



## Interpreting results from Rasch analysis 1. The “most likely” measures coming from the model

Luigi Tesio, Antonio Caronni, Dinesh Kumbhare & Stefano Scarano

To cite this article: Luigi Tesio, Antonio Caronni, Dinesh Kumbhare & Stefano Scarano (2023): Interpreting results from Rasch analysis 1. The “most likely” measures coming from the model, Disability and Rehabilitation, DOI: [10.1080/09638288.2023.2169771](https://doi.org/10.1080/09638288.2023.2169771)

To link to this article: <https://doi.org/10.1080/09638288.2023.2169771>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 05 Feb 2023.



[Submit your article to this journal](#)



[View related articles](#)



[View Crossmark data](#)

# Interpreting results from Rasch analysis 1. The “most likely” measures coming from the model

Luigi Tesio<sup>a,b</sup> , Antonio Caronni<sup>b</sup> , Dinesh Kumbhare<sup>c,d</sup>  and Stefano Scarano<sup>a,b</sup> 

<sup>a</sup>Department of Biomedical Sciences for Health, Università Degli Studi Di Milano, Milan, Italy; <sup>b</sup>IRCCS, Istituto Auxologico Italiano, Department of Neurorehabilitation Sciences, Ospedale San Luca, Milan, Italy; <sup>c</sup>Department of Medicine, Division of Physical Medicine and Rehabilitation, University of Toronto, Toronto, Ontario, Canada; <sup>d</sup>Pain Research Institute, Toronto Rehabilitation Institute, University Health Network, Toronto, Ontario, Canada

## ABSTRACT

**Purpose:** The present article summarises the characteristics of Rasch’s theory, providing an original metrological model for persons’ measurements. Properties describing the person “as a whole” are key outcome variables in Medicine. This is particularly true in Physical and Rehabilitation Medicine, targeting the person’s interaction with the outer world. Such variables include independence, pain, fatigue, balance, and the like. These variables can only be observed through behaviours of various complexity, deemed representative of a given “latent” person’s property. So how to infer its “quantity”? Usually, behaviours (items) are scored ordinally, and their “raw” scores are summed across item lists (questionnaires). The limits and flaws of scores (i.e., multidimensionality, non-linearity) are well known, yet they still dominate the measurement in Medicine.

**Conclusions:** Through Rasch’s theory and statistical analysis, scores are transformed and tested for their capacity to respect fundamental measurement axioms. Rasch analysis returns the linear measure of the person’s property (“ability”) and the item’s calibrations (“difficulty”), concealed by the raw scores. The difference between a person’s ability and item difficulty determines the probability that a “pass” response is observed. The discrepancy between observed scores and the ideal measures (i.e., the residual) invites diagnostic reasoning. In a companion article, advanced applications of Rasch modelling are illustrated.

## ARTICLE HISTORY

Received 26 February 2022  
Revised 13 January 2023  
Accepted 13 January 2023

## KEYWORDS

Questionnaires; Rasch analysis; physical and rehabilitation medicine; measurement; latent variables; psychometrics; metrology

## ► IMPLICATIONS FOR REHABILITATION

- Questionnaires’ ordinal scores are poor approximations of measures. The Rasch analysis turns questionnaires’ scores into interval measures, provided that its assumptions are respected.
- Thanks to the Rasch analysis, accurate measures of independence, pain, fatigue, cognitive capacities and other whole person’s variables of paramount importance in rehabilitation are available.
- The current work is addressed to rehabilitation professionals looking for an introduction to interpreting published results based on Rasch analysis.
- The first of a series of two, the present article illustrates the most common graphic and numeric outputs found in published papers presenting the Rasch analysis of questionnaires.

## The challenge of measuring a person’s behaviours and perceptions


This article is the first of two companion articles on Rasch Analysis (RA) [1] addressed to rehabilitation professionals, and other clinicians interested in understanding published results based on RA of cumulative questionnaires (or “scales”). Scales were initially applied to measure psychological variables (starting, perhaps, from the Stanford-Binet I.Q. assessing “intelligence” [2]). Likely for this reason, this scientific field is still named “psychometrics.” All areas of Medicine (including Physical and Rehabilitation Medicine – PRM) need to manage scales to measure patient-centred outcomes [3]. This is the case for pain, fatigue, independence, balance, mobility, continence, and cognitive capacities, to name a few.

RA provides a theoretically sound solution to a person’s measurement. However, because of its original conceptual approach [4], RA struggles to spill out of its circle of followers.

## Complementing the available literature

Guidelines for publishing Rasch results are available [5–7]: however, these are more addressed to researchers than the lay reader. Examples of Rasch-based articles and books are given below in the “How to learn more” paragraph. In the typical medical reports, the Methods section needs to be expanded to describe these (still unfamiliar) Rasch statistics, thus burdening the manuscripts themselves. Unfortunately, the explanation of the Rasch technicalities is difficult enough not to allow the typical reader to attain a critical

**CONTACT** Luigi Tesio  [l.tesio@auxologico.it](mailto:l.tesio@auxologico.it)  Department of Neurorehabilitation Sciences, Istituto Auxologico Italiano, IRCCS, Ospedale San Luca, via Giuseppe Mercalli 32, 20122 Milan, Italy

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/09638288.2023.2169771>

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

appraisal of the results. Exhaustive books are available, but they require a relevant investment in time, assume more than elementary statistical knowledge, and are not (obviously) focused on rehabilitation issues. In a sentence, they are not tailored to the clinical needs of rehabilitation professionals. Therefore, a series of three articles were realised to fill this gap. A published article summarised why questionnaire scores might be misleading [8]. The present article focuses on how the problem can be solved: it describes the principles of an “ideal” scale obtained through Rasch modelling the original raw data. A second companion article also focuses on Rasch modelling but explores the model technicalities more in-depth [1].

## **Classical Test Theory is not enough: the “latent trait” approach and the Item Response Theory**

### ***Why bother with questionnaire scores? After all, they provide numeric scores***

In questionnaires’ items, behaviours or perceptions deemed to represent a person’s variable are given numerical labels called “scores” (e.g., 0/1 = absence/presence or, 0/1/2/3 = severe/moderate/mild/absent).

In Classical Test Theory (CTT) [9–11], items’ scores are summed in the total questionnaire score to quantify the variable. The many flaws of this form of measurement are well known [12]. To the least, one should remember that numbers representing discrete counts are not necessarily proportional to the variable’s amount. Counting individual oranges provides numbers not proportional to their weight. In the same vein, on a scale of independence walking, the difference 3 (“use of rollator”) – 2 (“with someone’s support”) does not necessarily mean the same substantial difference in independence represented by the difference 4 (“fully autonomous”) – 3 (“use of rollator”).

### ***Searching for “latent” measures generating raw scores***

#### ***Do scores reflect measures or are they measures themselves? Two conflicting theoretical models***

For several decades, and still nowadays, the “latent trait” approach provided an innovative conceptual framework for measuring a person’s variable [13–15]. In essence: (1) the person’s variable is assumed to be “hidden/latent” within the person; (2) it can be observed only through a sample of potentially infinite behaviours. Therefore, based on observed scores, (3) inferences are necessary to estimate its amount [13,15,16]. This approach took the name of Item Response Theory (IRT). For CTT, the single score must be taken as the “truth” of that single observation [17]. By contrast, IRT adopts a probabilistic-inferential approach from scratch: the path from an item to a person’s single observed response is probabilistic. An error surrounds the estimates. Scores are (or they are supposed to be) ordered from “less-to-more” or vice versa, yet they remain arbitrary: “severe/moderate/mild/absent” can be labelled 0/1/2/3 or 10/23.5/48/122, or A/B/C/D.

#### ***The “item response theory”: scores as starting points to discover the latent measure***

Let’s consider only dichotomous items (no/yes, failed/passed) and numeric integer labels (i.e., the item’s score: 0/1). The “0” or “1” scores must be transformed into “the probability that one was observed.” This seems relatively easy to grasp, given that scores and probability both range from 0 to 1. However, their

substantive meaning is different. A person facing a given item might have a  $0.84 \pm 0.03$  probability of getting “1.” Still, the same probability might hold if the observable responses were labelled, say, 0/5, like in some items of the Barthel index of independence in activities of daily living. The reader needs to become familiar with the difference between the “observed” score (the response given to single items) and the “model-expected” score, i.e., the probability (coming from a statistical model) that a given response was observed. Also, the total score cumulated across items is transformed, in IRT jargon, into the person’s “measure.” The theory implies that the measures of quantities should be discovered through inferences from the observed scores. This theory also needs to be substantiated in statistical models. Following these models, a mathematical analysis transforms the arbitrary “raw” scores into measures based on the response probability. The analysis also provides many other related indexes (e.g., reliability, data-model fit, etc., as shown later). RA is often an umbrella term encompassing a measurement theory and statistical models. A rigorous epistemic study should treat these as related yet distinct concepts [18].

An effort will be made here to use the term “theory” to highlight the adherence of RA to fundamental measurement axioms, the term “model” to indicate the statistical equations stemming from the theory, and the term “analysis” to mean the calculations providing various measures and indexes. However, the umbrella term “Rasch Analysis,” so familiar in the medical literature, could not be entirely avoided.

## **Rasch’s theory: discovering measures underlying numbers**

Rasch’s theory is usually considered part of the IRT. The theory’s name comes from Georg Rasch, a Danish mathematician (1901–1980). His statistical model, substantiating an innovative theory, was initially published in Danish in 1960, but it caught the eye of the psychometric community only after its publication in English in 1980 [19]. The community of rehabilitation professionals probably started considering the model only after a seminal article published in 1989 [20]. Nowadays, a family of “Rasch models” is available, compliant with the original theory. It is worth citing here, along with the original model for the analysis of dichotomous items, models for the analysis of polytomous items (the “rating scale” and the “partial credit” models), and the “many-facet” model that takes into account, beyond items and persons, additional players (e.g., the raters). The number of articles in indexed Journals is increasing [21].

The companion article [1] will give more information on the modelling procedures. It will also face the analysis of consistency between observed data and the modelled scale (the data-model “fit” issue).

## **The Rasch theory is unique within the item response theory: a prescriptive, not a descriptive, theory**

The basic idea shared by all IRT models (RA included) is that the probability that a given person gets a given score in a given item is governed by the “ability” ( $\beta$ , the amount of any property/trait) of the person and the “difficulty” ( $\delta$ ) of the item. Unlike other IRT models, however, the Rasch theory imposes that only ability and difficulty govern the response probability (see below). The Rasch theory implies that more able people have a higher probability, compared to less able people, to pass any items; and that more

difficult items have a lower probability of being passed, compared to easier items, by any person.

Other IRT models complicate the model by considering the effect of different variables on the response probability. As an example, take the latent variable “knowledge of mathematics.” In a math test, a person more proficient in math is expected to pass items more difficult than those passed by a less proficient person. However, guessing or previous knowledge of one exercise, or other extraneous variables (e.g., tiredness), may affect the probability to complete the exercise correctly.

Of course, a statistical model can be modified to improve its fit to observed data. This is done in some IRT analyses (see [Supplemental Material, Note 1](#)) but is strictly avoided in the RA. For this reason, according to some, Rasch’s theory lie outside the IRT family.

In simple words, RA is prescriptive: it asks the data to fit the model, while other IRT are descriptive, asking the model to fit the data. This original standpoint can contribute to the scientific community’s resistance to RA [22], perhaps simply a variant of the well-known opposition of scientists to scientific discoveries [23].

### Rasch theory asks for measures independent of persons and items

In CTT, a sample of persons provides scores, and the distribution of the scores can be easily partitioned into “units,” e.g., percentiles. This distribution (and what a “unit” means) is unavoidably sample-dependent. This may have consequences on the generalisability of the findings. For instance, a person may “measure” in the 60th percentile when sample A is used for calibration but in the 80th percentile when sample B is used (if sample A persons are more proficient than those of sample B). Similarly, two persons with the same ability might get different total scores when facing scales with other items (if the items of the two tests have different difficulties, albeit tackling the same variable).

In his “separability” theorem, Rasch demonstrated [24] that measures become independent from the particular sample of items and persons if the data conform to his model, only. It is worth highlighting that an “if” reservation works in the background in every science, including physics: what “if” a plastic ruler gives different measures of the same object according to local temperature? Ideally, measures must be linear (what is meant by 2-1 equals what is meant by 1923–1922) and unidimensional (unbiased by extraneous variables). Of course, no measure is entirely linear or unidimensional in the real world. Again, Rasch’s model appeals to general measurement axioms and it estimates (in terms of probability) the discrepancy between empirical items’ scores and model-expected scores. Under this perspective, differences with respect to measurement in “hard” sciences fade away (see also below the paragraph titled “Limiting and widening the scope of the present and the companion articles”).

### Rasch theory turned into a statistical model: a simple yet very demanding, equation

The very intuitive Equation embodying the Rasch model is:

$$P(\text{pass}) = f(\beta - \delta) \quad (1)$$

This can be read as: the probability  $P$  that a “pass” is observed (i.e., that the higher of two alternative scores is deserved) is “a function of” the difference between ability and difficulty, and only of this difference. This function needs to be an increasing monotonic one, of course. “Probability to get score 1” (or whatever

label stands for a quantity of the variable) is the “expected score” (reaching probability 0 or 1 asymptotically). Suppose a person is proficient in math and expected to pass a very easy item with a probability of 0.94: what if he/she failed (observed score = 0)? Is this discrepancy statistically significant? The difference between expected and observed scores should stimulate diagnostic reasoning.

### Modelling probabilities to pass. The “item characteristic curve” – ICC

How to model the so far generic  $f$  function relating the ability-difficulty difference to the “pass” probability to a given item? The “function”  $f$  in [Equation \(1\)](#) can be rather complicated and may change according to different IRT models. Details will be provided in the companion paper [1].

### The Rasch model summarised by an equation estimating probabilities to pass or fail

In the model proposed by Georg Rasch [19,25], the function is represented graphically by an S-shaped curve ([Figure 1](#), panels A and B) asymptotically reaching 0 or 1 when  $(\beta - \delta)$  goes  $-$  or  $+$  infinity, respectively. [Figure 1](#) shows the so-called “item characteristic curves,” ICC. [Figure 1\(A\)](#) represents the more straightforward case for dichotomous items (0/1; no/yes; fail/pass), i.e., the issue considered by the original Rasch model.

The ordinate gives the probability of getting a score of 0 or 1 (e.g., “fail” =  $1 - P$ ; or “pass” =  $P$ ) as a function of the difference between a person’s ability and the item’s difficulty  $(\beta - \delta)$ , on the abscissa. In [Figure 1\(B\)](#), the ICCs for a series of dichotomous items are represented (only the probability to “pass” is displayed). The units on the abscissa, likely unfamiliar to the reader, are called “logits” (below, in this figure).

### The weird “logit” units: why complicate your life?

A logit is the natural logarithm of the pass/fail ratio (the odds, a quantity familiar to bookies and gamblers):

$$\text{logit} = \ln(P/(1 - P)) \quad (2)$$

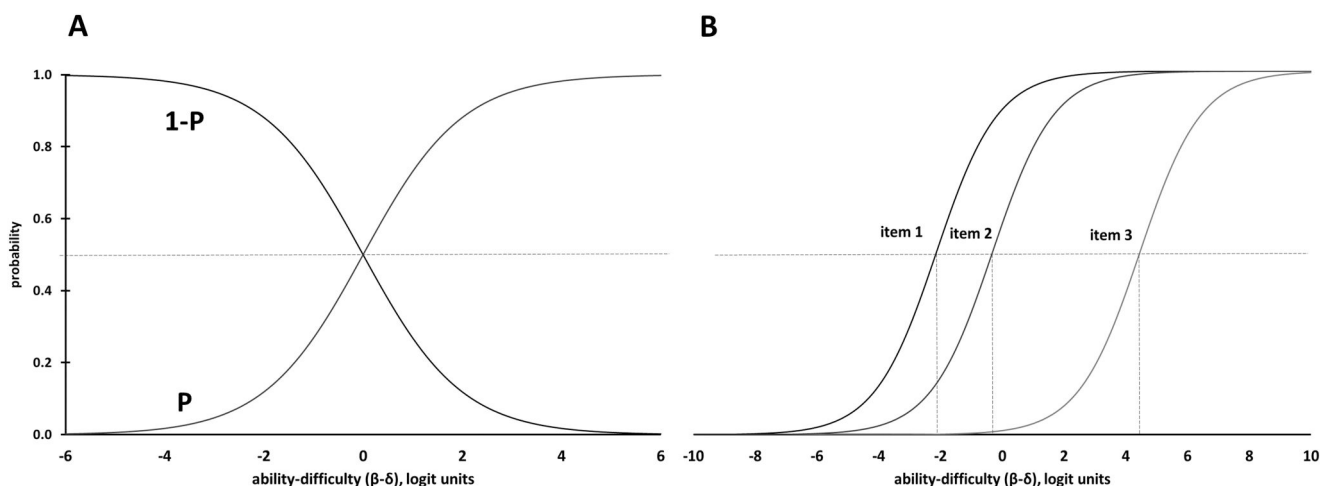
In [Equation \(2\)](#) above,  $P$  = pass probability and  $(1 - P)$  = fail probability, when only pass or fail are foreseen. Probabilities are confined between 0 and 1, but the difference between ability and difficulty may range from  $-$  to  $+$  infinity. Logits also may go  $-$  to  $+$  infinity, thus allowing a linear relationship with the  $\beta - \delta$  difference. Logits (not probabilities) are the conceptual equivalent of metric units on a ruler. The perhaps counterintuitive need for transforming probabilities into odds and finally into logits is motivated in the [Supplementary Material, Note 2](#).

### The Rasch model is summarised by an equation estimating linear measures of ability and difficulty

In RA, [Equation \(1\)](#) takes the form:

$$\ln(P/(1 - P)) = \beta - \delta = \text{logit} \quad (3)$$

The logits represent the measurement units. The logit measure (of items or persons) is also called “calibration” in the Rasch literature. In everyday language, calibration compares a measure to a traceable standard. A more proper term for the logit unit should be “calibrator” because it represents the bar over which ability and difficulty are confronted. Here, the lay reader only needs to



**Figure 1.** Item characteristic curves (ICC). Graphs show the ICC of dichotomous items according to the Rasch model. (A) The panel shows the relationship between the probability that response 1 (the P label; increasing curve) or 0 (the 1 – P label; decreasing curve) is observed, on the ordinate, as a function of the difference between the person’s ability and the item’s difficulty ( $\beta - \delta$ ), on the abscissa. The greater the person’s ability, the higher the probability of observing 1, and the lower the probability of observing 0. The ability value at which the curves cross (both responses are equally likely) is called a “threshold.” When  $\beta = \delta$  there is the same probability of observing either response (here, 50%). Probabilities reach 0 or 1 only asymptotically. (B) Three dichotomous items within a scale are represented (probability to get a score of 1 only; decreasing curves are not shown). Item 2 (the same represented in panel A) is more difficult than item 1, and item 3 is more difficult than item 2. The curves run parallel (not necessarily at equal distances on the abscissa). Whichever the person’s ability, the probability of observing response 1 is lower the more difficult the item is.

consider that the logit unit remains invariant along the variable’s span. Suppose a patient reduces his/her disability from 3 to 1 logit. This improvement is the same as another patient passing from 6 to 4 logit. Logits work like °C and °F (as well as cm and kg).

### When “pass” and “fail” have the same probability of occurring: understanding the “threshold” concept as a measure of item difficulty

The left panel in [Figure 1](#) shows that the more score 1 is probable, the less score 0 is probable, given that only 1 or 0 can be observed. The  $\beta - \delta$  value at which the two curves cross, i.e., the adjacent scores are equally likely, is called a “threshold” (more precisely, these 50-50 thresholds are referred to as Andrich’s modal thresholds, or  $\tau$  parameters). In dichotomous items (like those in [Figure 1\(A,B\)](#)), the threshold marks a 50% pass probability, meaning that ability and difficulty are the same sizes. Algebra tells that if  $\beta = \delta$ , then  $\beta - \delta = 0$  and  $\ln(P/(1 - P)) = 0$ , so that  $P/(1 - P) = 1$ , i.e.,  $P = (1 - P)$ . A 0-logit difficulty value is conventionally set at the mean value across thresholds on a scale made by several items. For illustrative purposes, the item in panel A has its threshold at 0 logits. Of course, not all items should have their thresholds at 0 logits on a scale. Panel B shows the ICC of three items. Item 2 is the exact item represented in panel A. Item 1 is less difficult, and item 3 is more difficult than item 2. Correspondingly, their thresholds (i.e., a 50% pass probability) are found at lower and higher logit values (higher persons’ ability levels). However, the curves’ slope (i.e., the item’s discrimination, the item’s capacity to assign different probabilities to persons with varying levels of ability) is the same across all items, so the curves are parallel. This is why the Rasch model is also called “1-PL” (1-parameter logistic) model by some authors. This property is peculiar to the Rasch model within IRT models (see [Supplementary Material, Note 1](#)). This graph indirectly shows that for any given ability – difficulty difference, probabilities to pass are progressively lower for items 1, 2, and 3, respectively: a stringent and peculiar requirement of Rasch modelling, which is in agreement with the intuitive expectation that the more difficult is

an item, the lower is the probability of passing it, no matter how “able” the subject is.

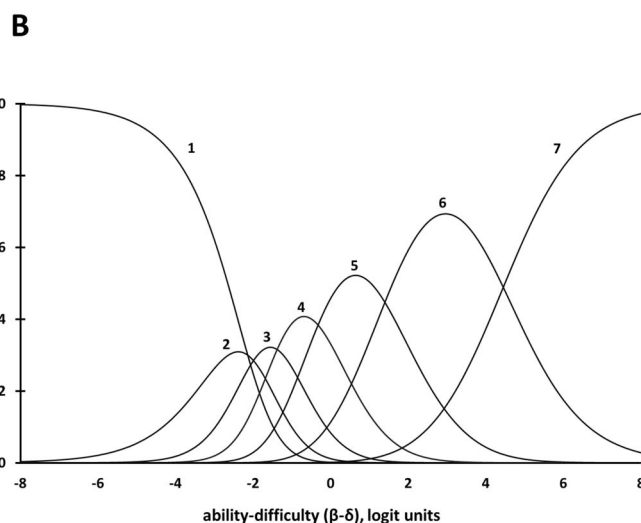
### Understanding outputs from Rasch analysis: the “category probability curves”

Items can be polytomous, i.e., graded on discrete levels (“categories”) such as - imagine an item in a pain scale - severe/moderate/mild/absent = 0/1/2/3, and the like. As anticipated, the labelling of categories is irrelevant, provided they are conceived as quantitatively ordered (e.g., from small to large, from rare to frequent, etc.). RA re-scores responses as 0/1/2/3, etc., and estimates thresholds across categories. When the probability of getting a score of 1 increases, the probability of getting a score of 0 decreases, etc. The combined graphic effect is a “hilly pattern,” like the one depicted in [Figure 2\(B\)](#): the Category Probability Curves (CPC), a visual output widespread in Rasch’s published articles.

[Figure 2](#) refers to the FIM™-Functional Independence Measure questionnaire (motor subscale), an instrument widely adopted worldwide [26] ([Figure 2\(A\)](#)). The FIM includes 13 items concerning self-care, sphincter management, mobility/transfer, and locomotion. Five more items cover communication and social cognition domains, not considered here [27]. According to a detailed manual, each item can be scored from 1 to 7: the higher, the greater the person’s independence in daily activities. The scores from 300 stroke patients at discharge from an inpatient rehabilitation unit were analysed (first author’s data, see [Legend for details](#)). In panel B, the curves give the model-expected probabilities that a given category is observed (on the ordinate) as a function of the difference between the subject’s ability and item difficulties (on the abscissa). The model (Andrich’s “rating scale” Rasch model, see the companion paper for details) assumes that the pattern of category difficulties is the same across items. The mean threshold difficulty in the graph is “centralised” at 0 logits so that the same chart represents all items. Of course, lower absolute probabilities are expected for each category when more difficult items are faced by a given patient (not illustrated). It can be seen that the categories work as intended: on average, higher

## A FIM™- Functional Independence Measure

- |  |   |
|--|---|
| <p><b>Self Care</b></p> <ul style="list-style-type: none"> <li>A. Eating</li> <li>B. Grooming</li> <li>C. Bathing</li> <li>D. Dressing-Upper body</li> <li>E. Dressing-Lower body</li> <li>F. Toileting</li> </ul> <p><b>Sphincter Control</b></p> <ul style="list-style-type: none"> <li>G. Bladder Management</li> <li>H. Bowel Management</li> </ul> <p><b>Mobility / Transfer</b></p> <ul style="list-style-type: none"> <li>I. Bed-Chair-Wheelchair</li> <li>J. Toilet</li> <li>K. Tub-Shower</li> </ul> <p><b>Locomotion</b></p> <ul style="list-style-type: none"> <li>L. Walk-Wheelchair</li> <li>M. Stairs</li> </ul> | <p><b>Communication</b></p> <ul style="list-style-type: none"> <li>N. Comprehension</li> <li>O. Expression</li> </ul> <p><b>Social Cognition</b></p> <ul style="list-style-type: none"> <li>P. Social Interaction</li> <li>Q. Problem solving</li> <li>R. Memory</li> </ul> <p><b>SCORING LEVELS</b></p> <ul style="list-style-type: none"> <li>7. Complete independence</li> <li>6. Modified Independence</li> </ul> <hr style="width: 50%; margin-left: 0;"/> <ul style="list-style-type: none"> <li>5. Supervision</li> <li>4. Minimal assistance</li> <li>3. Moderate assistance</li> <li>2. Maximal assistance</li> <li>1. Total assistance</li> </ul> |
|--|---|



**Figure 2.** Items of the functional independence measure. (A) The FIM™ is an 18-item questionnaire scoring the subject’s independence in daily activities. Scores may range from 1 to 7, the higher the patient’s independence in daily life. The 13 items A to M can be used as a “motor” subscale (score range: 13 to 91); the five items N to R can be used as a “cognitive” subscale (score range: 5 to 35) [26]. (B) The category probability curves of the FIM items. Given that there are n scoring options (categories), here there are  $n - 1 = 7 - 1 = 6$  “thresholds” (ability levels at which adjacent scores are equally likely). The 13 “motor” subscale items (A to M in panel A) were only considered. Scores were recorded in 300 post-stroke patients (157 men), mean age 69 (SD: 14) years, at discharge from an inpatient rehabilitation unit (first author’s data). In this analysis, all items are expected to present the same pattern of category difficulties (Andrich’s “Rating scale” Rasch model, see the companion paper). The probability of observing a given score (on the ordinate) is plotted against the difference between the patient ability and the item difficulty (logit units, on the abscissa). The mean difficulty level of the six thresholds is conventionally assigned (“centralised”) at 0 logit measure to foster comparisons of the thresholds’ patterns across items.

**Table 1.** FIM motor sub-scale (see Figure 2).

| Category | Adjacent categories | Threshold difficulty (logits) | Category measure (logits) |
|----------|---------------------|-------------------------------|---------------------------|
| 1        | –                   | –                             | –3.66                     |
| 2        | 1–2                 | –2.14                         | –2.38                     |
| 3        | 2–3                 | –1.99                         | –1.56                     |
| 4        | 3–4                 | –1.38                         | –0.69                     |
| 5        | 4–5                 | –0.33                         | 0.64                      |
| 6        | 5–6                 | 1.40                          | 2.96                      |
| 7        | 6–7                 | 4.45                          | 5.58                      |

Notes: First author’s data. From left to right, the columns give the category labels, the pairs of categories between which thresholds are placed, the thresholds’ values, and the average level of ability of persons selecting any given category (the “mean” category difficulty, or category measure), respectively.

scores (from 1 to 7) are more “difficult,” i.e., they represent “more” independence and require higher abilities to be achieved. In addition to categories, thresholds are also “ordered.” The boundary between scores 1 and 2 marks a lower ability level than the boundary between scores 2 and 3, etc. The graphic counterpart is that all categories “emerge” over the adjacent categories for at least some range of ability levels (a limited one for category 2, a large one for category 6, etc.).

It is interesting to consider the numeric counterpart of Figure 2(B). It must be highlighted that the “difficulty” of the category is the mean of the model-estimated ability levels of persons selecting a given category across all items (Table 1). Different software programs may provide these values in different graphic outputs.

It may be seen that thresholds are “ordered,” not less than the average measure of the categories. What “more” or “less” means is consistent with their intended ordering.

### Disordered thresholds can coexist with ordered categories: a counterintuitive property of the model

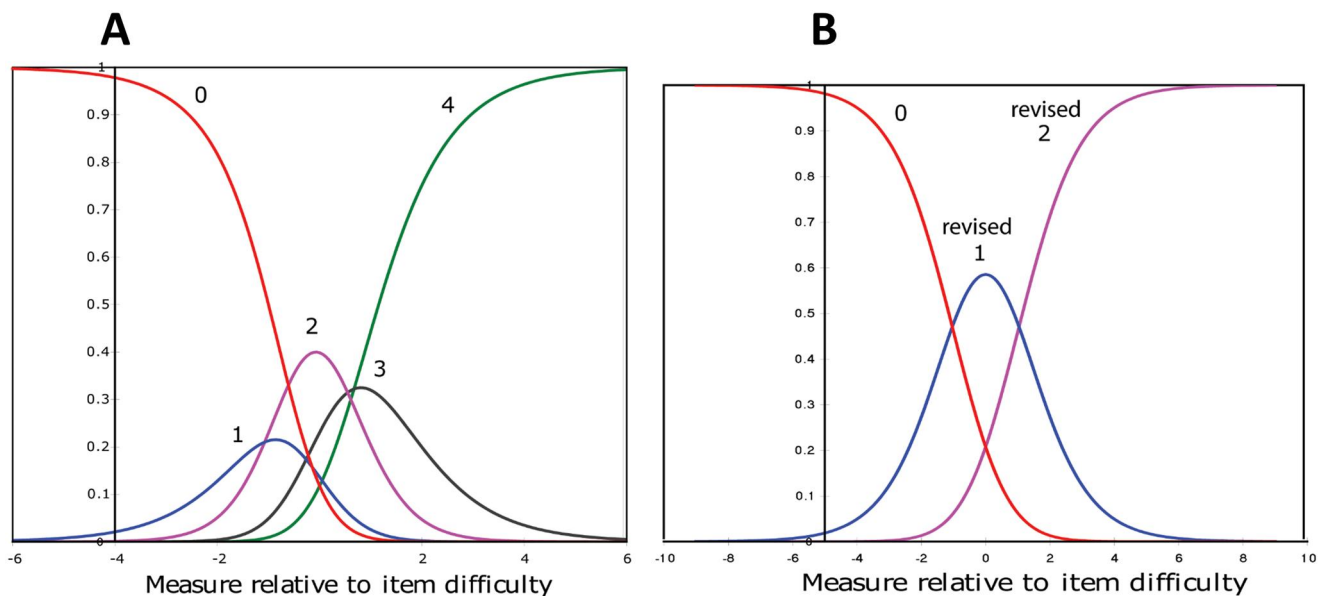
In building a new scale or testing the validity of an existing scale through RA, Andrich’s thresholds may often appear disordered,

i.e., not ordered “from-less-to-more” as intended, even when the difficulty of the categories remains ordered. This difference may originate when a category is seldomly chosen by respondents, a condition that can happen for several reasons, including a poor conceptualisation of the item’s categories. For instance, is “sometimes = 2” perceived sharply by the rater as “more” of a given property compared to “occasionally = 1”? Does the (ambiguous but popular) “don’t know” category represent an intermediate quantity across adjacent categories? Typically, one or more of the “hills” get submerged in such instances.

Figure 3(A) gives the CPC of items in a rating scale of health-related Quality of Life in Strabismus (AS-20, revised version) [28], applied to 584 adults.

The original scale consists of 20 items: illustrative items are “I feel inferior to others because of my eyes” and “I need to take frequent breaks when reading because of my eyes.” Items are scored as never = 0, rarely = 1, sometimes = 2, often = 3, always = 4. Therefore, on this scale, the higher the score, the worse the condition. Something unexpected pops up in Figure 3(A). Category 1-rarely did not emerge in the initial analysis (panel A). The never-rarely (0-1) threshold (the logit value at which the corresponding curves cross) flagged a worse condition than the rarely-sometimes (1–2) threshold. At the same time, it can be caught by eye that, on average, the ability measures subtended by the various categories are ordered. People responding “0” are, on average, less able than people responding “1,” who are less able than people responding “2,” etc. Ordered categories and disordered thresholds can thus coexist.

According to some scholars, threshold “disordering” is a minor problem if the average measures subtended by the categories remain ordered [29]. However, Andrich and other researchers affirm that thresholds should also be ordered to conclude that the item’s category structure functions as intended [22]. There is an actual controversy about these aspects of Rasch modelling.



**Figure 3.** Category probability curves of the items of the Adult Strabismus-20 questionnaire. Questionnaire items (20) are scored 0 to 4; the higher, the worse the condition. The category labelled 1 is submerged so that the 0–1 Andrich’s threshold is trespassed by persons with a worse condition than persons trespassing the 1–2 threshold. After collapsing some categories (the scoring pattern changed from 01234 to 01122), the thresholds became ordered (after [28]). See the text for the scale’s description.

### How to manage disordered thresholds

Disordered thresholds flag that the corresponding submerged category is never the category most likely selected by respondents (here, Category 1). In other words: the submerged category is never modal. A remedy for both disordered categories and thresholds consists in assigning to the submerged category the score of an adjacent category (“rescoring”), thus “collapsing” (a typical term in Rasch jargon) adjacent categories. The authors of the questionnaire represented in [Figure 3](#) “re-scored” category 2 by assigning score 1. Categories 1 and 2 were “collapsed” so that the scoring options were changed a-posteriori on the observed data from 01234 to 01123. Unfortunately, this made category 2 (former category 3) to submerge (not illustrated). Then, the authors also collapsed the new categories 2-often and 3-always: the scoring options, therefore, changed their pattern from 01234 to 01122 (0 = never; 1 = rarely or sometimes; 2 = often or always). Categories remained ordered, thresholds became ordered (numeric logit values of thresholds not shown), and all categories emerged ([Figure 3\(B\)](#)).

### The Rasch ruler

Again, let’s suppose we computed the  $\beta$  and  $\delta$  parameters in logit units for the respondents and the 13 items of the motor-FIM questionnaire, respectively (same sample of patients as in [Figure 2](#)). To assess the test functioning, one can start looking at the so-called “Rasch ruler” given in [Figure 4](#).

From bottom to top, items of increasing value are aligned in panel A, like ticks on a ruler, on the right. Ticks are irregularly spaced, like in a ruler where some ticks were cancelled here and there. Some items share the same difficulty level, indicating potential redundancy. Persons are aligned on the left of the ruler. Panel B is zooming (for clarity) on the upper ruler segment (0 logits or higher).

One may capture, at a glance, at least three properties of the scale:

- the matching between average persons’ ability and mean item difficulty (“targeting”).

- The matching between the density of ability and difficulty levels. “Gaps” in the ruler imply a less precise estimate of ability levels (i.e., lower discrimination).
- The spread of the scale. Ticks of the ruler (thresholds) should span the ability levels of most persons in the sample. In outcome studies, the spread of the ticks should also encase their ability changes. For instance, persons admitted to an acute rehabilitation inpatient unit show a spread of FIM measures in the order of 5 to 8 logits and an average increment of about 2 to 3 at discharge.

### The ruler ticks: thresholds, not items

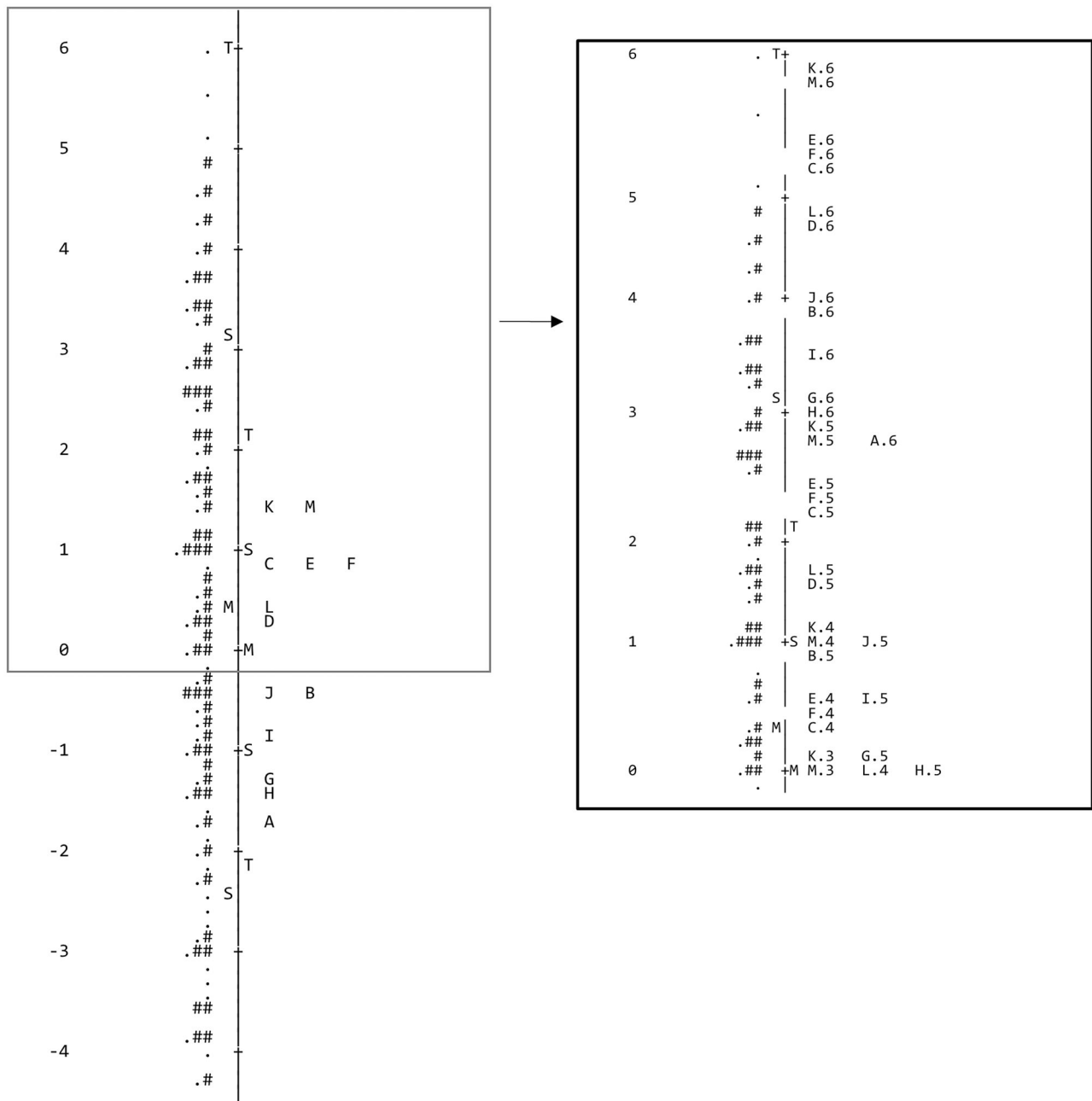
From the right panel of [Figure 4](#), the ruler’s “ticks” are the thresholds from Rasch modelling. Dichotomous items, in which item and threshold “difficulty” are coincident, can be seen as a particular case of polytomous items. Rasch modelling provides thresholds; therefore, dichotomous and polytomous items (also with different categories) may coexist on the same scale.

### Numbers underlying graphics

Many summary numeric indexes are usually shown and vary according to the software adopted. The floor/ceiling effect can be easily captured by the number of “extreme” scores (here, scores 13 or 91), which do not enter the RA for estimating the items’ calibration. Their measure cannot be calculated anywhere beyond the scale boundaries. One person scored 91, and 30 scored 13 (overall, 10.1% of the sample). No item was scored 1 or 7 by the entire sample.

[Table 2](#) provides refined numeric indexes deepening the information given by the graphic ruler.

In A, the spread of ability (for persons) and difficulty levels (for items) are given in standard error (SE) units (“separation”). For the sake of precision, SE, here, is the standard error of measurement, i.e., the standard deviation of a hypothetical distribution of measures centred on their “true” unknown value, a parameter quantifying the uncertainty of a single measurement.



**Figure 4.** The “Rasch ruler.” The linear “ruler” generated by the Rasch analysis of FIM scores from stroke patients discharged from an inpatient rehabilitation unit (first author’s data). The graphic output from Winsteps® Rasch Measurement software (4.4.5, [www.winsteps.com](http://www.winsteps.com)), one of the leading software for running the Rasch analysis, is displayed. Left panel. The vertical line with horizontal, equally spaced dashes represents the independence continuum, i.e., the variable measured by the FIM (lower end: low independence; top end: high independence). Like for rulers’ ticks, leftmost numbers (from -4 to +6) mark the logits’ position along the independence line. The position (i.e., measure) of groups of two (.) and three (#) persons is given along the independence line. Note that dots and hashtags show the frequency distribution of the patients’ measures. The “M” on the left of the independence line gives the person’s mean measure. “S” and “T” give the person’s one and two standard deviations, respectively. Similarly, “M,” “S,” and “T” to the right of the independence line show the mean, one and two standard deviations of the items’ measures. Note that 0 logits correspond to the items’ mean measures conventionally. Letters from A to M on the right report the position of the FIM items along the independence line (i.e., their calibration, “difficulty,” from bottom to top; see Figure 2 for legend). Right panel. This inset focuses on the positive portion of the ruler (from 0 to 6 logit) and gives the thresholds between categories. For instance, H.6 indicates the threshold between scores 6 and 7 in item H, “Bowel management.” A total of 269 persons are represented out of 300. Thirty-one persons were excluded because they received extreme scores (i.e., 13 or 91, see text).

### Understanding Rasch reliability

The general definition of reliability is the ratio of true variance to the observed (i.e., true + error) variance [17]. Stated more simply, it is an index of how much the differences across measures reflect real differences rather than measurement errors. In RA, this type of reliability corresponds to the “separation reliability,” analogous to the Cronbach- $\alpha$  index of internal consistency computed on raw

scores. Cronbach- $\alpha$ , however, may be inflated by a high number of items and by subsets of items correlated within, but not between, subsets [30], thus concealing multidimensionality (a problem discussed later on). The Reliability index allows computing how many “strata” of statistically distinct values can be discerned (means of adjacent “strata” should differ by 3 SE if significance is set at  $p < 0.05$ ). In practice, a Reliability  $> 0.7$  is



Table 2. Numeric indexes from Rasch analysis (RA) of 300 FIM records (see Figures 2 and 4).

|         | A) Summary indexes |             |        | B) Item statistics |             |             |         | C) Statistics of 6 representative persons |                     |             |             |         |          |  |
|---------|--------------------|-------------|--------|--------------------|-------------|-------------|---------|---|---------------------|-------------|-------------|---------|----------|--|
|         | Separation         | Reliability | Strata | Entry label        | Total score | Total count | Measure | Model SE                                  | Entry label         | Total score | Total count | Measure | Model SE |  |
| Persons | 6.3                | 0.98        | 8.7    | K                  | 970         | 300         | 1.46    | 0.08                                      | 448                 | 67          | 13          | 1.07    | 0.37     |  |
| Items   | 12.8               | 0.99        | 17.5   | M                  | 985         | 300         | 1.37    | 0.08                                      | 557                 | 64          | 13          | 0.68    | 0.35     |  |
|         |                    |             |        | E                  | 1060        | 300         | 0.91    | 0.08                                      | 464                 | 62          | 13          | 0.45    | 0.33     |  |
|         |                    |             |        | F                  | 1063        | 300         | 0.89    | 0.08                                      | 493                 | 61          | 13          | 0.34    | 0.33     |  |
|         |                    |             |        | C                  | 1074        | 300         | 0.82    | 0.08                                      | 412                 | 58          | 13          | 0.04    | 0.31     |  |
|         |                    |             |        | L                  | 1146        | 300         | 0.38    | 0.08                                      | 499                 | 56          | 13          | -0.15   | 0.30     |  |
|         |                    |             |        | D                  | 1172        | 300         | 0.22    | 0.08                                      | Extreme persons: 31 |             |             |         |          |  |
|         |                    |             |        | J                  | 1265        | 300         | -0.38   | 0.08                                      |                     |             |             |         |          |  |
|         |                    |             |        | B                  | 1271        | 300         | -0.42   | 0.08                                      |                     |             |             |         |          |  |
|         |                    |             |        | I                  | 1340        | 300         | -0.88   | 0.08                                      |                     |             |             |         |          |  |
|         |                    |             |        | G                  | 1393        | 300         | -1.25   | 0.08                                      |                     |             |             |         |          |  |
|         |                    |             |        | H                  | 1412        | 300         | -1.38   | 0.08                                      |                     |             |             |         |          |  |
|         |                    |             |        | A                  | 1459        | 300         | -1.72   | 0.09                                      |                     |             |             |         |          |  |
|         |                    |             |        |                    |             |             |         | Extreme items: 0                          |                     |             |             |         |          |  |

Notes: Extreme items (or persons): number of items (or persons) discarded from the analysis because they achieved minimal scores, only, or maximal scores, only. In A, the overall Reliability of the modelled scale is given (see text for details). In B, the difficulty measures of the items are shown in descending order from top to bottom. Entry label: order of items in the questionnaire (from A to M, see Figure 2(A) for labelling). Total score = sum of scores obtained by the items across the whole sample of analysed persons. Total count: number of persons tested (including "extreme," see below). Measure = item difficulty estimated through RA (logit units). Model SE: standard error of measurement, estimated by the model. In C, the same information is given (as in B) for six representative persons of intermediate ability levels. Total count = number of items answered (including "extreme" items, see below). Persons are listed in order of descending ability, from top to bottom.

needed to distinguish 2 "strata" of measures [31,32], thus allowing us to reject the hypothesis that all estimates reflect randomness. In the FIM example, persons' reliability is very high (0.98).

Researchers new to RA may wonder how a "Reliability" index can be computed without replicating measurements, the basis for variance estimation. This feature, shared with Cronbach- $\alpha$ , is explained in the [Supplementary Material, Note 3](#).

### The consistency between observed and expected scores: the "data-model fit" issue

A critical distinction must be made between the parameters characterising the model, illustrated so far, and the data set from which the model is "extracted." Given a dataset, Rasch software invariably returns ICCs, CPCs, and a Rasch ruler. In the Authors' experience, the software almost always displays at least 2 "strata" of persons' ability levels. The reader should be critical and ask: do the data at hand justify the model? Single strings of scores or even single scores (assigned from a subject to the series of items or obtained by an item from the series of subjects) can be flagged as "unexpected" ("misfitting," as per the Rasch jargon), given the ability of the subjects and the difficulty of the items. This may be the case for able subjects missing easy items or a difficult item being passed by low-ability subjects. RA is very efficient in highlighting even the most subtle data-model misfit. Conceptually, misfit raises the issue of "content" validity or, stated otherwise, the issue of "what" is measured rather than the issue of "how much" [33]. When misfit affects most items, they are suspected to reflect different variables. As an alternative, the very existence (outside the analyst's mind) of the variable itself must be suspected [34]. This complex issue goes beyond the scope of the present article and will be partially faced in the companion article. In any case, misfit is far from rubbish. Misfit stimulates diagnostic reasoning and generates a challenge: when is misfit "too much"? Should misfitting observations, or even misfitting subjects, or items, be ignored or removed from the analysis? Should the scoring structure of polytomous items be modified? The analyst's choices are partially subjective. The intricate and

crucial issue of the data-model fit will be addressed explicitly in the companion article.

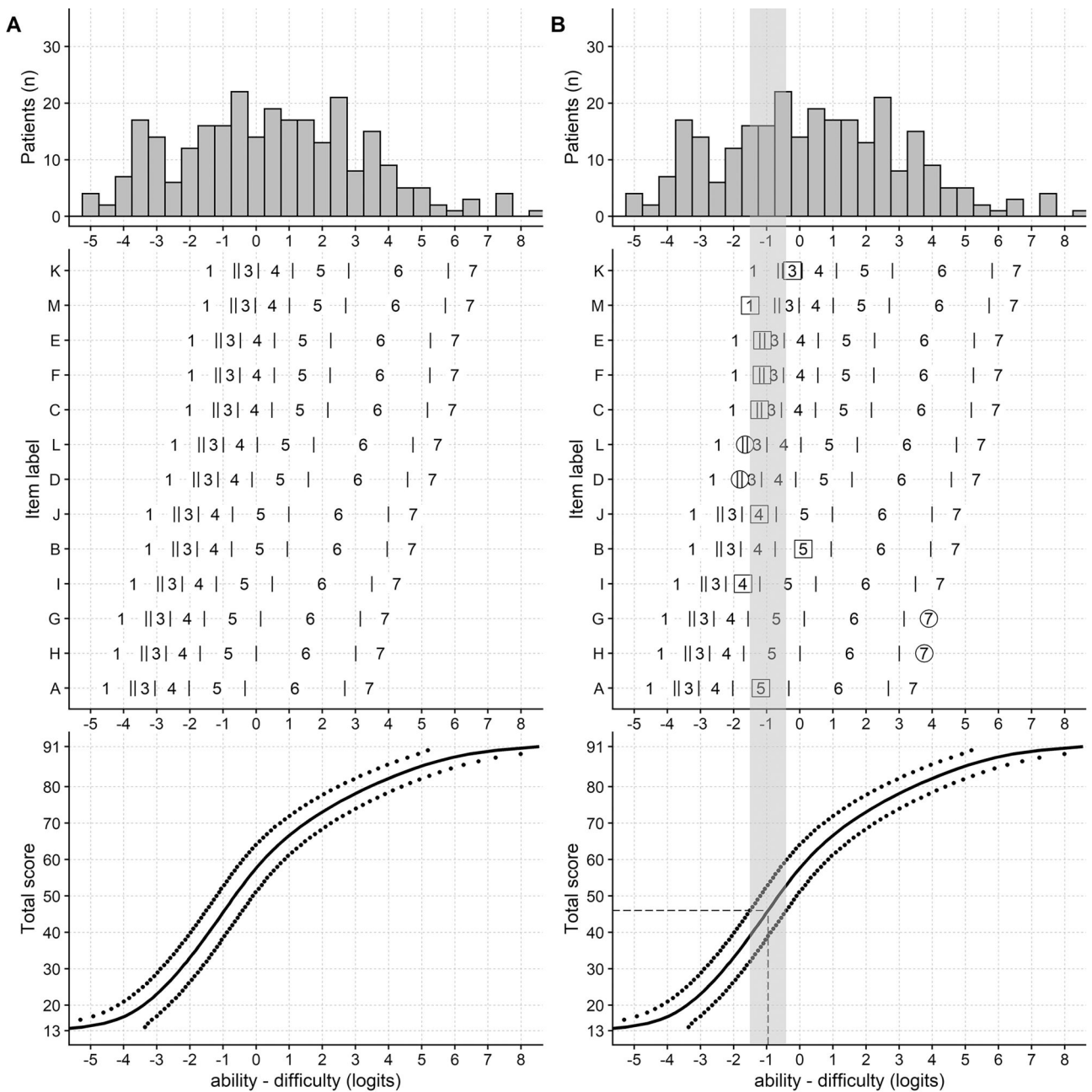
### The "response ogive" and the scoring pattern plot

Figure 5(A) shows another typical graphic output of Rasch software.

The motor-FIM scale is analysed (same sample of patients as in Figures 2 and 4). On the abscissa, the ability-difficulty values are given (logit units). From the bottom to the top panels, the ordinate shows the model-expected total score ( $\pm$  SE) of the 13 FIM items in order of increasing difficulty (bottom to top) and the frequency distribution of patients, respectively. As a note here, in RA the SE corresponds to the standard error of the measurement (not to be confounded with the standard error of the mean, which is also abbreviated with SE). The S-shaped curve highlights the concept that the closer the score is to the minimum or the maximum total score, the less the scores are allowed to differ across persons with different ability levels. Each category is shown over the corresponding logit measure in the middle panel. The vertical bars mark Andrich's thresholds, i.e., the ability levels making adjacent categories equally likely. In this "Andrich's rating scale" model, the distance pattern across categories is the same for all items: the threshold sequence is shifted to the right, to a greater extent the more difficult the item is.

Panel B provides an example of applying this graphic output to the visual analysis of the consistency of responses to the model expectations ("fit") in a single person scoring 46/91 on the motor-FIM scale (range from 13 to 91). From the observed score on the ordinate (dashed horizontal segment), the corresponding Rasch measure can be easily found on the abscissa (here,  $-0.96$  logits, dashed vertical line). The vertical grey band encases the measure  $\pm$  SE (here,  $0.28$  logits). The middle panel allows seeing the most likely individual score for each item. The squares surround the expected scores while the circles surround the unexpected ones (in the companion article, numeric values of "fit" will be given).

The "expected" score can be estimated once the logit measure is known. This may be useful when persons may only respond to



**Figure 5.** Rasch modelling of the expected responses to the items of the motor-FIM scale (see Figures 2 and 4). In panel A, the bottom graph gives the “response ogive,” i.e., the S-shaped function relating the total expected raw score (ranging from 13 to 91, on the ordinate) to the person’s ability level (on the abscissa, logit units). The score-to-measure function (continuous line) is surrounded by the standard error estimates (dotted lines). The histogram in the top panel gives the frequency distribution of abilities across the sample of persons. The middle panel provides each item’s seven response options (categories) as a function of the person’s ability. Items are listed from bottom to top, from easiest to most difficult. On each item, scores may range from 1 to 7. The “|” symbols mark Andrich’s thresholds between adjacent categories. Category 2 is most likely only across a minimal span of ability levels (see Figure 2B), so it could not be visualised between the adjacent “|” symbols. In panel B, this graphic representation is applied to the visual analysis of the “fit” of a person scoring 46 and measuring  $-0.96 (\pm 0.28)$  logits (see text). Panel obtained by combining graphic outputs from R software, 3.6.2.

some of the items, which makes nonsense the total score in CTT. RA can estimate very well the missing responses. This point will be developed in the companion article [1].

**How many items and thresholds?**

*More is not necessarily better*

Keeping stuck to the “metric ruler” metaphor, adding items (and/or thresholds) to an original questionnaire may (1) increase

the spread of the scale if the new “ticks” lower the floor and/or raise the ceiling of the scale or (2) the new “ticks” are nested between the existing ones, thus allowing greater accuracy of measurement or discrimination. However, this is not the case if passing the new thresholds implies the same ability level as passing the existing ones [35,36] (see the overlapping thresholds in Figure 3). The raw score is increased, but not the ability measure. Redundancy is only obtained, burdening the questionnaire (this problem is exemplified in [37], Figure 1, and [8], Figure 3). Redundancy is a double-edged blade. On the one edge, it can

increase the measure's reliability (scoring more behaviours reflecting the same ability level adds to your confidence in the measure). On the opposite edge, adding items and thresholds can bear the risk of introducing extraneous dimensions, increasing the data-model misfit.

### **Less is not necessarily better**

More and more articles are proposing "short forms" of original questionnaires [38]: item categories and whole items are removed. Nevertheless, too few "ticks" along the ruler may increase the uncertainty surrounding the ability estimate (hence, the information provided by the measurement and its generalisability). This trade-off should be carefully considered [39].

### **Three good reasons to prefer Rasch-consistent measures to raw scores**

Striving to fit the Rasch model may seem more like a statistical exercise than an effort of practical utility. Let's summarise a few good reasons which make the effort worth it. These reasons will receive heavier support in the companion article.

1. RA provides equal-interval (linear) units. There is also evidence that Rasch-transformed scores allow a higher accuracy (i.e., higher discrimination) in measuring the subject's ability levels [35,36,40-44].
2. RA can provide robust measures even in case of missing scores.
3. RA can tell you how trustworthy (i.e., "model-expected") the score assigned to a single item is. Unexpected scores may foster diagnostic investigation. For instance, unexpectedness (data-model "misfit") may come from scoring errors, guessing, individual or group peculiarities, etc.

### **Four hints to the lay readers: the figures and tables they should look at, first**

Different software packages may provide different graphic and numeric outputs. This issue is faced in the companion paper. However, all packages offer the same basic information about the scales' properties, seen from the Rasch perspective.

1. The most immediate representation of the scale's functioning is the Rasch "ruler" (Figure 4). According to the intuitive "ruler" metaphor, the spread, the precision, and the targeting of the scale with respect to the sample of persons appear immediately. For polytomous items, all the thresholds (the "ticks" of the ruler) should be provided.
2. The numeric counterpart of the ruler is given here in Table 2, columns A and B. Column A provides the discrimination of the "ruler" in statistically discernible units (here, "separation" and "strata") and its reliability. In essence, data in Column A answer the question: do the logit values of persons' and items' measures differ statistically from their overall, respective means? Significance is an on-off, risky cut-off. Complementing significance indexes, other indexes give the "amount" of spread around the mean. In general, persons' reliability  $> 0.8$  is deemed acceptable. Column B is even more interesting: it provides the items with their measures (the main ruler's ticks) and their SE. Items "fit" to the model is also commonly shown next to their calibration (not reported here; see the companion paper). Bulkier tables can also give the logit values (and SE) of the thresholds, i.e., the minor ruler's ticks. A look at persons' measures may also be enlightening (column C), even if

rarely reported in published papers. The number of people getting "extreme" scores is a valuable index of the scale's targeting. A person's "misfit" may also stimulate diagnostic reasoning (see the companion paper).

3. The Category Probability Curves represent the items' categories functioning for polytomous items (see Figures 2 and 3). The spread of and the relative distancing across categories is made evident, as well as the threshold disordering, if any. The numeric counterpart is given by a table analogous to Table 2, column B, but referring to categories rather than items. These tables provide the average measures of people selecting any given category (from which the categories' order can be checked), their SE, and the values of the thresholds between adjacent categories.
4. Once the scale looks promising, the clinician should ask: can I use this scale even without Rasch software? The "score-to-measure" conversion graph and table should answer this question. The unique requirement is that no scores are missing. This graph is shown here at the bottom of Figure 5. If the total raw score is available, the linear "measure" of the person, and his/her SE, are easily found. Usually, a numeric table accompanies the S-shaped graph to ease this conversion. If the item thresholds map is overlapped (see Figure 5, right column, middle panel), the "fit" of the subject's scores can also be estimated by eye. The "expected" scores of the subject in each item can be calculated through the score-to-measure conversion. The discrepancy between the observed and the expected items' scores strikes the eye, with no need for applying Rasch software.

### **Limiting and widening the scope of the present and the companion articles**

The present article deals only marginally with some issues relevant to the understanding of Rasch modelling and - of more importance here - to the critical appraisal of published Rasch-based articles. Perhaps the main practical problem needing further highlighting is the wide margin of choice left to the Authors in deciding (1) which kind of Rasch model to adopt in the analysis, (2) how to face any disordered category or threshold, and (3) how to face the inevitable data-model misfit. The reader will be further guided through some more Rasch technicalities in the companion article to be more critical about these points.

As a concluding remark, it is worth stressing that becoming more curious about Rasch modelling can be rewarding from a more general, cultural perspective. The Rasch theory and models must be encased within the general framework of measurement theory, a philosophical, not less than a mathematical/physical argument.

The Rasch theory offers a bridge between the measurement in human biology and a person's behaviour in Medicine [45]. The Rasch models pave the way for a common metrological framework between social and physical sciences [46,47].

### **How to learn more**

Beyond articles already cited in the previous paragraphs, further readings (a very concise list of examples) are suggested here. Advanced guidelines for Rasch authors will be mentioned in the companion article.

### Journal articles

The argument of framing Rasch's theory and models within the broader metrology domain is treated in [48].

The "latent-trait" approach is summarised in [13].

The general item-response theory is explained in [10].

A simple introduction to Rasch Analysis can be found in [37,49].

Examples of Rasch articles of methodological nature, yet addressed to health care professionals, are also available [50–52].

Many Rasch articles address various clinical conditions [43,52,53].

Of particular interest here, many articles face the disability assessment and the rehabilitation outcomes [54–59].

### Dedicated journals

These are the *Journal of Applied Measurement* and *Rasch Measurement Transactions*.

### Dedicated web portal

Relevant information can be found at [www.rasch.org](http://www.rasch.org).

### Books

At least the following books should be cited here:

Andrich D. *Rasch models for measurement*. Newbury Park, CA: Sage Publications, 1988.

Andrich D, Marais I. *A course in Rasch Measurement Theory. Measuring in the Educational, Social and Health Sciences*. Singapore: Springer Nature Singapore, 2019.

Bond T, Yan Z, Heene M. *Applying the Rasch model: fundamental measurement in the human sciences*. 4th ed. London/New York: Routledge, 2021.

Boone WJ, Staver JR, Yale ML. *Rasch Analysis in the Human Sciences*. Dordrecht: Springer, 2014.

Bang Christensen K, Kreiner S, Mesbah M (eds). *Rasch Models in Health*. Hoboken, NJ: John Wiley & Sons, Inc., 2013.

### Software

Most relevant software packages now include Rasch routines. Specific software programs are popular within the Rasch community (for a tentative list, see [rasch.org/software.htm](http://rasch.org/software.htm), accessed 12 August 2022). The issue of differences across software packages is dealt with in the companion paper.

### Concluding remarks

Rasch's theory and modelling provided a formal solution to the many problems caused by raw scores and proved empirically to raise the metric properties of questionnaires. Questionnaires represent a hard challenge in measurement theory and practice: persons measure persons, and, with respect to the physical measures, the uncertainty of results is higher, and the sources of uncertainty may be of various kinds. Problems may arise in clinical practice from the need for dedicated software and for the frequent omission of item and thresholds calibrations in the published scales, which are instead necessary for obtaining precise measures from newly collected questionnaires (see the [Supplemental Material, Note 4](#) for details on the "anchoring" procedure). Margins of decisions are (correctly) left to the analyst in extracting and interpreting the results. The companion article will illustrate different Rasch-compliant models and more

sophisticated metric applications. Hints to avoid misuse of the models will be proposed.

### Acknowledgements

The authors are indebted to Dr Franco P. Franchignoni and the anonymous reviewers for helpful criticisms of the manuscript.

### Disclosure statement

The authors declare that the research was conducted without any commercial or financial relationships construed as a potential conflict of interest.

### Funding

The Italian Ministry of Health supported the study, Ricerca Corrente 2021 (RESET project) and the Italian Foundation for Multiple Sclerosis (FISM), CORESETS (PROMOPRO-MS) project, 2012.

### ORCID

Luigi Tesio  <http://orcid.org/0000-0003-0980-1587>

Antonio Caronni  <http://orcid.org/0000-0003-3051-1031>

Dinesh Kumbhare  <http://orcid.org/0000-0003-3889-7557>

Stefano Scarano  <http://orcid.org/0000-0002-9217-4117>

### References

- [1] 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*. DOI: [10.1080/09638288.2023.2169772](https://doi.org/10.1080/09638288.2023.2169772)
- [2] Becker KA. History of the Stanford-Binet intelligence scales: content and psychometrics. In: *Stanford-Binet intelligence scales*. Itasca (IL): Riverside; 2003.
- [3] Porter ME, Stefan L, Lee TH. Standardizing patient outcomes measurement. *N Engl J Med*. 2016;374(6):504–506.
- [4] Andrich D. Understanding resistance to the data-model relationship in Rasch's paradigm: a reflection for the next generation. *J Appl Meas*. 2002;3:325–359.
- [5] van de Winckel A, Kozłowski AJ, Johnston M, et al. Reporting Guideline for RULER: Rasch reporting guideline for rehabilitation research: explanation and elaboration. *Arch Phys Med Rehabil*. 2022;103(7):1487–1498. v
- [6] Smith RM, Linacre JM, Smith EVJ. Guidelines for manuscripts. *J Appl Meas*. 2003;4:198–204.
- [7] Mallinson T, Kozłowski AJ, Johnston M V, et al. Rasch Reporting guideline for rehabilitation research (RULER): the RULER statement. *Arch Phys Med Rehabil*. 2022;103(7):1477–1486.
- [8] Tesio L, Scarano S, Hassan S, et al. Why questionnaire scores are not measures: a question-raising article. *Am J Phys Med Rehabil*. 2023;102(1):75–82.
- [9] Bock RD. A Brief history of item response theory. *Educ Meas Issues Pract*. 1997;16:21–33.
- [10] Carlson JE, von Davier M. Item response theory. In: Bennett R, von Davier M, editors. *Item response theory. Advancing human assessment. Methodology of educational measurement and assessment*. Princeton (NJ): Springer; 2017. p. 133–178.
- [11] Reise SP, Ainsworth AT, Haviland MG. Item response theory. *Fundamentals, applications, and promise in psychological research*. *Curr Dir Psychol Sci*. 2005;14:95–101.

- [12] Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: time to end malpractice? *J Rehabil Med*. 2012;44(2):97–98.
- [13] Hambleton RK, Cook LL. Latent trait models and their use in the analysis of educational test data. *J Educ Meas*. 1977;14:75–96.
- [14] Borsboom D, Mellenbergh GJ, van Heerden J. The theoretical status of latent variables. *Psychol Rev*. 2003;110(2):203–219.
- [15] Lord FM. The relation of test score to the trait underlying the test. *Educ Psychol Meas*. 1953;13:517–548.
- [16] Borsboom D. Latent variable theory. *Meas Interdiscip Res Perspect*. 2008;6:25–53.
- [17] Tesio L. Outcome measurement in behavioural sciences: a view on how to shift attention from means to individuals and why. *Int J Rehabil Res*. 2012;35(1):1–12.
- [18] Frigg R, Hartmann S. Models in science. In: Zalta EN, editor. *The Stanford Encyclopedia of philosophy*. [accessed 2022 Sept 14]. Available from: <https://plato.stanford.edu/archives/spr2020/entries/models-science/>
- [19] Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago (IL): University of Chicago Press; 1980.
- [20] Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. *Arch Phys Med Rehabil*. 1989;70(12):857–860.
- [21] Aryadoust V, Tan HAH, Ng LY. A scientometric review of Rasch measurement: the rise and progress of a specialty. *Front Psychol*. 2019;10:2197.
- [22] Andrich D. An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any ‘threshold disorder controversy. *Educ Psychol Meas*. 2013;73:78–124.
- [23] Barber B. Resistance by scientists to scientific discovery. *Science* (1979). 1961;134:596–602.
- [24] Rasch G. On General laws and the meaning of measurement in psychology. In: Neyman J, editor. *Berkeley symposium on mathematical statistics and probability*. Vol. 4.4. Berkeley (CA): University of California Press; 1961. p. 321–333.
- [25] Andrich D. Rasch models for measurement. Newbury Park (CA): Sage Publications; 1988.
- [26] Tesio L, Granger CV, Perucca L, et al. The Fim™ instrument in the United States and Italy: a comparative study. *Am J Phys Med Rehabil*. 2002;81:168–176.
- [27] Linacre JM, Heinemann AW, Wright BD, et al. The structure and stability of the functional independence measure. *Arch Phys Med Rehabil*. 1994;75(2):127–132.
- [28] Gothwal VK, Bharani S, Kekunnaya R, et al. Measuring health-related quality of life in strabismus: a modification of the Adult Strabismus-20 (as-20) questionnaire using Rasch analysis. *PLoS One*. 2015;10:1–19.
- [29] Adams RJ, Wu ML, Wilson M. The Rasch rating model and the disordered threshold controversy. *Educ Psychol Meas*. 2012;72:547–573.
- [30] Agbo AA. Cronbach’s Alpha: review of limitations and associated recommendations. *J Psychol Afr*. 2010;20:233–239.
- [31] W F Jr. Reliability, separation, strata statistics. *Rasch Meas Trans*. 1992;6:238.
- [32] Wright BD. Reliability and separation. *Rasch Meas Trans*. 1996;9:472.
- [33] Morel T, Cano SJ. Measuring what matters to rare disease patients - reflections on the work by the IRDiRC taskforce on patient-centered outcome measures. *Orphanet J Rare Dis*. 2017;12(1):171.
- [34] Stenner AJ, Stone MH, Burdick DS. Indexing vs. Measuring Rasch-related coming events. *Rasch Meas Trans*. 2009;22:1176–1177.
- [35] Tesio L, Cantagallo A. The functional assessment measure (FAM) in closed traumatic brain injury outpatients: a Rasch-based psychometric study. *J Outcome Meas*. 1998;2(2):79–96.
- [36] Linn R, Blair R, Granger C, et al. Does the functional assessment measure (FAM) extend the functional independence measure (FIM) instrument? A Rasch analysis of stroke inpatients. *J Outcome Meas*. 1999;3(4):339–359.
- [37] Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *J Rehabil Med*. 2003;35(3):105–115.
- [38] Hsieh YW, Hsueh IP, Chou YT, et al. Development and validation of a short form of the Fugl-Meyer motor scale in patients with stroke. *Stroke*. 2007;38:3052–3054.
- [39] Monticone M, Giordano A, Franchignoni F. Scale shortening and decrease in measurement precision: analysis of the pain Self-Efficacy questionnaire and its short forms in an Italian-Speaking population with neck pain disorders. *Phys Ther*. 2021;101:1–10.
- [40] Turner-Stokes L, Medvedev ON, Siegert RJ. Rasch analysis of the UK functional assessment measure in a sample of patients with traumatic brain injury from the UK national clinical database. *J Rehabil Med*. 2019;51:566–574.
- [41] Petrillo J, Cano SJ, McLeod LD, et al. Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. *Value Health*. 2015;18(1):25–34.
- [42] Kersten P, White PJ, Tennant A. Is the pain visual analogue scale linear and responsive to change? An exploration using Rasch analysis. *PLoS One*. 2014;9(6):e99485.
- [43] Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess*. 2009;13(12):iii, ix–x, 1–177.
- [44] Baker K, Barrett L, Playford ED, et al. Measuring arm function early after stroke: is the DASH good enough? *J Neurol Neurosurg Psychiatry*. 2016;87(6):604–610.
- [45] Tesio L. Measuring in clinical vs. biological medicine: the Rasch model as a bridge on a widening gap. *J Appl Meas*. 2004;5(4):362–366.
- [46] Pendrill L. Man as a measurement instrument. *J Meas Sci*. 2014;9:24–35.
- [47] Pendrill L. Quality assured measurement: unification across social and physical sciences. Cham: SpringerLink; 2019.
- [48] Pendrill LR. Assuring measurement quality in person-centred healthcare. *Meas Sci Technol*. 2018;29:034003.
- [49] Kersten P, Kayes NM. Outcome measurement and the use of Rasch analysis, a statistics-free introduction. *NZ J Physiother*. 2011;39:92–100.
- [50] Hobart JC, Cano SJ, Zajicek JP, et al. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet Neurol*. 2007;6(12):1094–1105.
- [51] Vanhoutte EK, Hermans MCE, Faber CG, et al. Rasch-ionale for neurologists. *J Peripher Nerv Syst*. 2015;20(3):260–268.
- [52] Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care Res*. 2007;57:1358–1362.

- [53] Caronni A, Sciumè L, Donzelli S, et al. ISYQOL: a Rasch-consistent questionnaire for measuring health-related quality of life in adolescents with spinal deformities. *Spine J.* 2017; 17(9):1364–1372.
- [54] Baylor C, Hula W, Donovan NJ, et al. An Introduction to item response theory and Rasch models for speech-language pathologists. *Am J Speech Lang Pathol.* 2011;20(3): 243–259.
- [55] Tesio L, Valsecchi MR, Sala M, et al. Level of activity in profound/severe mental retardation (LAPMER): a Rasch-derived scale of disability. *J Appl Meas.* 2002;3(1):50–84.
- [56] Franchignoni F, Horak F, Godi M, et al. Using psychometric techniques to improve the balance evaluation systems test: the mini-BESTest. *J Rehabil Med.* 2010; 42(4):323–331.
- [57] Penta M, Tesio L, Arnould C, et al. The ABILHAND questionnaire as a measure of manual ability in chronic stroke patients: Rasch-based validation and relationship to upper limb impairment. *Stroke.* 2001;32(7):1627–1634.
- [58] la Porta F, Franceschini M, Caselli S, et al. Unified Balance scale: an activity-based, bed to community, and aetiology-in dependent measure of balance calibrated with Rasch analysis. *J Rehabil Med.* 2011;43:435–444.
- [59] Khoury G e, Barbier O, Libouton X, et al. Manual ability in hand surgery patients: validation of the ABILHAND scale in four diagnostic groups. *PLoS One.* 2020;15:1–17.