physics has been built. In three steps:
0: no repeatability: everything is singular; nothing is invariant;
1. repeatability conditions are discovered: local, specific invariants;
2. reproducibility conditions are discovered: (more) global, (more) general invariants.
In the context of measurement in the social sciences, in one sense, the first step corresponds to the Classical Test Theory, where item difficulties (for example) are deemed to be specific to an instrument and a sample of persons. Then, in the second step, with the Rasch model, and other Item Response Theory (IRT) models, we gain the possibility of having item difficulties that are useful across local contexts, so that we can have a calculus of item difficulties that corresponds to possible selections of items into an instrument. But then, at the third step, some new issues arise (such as Differential Item Functioning, the phenomenon that, for a particular item, the transduction function is different for persons from different observed groups) where there are limits to the item-person invariance that is hypothesized in the second step, and we have possibilities such as gender, content, or rater effects on item difficulty, as well, at the ultimate point, of having enough predictors to very precisely predict item difficulty.
In metrology, physical laws convey the evidence of reproducibility, and this allows us to build measuring instruments whose behavior, provided that the specified measurement procedure is followed appropriately, is indeed highly reproducible. The analogy, pursued here in Construct Modeling, is that the sequence of levels operates in the validation argument in the place of a physical law, in the sense that it provides us with something to validate against. It is not familiar-looking, as the set of levels is an ordered set of locations on a continuum, so it is not so elegant to write down as a physical law, but it is indeed a mathematical rule, thus:

if f is such that $f < k_1$, then the person is in level 1;
if f is such that $k_1 \leq f < k_2$, then the person is in level 2;
...
if f is such that $k_{N-1} \leq f < k_N$, then the person is in level N.

Of course, there are other ways to pursue greater complexities, and indeed to deconstruct the history of social science measurement, for example by adopting more complex Rasch-type models which, e.g., admit polytomous items or take into account influence properties / quantities [29]. The simplest case of a Rasch

model that we have presented here builds upon a Guttman model and adds the two assumptions introduced at the beginning of Section 4, i.e.:
(i) the result of the transduction is probabilistic, instead of deterministic; and
(ii) the measurand is a ratio-scale quantity, instead of an ordinal-scale property.
Since these assumptions are, in principle, independent of one another, the following four cases emerge as structurally possible:
A. *deterministic indications obtained from the transduction of an ordinal-scale measurand*, the case of the Guttman model in psychometrics and of Mohs hardness, Beaufort wind strength, etc. in metrology;
B. *probabilistic indications obtained from the transduction of an ordinal-scale measurand*, as for probabilistic Guttman models [30]; all ordinal measurements whose measurement model takes into account measurement uncertainties would be classified here;
C. *deterministic indications obtained from the transduction of a ratio-scale measurand*, the case of physical quantities where measurement uncertainties are neglected;
D. *probabilistic indications obtained from the transduction of a ratio-scale measurand*, the case of located latent class models [31] and, of course, Rasch models in psychometrics; this is the canonical case of the measurement of physical quantities where measurement uncertainties are kept into account.
In this perspective Rasch models belong to the same class that metrologists consider paradigmatic of measurement. Moreover, the fact that in Rasch models the non-deterministic contribution comes, even independently of influence properties / quantities, from the transduction function – which is unusual for metrology – makes such models "templates" for a possible generalized, because probabilistic, theory of measurement.

References

[1] F.A. von Hayek, The pretence of knowledge, Nobel Prize Lecture, 11 Dec 1974 (http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1974/hayek-lecture.html).
[2] J. Michell, Measurement in psychology: A critical history of a methodological concept, Cambridge University Press, Cambridge, 1999.
[3] G.B. Rossi, Measurability, Measurement, 40 (2007) 545–562.
[4] S.S. Stevens, On the theory of scales of measurement, Science, 103, 2684 (1946) 677–680.
[5] O. Holder, The axioms of quantity and the theory of measurement, 1901, English translation in: J. of Mathematical Psychology, 40 (1996) 235–252.
[6] R.T. Cox, The algebra of probable inference, Johns Hopkins Press, Baltimore, 1961.
[7] JCGM 200:2012, International Vocabulary of Metrology – Basic and general concepts and associated terms (VIM), 3rd Edition (2008 version with minor corrections), Joint Committee for Guides in Metrology, 2012 (http://www.bipm.org/en/publications/guides/vim.html).
[8] D. Andrich, Rasch models for measurement, Sage, Newbury Park, 1988.
[9] A. Giordani, L. Mari, Property evaluation types, Measurement, 45 (2012) 437–452.
[10] O.D. Duncan, Notes on social measurement. Historical and critical, Russell Sage Foundation, New York, 1984.
[11] L. Guttman, A basis for scaling qualitative data, American Sociological Review, 9 (1944) 139–150.
[12] G.B.Rossi, A probabilistic theory of measurement, Measurement, 39 (2006) 34–50.
[13] P. Holland, On the sampling theory foundations of item response theory models, Psychometrika, 55, 4 (1990) 577–601.
[14] T.S. Kuhn, The function of measurement in modern physical science, Isis, 52, 2 (1961) 161–193.
[15] P.W. Bridgman, The logic of modern physics, Macmillan, New York, 1927.
[16] L. Finkelstein, Widely, strongly and weakly defined measurement, Measurement, 34 (2003) 39–48.
[17] L. Mari, V. Lazzarotti, R. Manzini, Measurement in soft systems: epistemological framework and a case study, Measurement, 42 (2009) 241–253.
[18] J.C. Nunally, I.H. Bernstein, Psychometric theory, 3rd ed., McGraw-Hill, New York, 1994.
[19] L.J. Cronbach, Essentials of psychological testing, 5th ed., Harper & Row, New York, 1990.
[20] American Educational Research Association, American Psychological Association, National Council for Measurement in Education (AERA, APA, NCME), Standards for Educational and Psychological Testing, American Educational Research Association, Washington, DC, 1999.
[21] A.E. Raczek, J.E. Ware, J.B. Bjorner, B. Gandek, S.M. Haley, N.K. Aaronson, G. Apolone, P. Bech, J.E.

Brazier, M. Bullinger, M. Sullivan, Comparison of Rasch and summated rating scales constructed from the SF-36 Physical Functioning items in seven countries: Results from the IQOLA Project, J. of Clinical Epidemiology, 51 (1998) 1203–1211.
[22] J.E. Ware, B. Grandek, Overview of the SF-36 Health Survey and the International Quality of Life Assesment (IQOLA) Project, J. of Clinical Epidemiology, 51 (1998) 903–912.
[23] C.A. McHorney, J.E. Ware, J.F. Rachel Lu, C.D. Sherbourne, The MOS 36-item short-form health survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups, Medical Care, 32 (1994) 40–66.
[24] M. Wilson, Constructing measures: An Item Response Modeling approach, Erlbaum, Mahwah, NJ, 2005.
[25] M. Wilson, Seeking a balance between the statistical and scientific elements in psychometrics, Psychometrika, 78, 2 (2013) 211–236.
[26] L.L. Thurstone, A method of scaling psychological and educational tests, J. of Educational Psychology, 16 (1925) 433–451.
[27] G. Rasch, Probabilistic models for some intelligence and attainment tests, Danish Institute for Educational Research, Copenhagen, 1960.
[28] R.D. Bock, A brief history of item response theory, Educational Measurement: Issues and Practice, 16, 4 (1997) 21–32.
[29] P. De Boeck, M. Wilson (eds.), Explanatory item response models: A generalized linear and nonlinear approach, Springer-Verlag, New York, 2004.
[30] M.A. Croon, Latent class analysis with ordered latent classes, British J. of Mathematical and Statistical Psychology, 43 (1990) 171–192.
[31] J.K. Vermunt, The use of restricted latent class models for defining and testing nonparametric and parametric IRT models, Applied Psychological Measurement, 25 (2001) 283–294.