OPEN

Why Questionnaire Scores Are Not Measures

A Question-Raising Article

Luigi Tesio, MD, Stefano Scarano, MD, Samah Hassan, MD, MSc, PhD,
Dinesh Kumbhare, MD, PhD, FRCPC, FAAPMR, and Antonio Caronni, MD, PhD

Abstract: Any person is provided by characteristics that can be neither located in body parts nor directly observed (so-called latent variables): these may be behaviors, attitudes, perceptions, motor and cognitive skills, knowledge, emotions, and the like. Physical and rehabilitation medicine frequently faces variables of this kind, the target of many interventions. Latent variables can only be observed through representative behaviors (e.g., walking for independence, moaning for pain, social isolation for depression, etc.). To measure them, behaviors are often listed and summated as items in cumulative questionnaires ("scales"). Questionnaires ultimately provide observations ("raw scores") with the aspect of numbers. Unfortunately, they are only a rough and often misleading approximation to true measures for various reasons. Measures should satisfy the same measurement axioms of physical sciences. In the article, the flaws hidden in questionnaires' scores are summarized, and their consequences in outcome assessment are highlighted. The report should inspire a critical attitude in the readers and foster the interest in modern item response theory, with reference to Rasch analysis.

Key Words: Questionnaires, Personmetrics,
Physical and Rehabilitation Medicine, Measurement, Rasch Analysis

(*Am J Phys Med Rehabil* 2023;102:75–82)

Measurement is a fundamental part of the patient's assessment in medicine and thus in physical and rehabilitation medicine (PRM). While some (biological) patients' characteristics, for example, weight or blood glucose concentration, are entirely and directly measured, others are not. For example, the "ability in daily activities" cannot be directly measured, and for this reason, it is described as an example of latent characteristics or traits. However, this variable can be quantified as well. This is possible by observing the patients' behavior.

Strictly speaking, observable motor behaviors, be they eye blinking, marking an answer to a questionnaire, speaking, or running, reflect a "latent" property (or trait), allowing inferences on the existence and the quantity of the generating variable.[1,2] These behaviors can be listed as "items" in questionnaires and counted. Questionnaires that provide cumulate numeric scores are dubbed "scales." The "traditional test theory" (TTT, also known as classical test theory [CTT]), which conventional psychometrics is based on (see Supplemental Digital Content 1, http://links.lww.com/PHM/B666, for a history of the name), assimilates these numerals to numbers and thus to actual quantities. Physiatrists need to understand and incorporate clinically meaningful and scientifically rigorous measurements of a patient's latent traits. The lay clinician might be satisfied with numeric outputs, heralding precision and objectivity, the "true" science hallmarks. Many complete books on traditional psychometrics (i.e., TTT) exist (for a classic manual, see the book by Nunnally and Bernstein[3]).

Measurement theory shows that TTT numerals can only approximate a true measure. For a seminal article on this topic, see the study by Luce and Tukey.[4] The item response theory,[5] from which Rasch analysis (RA) is inspired, aims to provide a formal solution to this issue. Rasch analysis takes its name from the Danish mathematician Georg Rasch (1901–1980).[6] For a benchmark article on the application of RA to physical and rehabilitation medicine, see the study by Wright and Linacre.[7] The relevance of RA[8] for professionals in the field of PRM cannot be underestimated, given that RA is more and more applied to many related areas, such as psychology, speech and occupational therapy, education, sports medicine, and pain medicine. The RA points out the problems with the TTT pseudo-measures. It points out when and how real (interval) measures can be obtained from (ordinal) scores attributed to persons' behaviors. However, physiatrists, and all other physicians will not accept the Rasch solution (apparently, unorthodox and weird) unless they acknowledge that some problems exist in questionnaires' scores.

The current article highlights the conceptual and empirical pitfalls of questionnaires' total scores. The work, aimed to raise the awareness that these scores are not measures, is not meant to be an updated review of the most recent results on this topic. Instead, references to seminal works in measurement theory

From the Department of Biomedical Sciences for Health, Università degli Studi di Milano, Milan, Italy (LT, SS); Istituto Auxologico Italiano, IRCCS, Department of Neurorehabilitation Sciences, Ospedale San Luca, Milan, Italy (LT, SS, AC); Division of Physical Medicine and Rehabilitation, Department of Medicine, University of Toronto, Toronto, Canada (SH, DK); and Pain Research Institute, Toronto Rehabilitation Institute, University Health Network, Toronto, Canada (SH, DK).

All correspondence should be addressed to: Luigi Tesio, MD, Istituto Auxologico Italiano, IRCCS, Department of Neurorehabilitation Sciences, Ospedale San Luca, via Giuseppe Mercalli 32, 20122 Milan, Italy.

Dr Khumbare is an associate editor in the *American Journal of Physical Medicine and Rehabilitation*.

The study was supported by the Italian Ministry of Health, Ricerca Corrente 2018 (RESET project).

Financial disclosure statements have been obtained, and no conflicts of interest have been reported by the authors or by any individuals in control of the content of this article.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.ajpmr.com).

ISSN: 0894-9115
DOI: 10.1097/PHM.0000000000002028

and the measurement of latent traits were preferred. Hopefully, the work will also raise the reader's interest in Rasch modeling. Two recent review articles are suggested, introducing RA to PRM professionals.[9,10]

The study, theoretical in nature, uses several examples to make plainer the concepts explained. However, no real patient or healthy participant was recruited. These examples should actually be considered "thought experiments." For these reasons, no approval from the local ethical committee was deemed necessary, and consent was waived.

MEASUREMENT PROBLEMS NEEDING SOLUTIONS

Counting objects is necessary but insufficient to measure them.[7] For example, when buying oranges to prepare juice, you can easily count their number. However, it would be much better to pay for the weight of juice you can squeeze from each orange. Thus, according to a classic metaphor in measurement theory, it seems preferable to pay for their weight (an orange is about half juice by weight[11]), because counting oranges is only a rough approximation to the juice weight.

Price and weight are abstract concepts superimposed or predicated on concrete oranges. Measurement is also an abstract concept and activity: the idea and the work of concatenating objects (i.e., adding quantities) along a conceptual and continuous gradient "from less to more." Therefore, it seems acceptable that intangible, abstract features of persons (e.g., pain, fatigue, balance, continence, mobility, depression, intelligence, pain, knowledge of math, and the like) can be measured. The difference between the physics and latent variables measures should be seen more as an empirical (i.e., practical) than an ontological (i.e., conceptual) issue.

Psychometrics or Person Metrics?

The study of a person's variables through questionnaires began perhaps in the early 20th century, with the famous Stanford-Binet IQ.[12] Scales were primarily developed for cognitive (e.g., knowledge) and psychological (e.g., mood) variables, hence, the name "psychometrics." However, motor behaviors are needed to manifest all person's variables. In addition, many variables are overtly "physical" (e.g., motor abilities, balance, continence). For this reason, the term "person metrics" has been suggested as more appropriate.[13]

Counting as an Approximation to Measurement

From Observations to a Scale

Cumulative questionnaires (or "scales") consist of a list of observable behaviors (i.e., the questionnaire's items) considered as the manifestations (effects) of a shared latent property. When the number of behaviors expressed by an individual is counted (such as when a questionnaire's total score is calculated, see hereinafter), three assumptions are implicitly accepted: (1) that all observations are caused by the same variable; (2) that all observations are caused by a single variable; and (3) that the larger the sum score, the more of the latent variable is there. Remember with this regard the oranges' example reported previously.

In the most straightforward "dichotomous" items, a 0/1 labelling (no/yes, pass/fail) is assigned to absence/presence of a given behavior (e.g., wrong/correct answer to an algebraic question in a knowledge of math scale; being dependent/independent in walking in a disability questionnaire, etc.). By "observed behavior," a "yes = 1" endorsement by the subject or, passing a given test, is intended. The symbol "1" is simply a label. "How many times" a "1" is observed, that is, how many times a subject manifests these behaviors across items (the cumulated, total scale score), provides the estimate of subjects' overall "ability." Conversely, "how many times" an item receives a 1/yes/pass label across subjects provides an estimate of item "difficulty" (the lower the item score across subjects, the "more difficult" the item).

Together with dichotomous items, "polytomous" items are also widespread (e.g., "with someone's help/with orthotic aid/with supervision/autonomous" = 0/1/2/3 or, "fully disagree/partially disagree/moderately agree/fully agree" = 0/1/2/3). If they are made of polytomous items, questionnaires are dubbed "rating scales." The concept of "counting" does not change. The total scale score comes from summing "how many times" a "1" was observed, plus "how many times" a "2" was observed, etc. The difference lies in the "weight" assigned to the counts. Observing the behavior labeled "2" counts twice as much, compared with observing the behavior scored 1, etc.

It seems advisable to build scales assigning higher scores to a better condition.[14] Moreover, a mixture of items positively (e.g., "pain is disturbing sleep") and negatively (e.g., "pain does not limit my mobility") worded should be avoided in the same scale or questionnaire, because this makes less immediate the calculation of the total scale score (and errors more likely). In addition, endorsing a statement does not bear the same quantitative meaning of not endorsing its opposite ("feeling good" does not equate "feeling not bad").[15]

Are scale scores measures? Unfortunately, not. The true difference between 1 and 0, between 2 and 1, and either between or within items remains unknown.

Consider, for example, the trunk control test, a famous test of mobility applied to stroke and brain-injured patients in the early stages of recovery.[16,17] The test consists of four items (1, rolling to weak side; 2, rolling to strong side; 3, balance in a sitting position; 4, sit up from lying down), each with three categories and labeled: 0 (unable to complete the task without assistance), 12 (the patient meets the task weirdly), and 25 (completes the task normally).

Is there any reason to allow scores 0, 12, 25, rather than, say, 3, 7.5, 18.3? There is none. Does a change from 12 to 25 (i.e., 13 points) mean a difference more remarkable than that represented by a shift from 0 to 12 (i.e., 12 points) within each item? Nobody knows.

Some Clarifications on "Difficulty" and "Ability"

"Easier" and "more difficult" in the psychometric jargon do not necessarily indicate the motor or cognitive difficulty of the task described by an item. Instead, these adjectives stand for easier or more difficult to endorse, i.e., "more" or "less" likely to be observed. Mirror reasoning applies to subjects' "ability." This is a general term indicating the amount of a given trait or property owned by the subject (e.g., on a scale of disability, a person with higher scores can be said to be "more able" than subjects getting lower scores).

TABLE 1. Two hypothetical "scales" of mobility (A, B, and C, respectively)

A

	Sit	Stand up	Walk	Run	Total
Subject A	1	1	1	1	4
Subject B	1	1	1	0	3
Subject C	1	1	0	0	2
Subject D	1	0	0	0	1

B

	Sit	Stand up	Walk	Run	Total
Subject A	1	1	0	0	2
Subject B	0	0	1	1	2

C

	Sit	Walk	Speaking Italian	Total
Subject A	1	1	0	2
Subject B	1	0	1	2

Score 1 means that the behavior defined by the item names (columns) was observed by a rater (or endorsed by the subject); score 0 means: not observed/not endorsed.

A Higher Score May Not Imply a Higher Subject's "Ability"

Tables 1A and B show a simplistic didactic example. We want to measure the quantity of each subject's "mobility" rather than, like in the oranges metaphor, weight. From left to right, items are aligned to assume increasing "difficulty." This intuitive item ordering will be proposed again in this article, but one should consider that TTT in itself does not provide any formal estimation of the items' difficulty. Based on their raw scores, subjects are aligned from top to bottom in a (supposed) order of decreasing "ability," based on their raw scores. The questionnaire in Table 1A makes sense. As the overall score of the subjects rises (from bottom to top), progressively more difficult items are "passed." More able subjects only, pass more difficult items.

It is important to note that this sensible pattern implies that the items do not share the same difficulty level, although they can only be scored 0 or 1. This point is neglected whenever the overall score is only considered, regardless of which items (difficult or easy?) were passed.

This ideal pattern is dubbed the "diagonal" pattern ("1" and "0" scores lie on opposite sides of a perfect bottom-left to top-right diagonal line) or "Guttman" pattern (after the name of Louis Guttmann, 1916–1987, a famous statistician). Unfortunately, this pattern is virtually impossible to be observed in actual questionnaires, where some "0" and some "1" always happen to take the wrong place. An extreme case is shown in Figure 1B. Subjects A and B seem to share the same "mobility" level (both are scoring 2 of 4), but it is hard to believe that subject B cannot sit and stand while he/she can walk and jump. In subject B, items' difficulty and subject's ability do not match. Did the rater inadvertently reverse the scores? Are there any reasons unrelated to "mobility" for this weird behavior? Algebraic expedients may attenuate the problem, but TTT does not provide formal solutions, which would require a proper estimation of the item difficulty levels.

"MORE" OR "LESS"... BUT OF WHAT?

Table 1C introduces the requirement of "unidimensionality" of a measure, a cornerstone of Rasch modeling. Unidimensionality simply means that all the items included in a scale reflect (i.e., are indicators of, see hereinabove) of the same trait. A case is shown in which problems in scoring mobility arise, presumably, from the questionnaire, not from respondents. "Speaking Italian" requires some mobility (after all, one must use muscles to provide voice and speech). Still, for sure, this item reflects another variable, that is, knowledge of the Italian language, much more than "mobility."[13] Again, patients A and B seem to share the same mobility level: quite an absurd conclusion.

The three examples of Figure 1 are intentionally simplistic. Still, the reader can easily imagine how treacherous such "bugs" can be once hidden in a large data matrix with tenths of items administered to hundreds of people.

From Counts to Measures: An Unsurpassed Cleft

Let us suppose that a questionnaire is unidimensional and respects the ideal Guttmann pattern. The issue of quantity remains open. There are no a priori reasons to assume that "intervals," the "local distances," are invariant, that is, that 4–3 = 3–2. There are no reasons to assume that all oranges have the same weight nor that all scores represent the same amount of advancement, in the Table 1 examples, along the "mobility" continuous gradient. Does improving from standing to walking mean the same improvement in mobility represented by a change from walking to running? Probably, not. By contrast, a difference of 1 m means the same length difference, at whatever absolute length, for example, (1001–1000) m = (2–1) m. This concept is highlighted in a seminal article in the PRM literature, "Observations are always ordinal; measurement, however, must be interval."[7] "Interval," here, means "equal" intervals, that is, that the difference in quantity represented by a unit measurement is invariant. "Objective measurement is the repetition of a unit amount that maintains its size, within an allowable range of error, no matter which instrument, intended to measure the variable of interest, is used and no matter who or what relevant person or thing is measured" (from: rasch.org/define.htm, accessed April 30, 2022).

The TTT, that is, conventional psychometrics, was conscious that scores are rater dependent and nonlinear and strove from its birth (see Supplemental Digital Content 1, http://links.lww.com/PHM/B666) to estimate empirically the "objective linear measures" concealed by raw "scores." The TTT found sophisticated empirical solutions.[3,19] However, Rasch's analysis provided a formal answer to this problem.

Nonlinearity: Floor and Ceiling Effects

All concrete measurement instruments only allow a finite range of measures. This limitation entails a well-known floor-ceiling effect. Close to the extremes, changes of the variables' measures are underestimated: scores tend to "saturate" and crowd looking similar. This phenomenon holds for a bathroom scale as well as for a questionnaire. Figure 1[18] summarizes this and other issues. The abscissa gives the "true" measures, in unfamiliar "logit" units, adopted by RA. Here, it is sufficient to accept that logits are linear.[9,10] Like meters in length

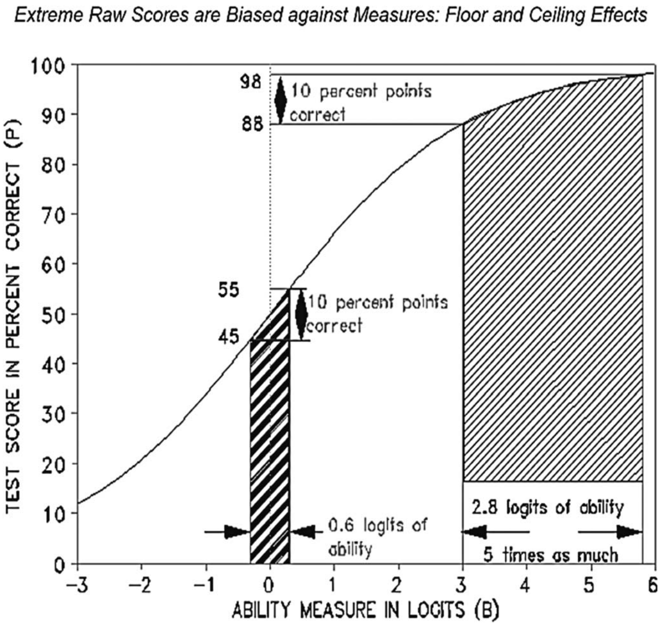

FIGURE 1. The raw scores from a 0–100 cumulative questionnaire (on the ordinate) are given as a function of the "true" linear measure (on the abscissa) of the variable tackled. Suppose scores are the number of correct responses (or, interchangeably here, the percent of correct answers). The true measure is given in linear logit units from RA. The increase in scores is roughly proportional to the rise in measure only in the "central" portion of the S-shaped (here, logistic) function (heavily hatched area). At the extremes (here, the "ceiling" of the scale, not the "floor," is highlighted), the same increase of 10 score points corresponds to a much higher increase in true measure (lightly hatched area). The scale's capacity to discriminate across subjects is lower the more the scores approach the ceiling of the questionnaire (reprinted with permission from Linacre, Fig. 1 in Wright and Linacre[18]).

measurement, a 1 logit difference means the same quantity change at whichever variable level. The ordinate gives the raw score (counts of observations), shown for convenience on a 0–100 scale. This S-shaped score-to-measure "logistic" relationship affects any concrete measurement. This is critical as in some patients, their ratings reach the finite range of measures even at the first visit. In these cases, assessing changes over time may be difficult to capture, sending false conclusions that no change has occurred.

Nonlinearity: Noninvariance of the Measurement Unit

The "metric ruler" metaphor, another cornerstone of Rasch modeling, needs to be anticipated here. Figure 2 shows a ruler bearing a familiar length scale in millimeters and centimeters on the top side. Numbers are proportional to length by construction. The difficulty level of items in a hypothetical questionnaire is marked on the bottom. There is no reason to assume that these marks are equally spaced. Take two children, A and B. On the upper metric scale, subject A increased his/her stature by 39 mm (from level A to A′), while subject B increased his/her stature by 16 mm (from level B to B′).

Now suppose that both children improved their "knowledge of math score" by 5 points as suggested by the "psychometric" scale on the ruler's bottom. Given the different "difficulty" levels represented by scores along the ruler, it is clear that subject A improved much more than subject B. Counting events is only a rough approximation to the measure of the variable causing those events. While TTT fails to consider this concept, Rasch modeling recognizes the difference.

The "Sufficiency" Requirement and the Issue of Redundancy

The actual difficulty of each item is only roughly estimated with TTT. Could we better know this difficulty level, we might predict, given the total score achieved by a subject, which is the

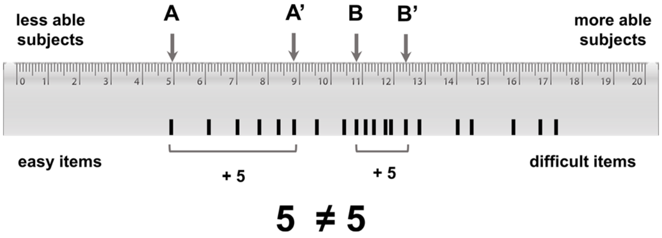

FIGURE 2. The metric ruler metaphor. Length units (millimeters or centimeters, top scale) are equally spaced by construction (i.e., by concatenating the measurement unit). The difficulty level of questionnaire items (bottom scale) can be assimilated to ticks along a ruler, but the distance between consecutive ticks is unknown. Subjects A and B increased their stature by a different amount. If they are compared along a scale of—say—math knowledge, where ticks represent item difficulties (bottom side of the ruler), they show an equal improvement in scores. However, their advancement along the knowledge gradient is very much different.

score expected in each item (see the ideal Guttmann pattern on Table 1A) and vice versa. This property is called "sufficiency." Suppose two or more items share the same difficulty level. In that case, they indicate the same overall ability level: if their scores are summed anyway, as in the case of TTT, then an illusion of a greater ability is by fault considered.

Take the following simplistic example of a 6-item math test. A score of 1 is given to each correct answer; 0 is assigned to wrong responses.

A) 3 + 2 =?; B) 5–4 =?; C) 4*5 =?; D) 9/3 =?; E) 6/2 =?; F) 4/1 =?

It is clear that once the respondent can answer correctly to item D, the first of the three divisions, he/she will likely answer correctly also to items E and F, the other two divisions (all with 1-digit factors and an integer quotient). The respondent will get three score points for the same ability level ("ability to solve this kind of division"). His/her score will artificially jump from 3 to 6. This "jackpot" passes easily undetected in many questionnaires but makes it hazardous to compare changes within and between subjects. Rasch analysis solves this problem elegantly.

The Puzzle of Missing Values

Missing items are common in questionnaires, resulting from different reasons. Typical examples are made by subjects who may inadvertently skip over some questions or may not have enough time to complete the test. Subjects would prefer not to answer if higher penalties are assigned to wrong than missing answers. Sometimes, some items are reserved to some respondents, for example, those who attempted in the past a given activity or have other characteristics such as sex or language.

When missing answers are found, the easiest and most practiced solution would be to transform scores into the percentage of the total score given by the answered items. For example, if a subject misses two items and totals seven on a questionnaire of 10 dichotomous items, the subject's "total score" is 87.5% (7/8). Note that this procedure is equivalent to the common "mean substitution" procedure,[20] assigning the missing items the average score observed in the other items. This method relies on the (strong and unlikely) assumption that questionnaires' items are all indicators of the same quantity of the latent variable of interest and thus exchangeable (i.e., they all share the same difficulty level).

Table 2 provides an example showing why this solution does not work. Suppose 10 items again, each scored 0 or 1, aligned from left to right for increasing difficulty. The "x" symbol flags a missing response. According to traditional psychometrics, subjects A and B are assigned the same score. Subject C, providing two missing answers, is assigned a lower score. However, it seems that subject A's missing response should be given a higher score estimate, given that the subject passed five more difficult items. In contrast, subject B failed three items easier than the missed one. It is doubtful that subject C deserves a lower score than subject B: he/she passed four items more difficult than the missed items.

Presumably, subjects A and C were careless in their answers or had not enough time to complete the test.

This example is intentionally simplistic. The treatment of missing data is a complex research field,[21] faced by the whole field of statistics.[22] As far as psychometric statistics is concerned, RA provides a satisfactory answer to this problem.[10]

The Issue of Reliability

Any measure is affected by error, even the seemingly precise physics measures. The only way to estimate the error surrounding the "true" measure underlying any observation is repeating the observation (across multiple times, multiple observers, or both). In the meanwhile, it is assumed that no systematic changes are occurring. In medicine, this means that no changes in the measured variable occur, for example, because of treatments or the disease's progression. Statistics can manage these changes and provide indexes of reliability.

The reliability is a proportion, ranging from 0 to 1:

$$\text{Reliability} = \text{true variance}/(\text{true variance} + \text{error variance}) \quad (\text{Eq. } 1)$$

Variance is the mean of the (squared) differences between individual values and their mean value. Thus, reliability tells how much the differences in measures across subjects reflect true differences in the amount of the measured variable. The "error" variance includes sources of variance modeled as "random." In addition, it may consist of extraneous systematic influences: these can affect all observations or stem from unsuspected interactions between some subjects and some items (see hereinafter).

Because of the inevitable measurement error, the total variance (the denominator in Eq. 1) is always larger than the "true" variance. In short, reliability = 1 means that all differences in scores reflect differences due to the amount of trait owned by the subjects. Reliability = 0 means that all differences reflect error (random and/or systematic).

Reliability is a property of the entire measurement process: it can change, for the same measuring instruments, depending on the procedures adopted (e.g., if a mean across several ratings and/or several raters, instead of single measurements, is considered; if reliability is tested across different time points, etc.).

TABLE 2. Estimation of missing values, based on average scores across answered items

		Item 1	**Item 2**	**Item 3**	**Item 4**	**Item 5**	**Item 6**	**Item 7**	**Item 8**	**Item 9**	**Item 10**	**Total**	**% Answered**
Subject	A	1	x	1	1	1	1	1	0	0	0	**6**	66.7%
	B	1	1	1	1	1	1	0	0	0	x	**6**	66.7%
	C	1	x	x	1	1	1	1	0	0	0	**5**	62.5%

Items (in bold) are aligned from left to right in order of increasing difficulty.

From top to bottom, subjects are aligned in order of decreasing ability. Allowable answers are 0 or 1. The "x" symbol flags missing responses. Total: total raw score. % answered: average percentage score computed on answered items, only.

Estimating reliability is estimating the measurement error concerning an assumed "truth." Conventionally, repeated measurements are run assuming that random error will gradually attenuate, and nonrandom values will not.

Errors May Be Systematic: How to Detect Them?

The systematic error components of scores are of the utmost relevance in behavioral sciences. Let's cite two reasons:

a) In all behavioral assessments, the subject accumulates learning and fatigue during repeated tests, which are in themselves a source of bias that may affect the true scores.
b) Some subjects may have peculiar reasons to get a given score in a given item (so-called "special knowledge" or "idiosyncrasies," see hereinabove). For instance, an item on a balance scale (e.g., "picking up an object from the floor") can be more difficult for visually impaired subjects than other participants. A math item including wordy descriptions may be more difficult for subjects with lower language proficiency.

"Errors" and "Residuals"

In TTT, if a single person scores "3" to an item, this is viewed as one of many possible replicated scores, the unattainable truth lying somewhere around it. In RA, even a single observation of score "3" can be considered reliable as long as it approaches "truth," which is defined a priori through a theoretical model (an "expected" score). Take the analogy of the ratio between the length of the diagonal and sides of a square. In a TTT perspective, this ratio would be measured in a sample of concrete (wooden, metal, paper) squares. Then, an empirical equation, unavoidably sample dependent, would be suggested. In a Rasch perspective, the "truth" would be provided by Pythagora's theorem. The measured ratio would be as true as it approaches the theory's prediction. A theory is conceived by humans based on empirical observations, of course. Still, after that, it must stay standstill with respect to the experience: in particular, it must stand sample independent (no concrete wooden or plastic triangles will ever invalidate the theorem). In questionnaires, once the Rasch model defines the expected score and (under a probabilistic theory) its error, reliability comes from the difference between the observed and the expected score. By contrast, TTT requires that repeated scores are collected, and their mean and SD computed. The TTT would provide errors with respect to an empirical (sample) mean, taken as inevitably provisional truth; RA provides "residuals," representing the distances between the (single) observed and the theory-expected (true) value. If the model works (like Pythagora's and Rasch's theories do), it provides a firm anchor (hence, a sounder form of reliability) to empirical measures.[9,10]

Repeatability Inflates Reliability if Items Are Too Easy or Too Difficult

Repeatability (not reliability) is easier to achieve if the difference between subjects' abilities and item difficulties is large. If a test is too difficult or easy for most subjects, total scores will be highly repeatable across raters and time points. Paradoxically, indexes based on repeatability of raw scores will be higher the less information the scale provides. If one already knows that a subject will not pass (or fail) a given item, that item is uninformative.

Things may be even subtler than that: repeatability may be inflated by local gaps in the ruler, thus being different depending on the overall ability of the individual subject. Suppose the actual "ruler" representing a questionnaire is depicted in Figure 3. In order of increasing difficulty, items are labeled from 0 to 6. Score 3 encases a wide range of ability levels. This situation will make this score highly "repeatable" across raters, time, etc. However, the measures' precision is low. Rasch analysis does not inflate reliability by this kind of forced repeatability.

Beware "Factors": They Challenge Unidimensionality

The concept of "systematic error" and "idiosyncratic" subject-item interactions sends us back again to the fundamental requirement of unidimensionality. Which are the "extraneous" dimensions or "factors" or "components" (here, other synonyms for variables or traits) steering the scores in the wrong direction? In physics, not less than in person metrics, measurement should only reflect the amount of measured variable (of course, this is an ideal property that can only be approximated in practice). The weight measure provided by a bathroom scale must not be influenced by room temperature and humidity or by the optical distortion of the lens magnifying the scale numbers or the magnitude of the subject's weight. The score on a pain questionnaire should not be influenced by depression, sex, language, etc. Finding "factors" in questionnaire scores is a standard procedure adopted to "validate" the questionnaires themselves. It is often forgotten that the stronger and more numerous factors are found, the more the scale is multidimensional.[23] Scales with complex factorial structures

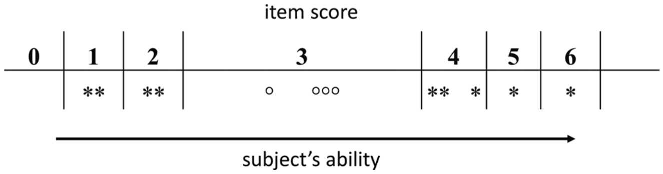

FIGURE 3. A ruler with irregularly spaced "ticks," representing items marking ranges of "ability," flagged by ordered numerals. The position of the bottom symbols along the trait continuum gives subjects' ability levels. Subjects "trapped" within the "3" ability interval (degree symbols) will likely receive the same score regardless of the raters, the time of assessment, etc., given that a massive change in ability is required to change to a score of 2 or 4. By greater force, subjects scoring "2" will hardly leap to score "4," and subjects achieving "4" will hardly leap to score "2." Randomness will hardly lead to a score change. A "reliability" index will be inflated while giving less information on the actual ability level of subjects: what seems a virtue is a fault.

usually boast of attaining "comprehensiveness" or "construct validity." However, this apparent virtue is obtained at the expense of unidimensionality, that is, of a fundamental property of any measure.

The "differential item functioning" is another index of multidimensionality. The example of cross-cultural differences is enlightening. Across linguistic/ethnic/national groups, total scores in a scale may be similar, but the position of the items along the ruler (i.e., their "calibration") may not. Rulers are qualitatively different. Take the familiar functional independence measure scale of independence in daily life, scored in a patient undergoing hospital rehabilitation. The "eating" item is more difficult in some East Asian countries, where chopsticks are preferred, than in Western countries, where cutlery is prevalently adopted.[24] "More difficult" means here that given an East Asian patient and a patient from a Western country with the very same total score of independence in daily life, the Asian one will score lower on the "eating" item (and higher in other items), compared with the Western patient (a consequence of differential item functioning). The scale conceals two distinct scales tackling distinct, incommensurable variables (independence in East Asian vs. Western cultures?).

Different Variables Can Score the Same: How Much of What?

The functional independence measure example above shows that the purely linguistic translation does not warrant equivalence of the "metric" meaning ("how much" of the variable are they representing) of the questionnaires compared. If this meaning is different, conceptually distinct variables are concealed under the same name.

The same reasoning applies to questionnaires applied to different diagnostic groups. The differential item functioning problem may find a solution within the Rasch modeling approach. See, for example, the studies by Arnould et al.[25] and Simone et al.[26]

The False Promise of Visual Analog and Numeric Rating Scales

The nature of the measured variable is even more questionable when the familiar Visual Analog Scales (VAS) are adopted. Here, scores are obtained by ticking a short straight segment with extremes labeled "no pain = 0" to "the worst imaginable pain = 100" and the like. However, the infinite precision promised by the continuity of the VAS graphic segment is illusory. For more than 60 years, psychology has taught that people cannot discriminate across more than 4–7 alternatives.[27,28] "Numeric Rating Scales" are conceptually equivalent to VAS: at least, they do not promise infinite precision. The "continuum" represented by a segment is fractionated into ordered levels (e.g., integers from 0 to 10). One-item scales are conceptually similar: this is the case for the Extended Disability Status Scale for multiple sclerosis (scored 0 = normality to 10 = death) or the Borg rating of perceived exertion category scale, scored 6 (very, very light) to 20 (very, very hard; see the complete *sralab.org* website for details on these and many other scales). Of more interest here (and often overlooked) is that the "variable" to be measured through VAS or Numeric Rating Scale instruments presumably is not the same across respondents. Two persons may choose to tick a 100-mm VAS-pain segment at 67 mm from the left origin; but by "pain," one person intends its peak intensity, the other its disabling impact, etc.[29]

To sum up, a question always lies in the background: how much of what? In itself, the "26" thick on a gauge may mean 26°C, 26 V, 26 secs, etc.[29] This introduces perhaps the most arduous argument, that is, the validity debate.

The "Validity" Debate

"Validity" is perhaps the most controversial property of scales because it is the most vaguely defined. In the literature, several forms of "validity" are proposed: "construct" validity (consistency of scores with other conditions supposed to determine them), "concurrent" validity (covariation with scores of already "validated" scales), "predictive" validity, etc. A historical definition, still widely accepted, is the one proposed by TL Kelley, who, in 1927, stated that a test is valid if it measures what it purports to measure.[30] This old definition, as it stands, sounds circular to the least. An ultrapragmatic (not to say simplistic) definition of validity comes from Guilford, who in 1946 argued that "a test is valid for anything with which it correlates."[31] Some covariation is necessary (measures in kilograms must covary with objects' volume). Still, it looks far from sufficient (what if objects are made of different materials, or the scale is sensitive to room temperature?).

We instead embrace the position that "a test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure."[32] The concepts of "existence"[33] and "causal relationship" between variable and measure are imbued with philosophical implications.[34] Suffices it to say here that researchers can invent a variable that does not exist outside their minds.

It should be stressed that validity is not (only) a statistical issue. Linearity and unidimensionality are necessary (mathematical) characteristics of measures, clearly insufficient for granting validity in the sense we endorsed.

CONCLUDING REMARKS

Questionnaires are the only available method to assess patients' latent traits.

Simply assuming that questionnaires' scores represent (unidimensional, linear) measures remain the root of approximations forecasting errors in many instances. A paradigm shift is needed to transform observed counts of events (the raw questionnaires' scores) into linear measures. Rasch statistical modeling provides this new paradigm. With linear measurements, all behavioral disciplines, including PRM, will have the possibility to bridge the scientific gap between biomedicine and clinical sciences.[35]

REFERENCES

1. Hambleton RK, Cook LL: Latent trait models and their use in the analysis of educational test data. *J Educ Meas* 1977;14:75–96
2. Steyer R, Schmitt M, Michael E: Latent state-trait theory and research in personality and individual differences. *Eur J Pers* 1999;13:389–408
3. Nunnally JC, Bernstein IH: *Psychometric Theory*, 3rd ed. McGraw-Hill, Inc, 1994
4. Luce RD, Tukey JW: Simultaneous conjoint measurement: a new type of fundamental measurement. *J Math Psychol* 1964;1:1–27

5. Carlson JE, von Davier M: Item response theory, in Bennett R, von Davier M (eds): *Item Response Theory. Advancing Human Assessment. Methodology of Educational Measurement and Assessment*. Springer, 2017:133–78. doi:10.4324/9781410605269
6. Rasch G: *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press, 1980
7. Wright BD, Linacre JM: Observations are always ordinal; measurements, however, must be interval. *Arch Phys Med Rehabil* 1989;70:857–60
8. Andrich D: *Rasch Models for Measurement*. Sage Publications, 1988
9. Tesio L, Caronni A, Kumbhare D, et al: Interpreting results from Rasch analysis 1. The "most likely" measures coming from the model. *Disabil Rehabil* 2022. (submitted)
10. Tesio L, Caronni A, Simone A, et al: Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. *Disabil Rehabil* 2022. (submitted)
11. Wright B: Measuring and counting. *Rasch Meas Trans* 1994;8:371–1
12. Becker KA: History of the Stanford-Binet Intelligence Scales: content and psychometrics, in: *Stanford-Binet Intelligence Scales*, 5th ed. Riverside Publishing, 2003
13. Tesio L: Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *J Rehabil Med* 2003;35:105–15
14. Tesio L, Alpini D, Cesarani A, et al: Short form of the Dizziness Handicap Inventory: construction and validation through Rasch analysis. *Am J Phys Med Rehabil* 1999;78:233–41
15. Yamaguchi J: Positive vs. negative wording. *Rasch Meas Trans* 1997;11:567–7
16. Collin C, Wade D: Assessing motor impairment after stroke: a pilot reliability study. *J Neurol Neurosurg Psychiatry* 1990;53:576–9
17. Franchignoni F, Tesio L, Benevolo E, et al: Psychometric properties of the Rivermead Mobility Index in Italian stroke rehabilitation inpatients. *Clin Rehabil* 2003; 17:273–82
18. Wright B, Linacre MJ: Fundamental measurement for outcome evaluation. *MESA Memorandum 66*. 1997. Available at: https://www.rasch.org/memo66.htm. Accessed April 30, 2022
19. Hassan S, Kumbhare D: Validity and diagnosis in physical and rehabilitation medicine: critical view and future perspectives. *Am J Phys Med Rehabil* 2022;101:262–9
20. Schafer JL, Graham JW: Missing data: our view of the state of the art. *Psychol Methods* 2002; 7:147–77
21. Chatfield C, Little RJA, Rubin DB: Statistical analysis with missing data. *J R Stat Soc Series A* 1988;151:375–6
22. Little RJA, Rubin DB: *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc, 2002. doi:10.1002/9781119013563
23. Franchignoni F, Mandrioli J, Giordano A, et al, ERRALS Group: A further Rasch study confirms that ALSFRS-R does not conform to fundamental measurement requirements. *Amyotroph Lateral Scler Frontotemporal Degener* 2015;16(5–6):331–7
24. Tennant A, Penta M, Tesio L, et al: Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. *Med Care* 2004;42:I37–48
25. Arnould C, Vandervelde L, Batcho CS, et al: Can manual ability be measured with a generic ABILHAND scale? A cross-sectional study conducted on six diagnostic groups. *BMJ Open* 2012;2:e001807
26. Simone A, Rota V, Tesio L, et al: Generic ABILHAND questionnaire can measure manual ability across a variety of motor impairments. *Int J Rehabil Res* 2011;34:131–40
27. Miller GA: The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol Rev* 1956;63:81–97
28. Ma WJ, Husain M, Bays PM: Changing concepts of working memory. *Nat Neurosci* 2014;17: 347–56
29. Franchignoni F, Salaffi F, Tesio L: How should we use the Visual Analogue Scale (VAS) in rehabilitation outcomes? I: how much of what? The seductive VAS numbers are not true measures. *J Rehabil Med* 2012;44:798–9
30. Kelley TL: *Interpretation of Educational Measurements*. Yonkers-on-Hudson, NY, World Book Company, 1927
31. Guilford JP: New standards for test evaluation. *Educ Psychol Meas* 1946;6:427–38
32. Borsboom D, Mellenbergh GJ, Van Heerden J: The concept of validity. *Psychol Rev* 2004; 111:1061–71
33. Agazzi E: *Scientific Objectivity and Its Contexts*. Heidelberg, Springer International Publishing, 2014. doi:10.1007/978-3-319-04660-0
34. Buzzoni M, Tesio L, Stuart MT: Holism and Reductionism in the Illness/Disease Debate, in Wuppuluri S, Stewart I (eds): *From Electrons to Elephants and Elections. The Frontiers Collection*. Cham, Springer, 2022:743–78
35. Tesio L: Measurement in clinical vs. biological medicine: the Rasch model as a bridge on a widening gap. *J Appl Meas* 2004;5:362–6