



Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment

Luigi Tesio, Antonio Caronni, Anna Simone, Dinesh Kumbhare & Stefano Scarano

To cite this article: Luigi Tesio, Antonio Caronni, Anna Simone, Dinesh Kumbhare & Stefano Scarano (2023): Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment, *Disability and Rehabilitation*, DOI: [10.1080/09638288.2023.2169772](https://doi.org/10.1080/09638288.2023.2169772)

To link to this article: <https://doi.org/10.1080/09638288.2023.2169772>



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



[View supplementary material](#)



Published online: 06 Feb 2023.



[Submit your article to this journal](#)



[View related articles](#)



[View Crossmark data](#)

Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment

Luigi Tesio^{a,b} , Antonio Caronni^b , Anna Simone^b , Dinesh Kumbhare^{c,d}  and Stefano Scarano^{a,b} 

^aDepartment of Biomedical Sciences for Health, University of Milan, Milan, Italy; ^bIRCCS, Istituto Auxologico Italiano, Department of Neurorehabilitation Sciences, Ospedale San Luca, Milan, Italy; ^cDivision of Physical Medicine and Rehabilitation, Department of Medicine, University of Toronto, Toronto, Ontario, Canada; ^dPain Research Institute, Toronto Rehabilitation Institute, University Health Network, Toronto, Canada

ABSTRACT

Purpose: The present paper presents developments and advanced practical applications of Rasch's theory and statistical analysis to construct questionnaires for measuring a person's traits. The flaws of questionnaires providing raw scores are well known. Scores only approximate objective, linear measures. The Rasch Analysis allows you to turn raw scores into measures with an error estimate, satisfying fundamental measurement axioms (e.g., unidimensionality, linearity, generalizability). A previous companion article illustrated the most frequent graphic and numeric representations of results obtained through Rasch Analysis. A more advanced description of the method is presented here.

Conclusions: Measures obtained through Rasch Analysis may foster the advancement of the scientific assessment of behaviours, perceptions, skills, attitudes, and knowledge so frequently faced in Physical and Rehabilitation Medicine, not less than in social and educational sciences. Furthermore, suggestions are given on interpreting and managing the inevitable discrepancies between observed scores and ideal measures (data-model "misfit"). Finally, twelve practical take-home messages for appraising published results are provided.

ARTICLE HISTORY

Received 26 February 2022
Revised 10 November 2022
Accepted 10 January 2023

KEYWORDS

Rasch analysis; data-model misfit; Rasch model advanced applications; critical interpretation; latent variables; psychometrics; neurorehabilitation; metrology

► IMPLICATIONS FOR REHABILITATION


- The current work is the second of two papers addressed to rehabilitation clinicians looking for an in-depth introduction to the Rasch analysis.
- The first paper illustrates the most common results reported in published papers presenting the Rasch analysis of questionnaires.
- The present article illustrates more advanced applications of the Rasch analysis, also frequently found in publications.
- Twelve take-home messages are given for a critical appraisal of the results.

Questionnaires are very common in Medicine, including Physical and Rehabilitation Medicine, not less than in social and educational sciences. They are made by lists of observations (items), each graded on ordinal scores, which are summed to provide cumulative scores. Questionnaires are necessary to estimate a whole person's attributes like perceptions, skills, knowledge, and the like ("abilities") [1]. Unfortunately, at best, cumulative raw scores approximate true linear measures [2]. A formal solution to measurement is provided by Rasch Analysis (RA). RA is nowadays an umbrella term encompassing a theory on measurement, statistical models, and a series of related algebraic techniques [3]. The present article complements a companion article [3] aiming at helping clinicians understand the most common published results based on RA of questionnaires.

RA, named after the Danish mathematician Georg Rasch [4], is based on a statistical model that transforms ordinal scores from a

questionnaire's items into true linear measures surrounded by error estimates [5]. The Classical Test Theory (CTT) assumes that scores represent measures proportional to the amount of the variable they are purported to represent. This assumption is optimistic to the least and may lead to misleading results, as exemplified elsewhere [2,6]. The previous companion article [3] introduces the reader to Rasch's statistical modelling, proposed as the best solution to overcome these limitations. RA renders the "most likely" measures of a person's ability and items' difficulty that can be "extracted" from questionnaire data. But how much do the data at hand justify the new measures? The present article illustrates more advanced applications of RA, frequently found in publications, and faces the data-model consistency (fit) issue. Some repetitions between the two companion papers were inevitable, but they were kept to a minimum.

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

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/09638288.2023.2169772>.

© 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.

From raw data to ideal/modelled data

The original Rasch model prescribes which will be the probability of a response to a dichotomous item with a given “difficulty” (e.g., a “yes” vs “no” answer, “pass” vs “fail,” 1 vs 0) for a person of a given “ability.” What the model “expects,” if only 1 = pass and 0 = fail can be observed, is summarised by a rather simple equation:

$$P(X = 1|0, 1) = e^{\beta - \delta} / (1 + e^{\beta - \delta}) \quad (1)$$

The equation is read as: “the probability (P) that response X is observed to be 1, given that (| symbol) X may be either 0 or 1, depends on subject’s ability (β) and item’s difficulty (δ), according to the relationship...” (see the right side of the equation). The term “e” is the base of natural logarithms (2.718...) [1].

An invariant measurement unit is needed for linear measurement. After simple algebraic transformation, Equation (1) can be rewritten as

$$\ln(P/(1 - P)) = \beta - \delta \quad (2)$$

where “ln” stands for natural logarithm and $(1 - P)$ is the “fail” probability. The linear nature of the unit, i.e., one of a difference, becomes evident. The quantity $P/(1 - P)$ is the familiar odd ratio (the win/lose ratio bookies apply to bets). Its natural log (ln) is the invariant measurement unit called “logit.” The reason for resorting to logit units rather than proportions is the need to overcome the nonlinearity of probabilities bound from 0 to 1 (see the [Supplemental Material, Note 1](#)). The Rasch “separability” theorem demonstrates that the formulation given in Equation (1) is the only one estimating ability independently from difficulty. This property respects a fundamental measurement axiom: the measure of the persons, like measures of length or time, must be independent of the particular instrument (ruler, clock, thermometer) adopted for measurement. Conversely, the measure provided by an instrument (length, time, temperature) must be independent of the object measured [7].

The magic β and δ : where do they pop up from?

The β and δ symbols refer to the amount of a person’s property (“ability” as per the Rasch jargon) and the intrinsic difficulty of the item, respectively. The difficulty level of the item is called “threshold” when the lower expected score in one item is equally likely than the next higher score. “Zero” logits mean the two alternative outcomes (e.g., 0/1, fail/pass) share the same probability of being observed. Suppose the person has a 50% (or 0.5) probability of passing or failing: $\ln(0.5/0.5) = \ln(1) = 0$. In this case, a person’s ability equals the item’s difficulty ($\beta - \delta = 0$). Negative logits mean pass probabilities < 50%. Logits discourage most clinicians. However, it is sufficient to assimilate them to Celsius degrees of temperature (also foreseeing negative values below a conventional “zero”). The parameters are “extracted” from the matrix (items by persons) of raw scores in logit units. The basic algebraic principles are shown in the [Supplemental Material, Note 2](#).

It must be recalled that the statistical process of “extraction” of the β and the δ parameters aims to provide the parameters “most likely” fitting the Rasch model based on the empirical data. The properties of the model were highlighted in the companion paper [3]. These include unidimensionality and local independence (responses should be correlated only by the unique, shared unidimensional variable). Violations of these strict model requirements are inevitably found in real data and will be discussed further.

Estimating missing scores: the future of questionnaires

a) Incidental missing responses

Scores may be missing for various reasons: the rater’s carelessness, the subject’s unwillingness to respond, limited time for answering a whole set of questions, etc. The most straightforward remedy (see [2]) is assigning the missing score in one item the mean value of non-missing scores. This procedure assumes that all items share the same difficulty. Rasch modelling needs not this unrealistic assumption. Ability and difficulty are jointly estimated from the observed scores. Then, missing scores are estimated, considering who missed which item. A subject’s “pass” response to an easy item is more likely if the same person gave several “pass” responses to more difficult items. Of course, the higher the number of missing responses, the higher the measures’ error estimate.

b) Scales foreseeing missing responses

More and more frequently, Rasch-compliant questionnaires explicitly foresee that only some items should be administered to, or answered by, all subjects. There may be a-priori reasons for not requiring responses to some items by some persons. For instance, in the ABILHAND questionnaire, persons are asked to rate the perceived difficulty of completing only the manual activities they attempted during the last three months [8]. In cross-cultural studies, some items may not have the exact semantic counterpart (see the case for the “aching” items in pain questionnaires [9]) or the same quantitative meaning in different cultures [10] nor to single persons [11]. Also, different subjects may be requested to select some items based on their familiarity [12].

Item banking: the future of a person’s measurement

a) Item banking: the way to computerised adaptive testing

The problem motivating the solution of item banking is straightforward. There is no point in repeatedly proposing elementary items to very able persons and presenting challenging items to persons of low ability. Their responses will be too predictable, hence uninformative. Also, their responses will push the total score towards the ceiling or the floor of the scale, thus abating its discriminative power [2]. Why not deposit calibrated items to be “tailored” to the subject’s ability?

The heaviest work of “banking” consists in building a bulky scale in advance, in which items and thresholds cover a wide span of difficulty, and persons cover a wide span of ability. The “density” of difficulty values of items/thresholds (like ticks on a metric ruler) must be high. After that, in subsequent studies, the analyst can draw items on this “bank.” The person’s ability is probed through adequate software by administering a few items with very distant difficulty levels. If the person passes the easier ones and misses the more difficult ones, an algorithm selects items with difficulty progressively closer to the person’s ability (responses become less deterministic). The final scale is then targeted to the specific person, maximising information with a few tailored items [13,14]. Banks should be dynamic. New items can be added, and the item calibrations can be updated through the conflation of new assessments. Building online sharable items banks, rather than inventing new questionnaires for each new study, is a challenge worth facing in the future of measurement.

b) Item banking on raters: the future of multi-examiners assessments

Evaluations by a panel of multiple raters are widespread in education, clinical research, sports, and project funding. Parameters of raters’ “severity” and estimates of their “fit” can enter Rasch modelling as a linear modifier of a person’s ability measure (the more

severe the rater, the higher the person's ability estimate) [15] (see below, the paragraph on Many-facets Rasch modelling). Raters' severity and fit can give rise to a "raters' bank." Centralised item and rater banks seem quite exceptional in Rehabilitation Medicine. The only such bank known to the Authors is the one subtending the Assessment of Motor and Process Skills (AMPS), realised by Anne Fisher. Regardless of their workplace and country, raters follow a credentialing course and are credited by calibrating their severity and fit, which is periodically updated. This allows them to provide scores adjusted for the rater's severity and fit comparable across different facilities and cultural contexts [12,16].

Testing data-model fit

Consistent with fundamental measurement axioms, RA prescribes that measures be unidimensional, i.e., RA only applies the "less-to-more" concept to one variable. The Rasch model is axiomatic and, therefore, "prescriptive" (of how observed scores should be) rather than "descriptive" for the sample at hand. This characteristic is perhaps the origin of the reluctance of the scientific community to accept the model itself [17]. No factual data will ever conform to the model; nevertheless, it remains true that data should approximate the model, not the reverse.

From an algebraic standpoint, the Rasch model for dichotomous items is the same as the one-parameter logistic model of the Item Response Theory (IRT) [18,19]. However, their conceptual differences are numerous and profound to the point that, not without reason, authoritative scholars consider the Rasch model as extraneous to the IRT family [20–22]. The main feature that makes the Rasch analysis substantially different from the IRT is precisely its being "prescriptive." In conventional statistical modelling, the analyst strives to find the best model for summarising the data. In Rasch analysis, the analyst checks if the data sufficiently meet the expectations of the Rasch model. If this is not the case, data undergo a critical analysis and are discarded if needed (data should fit the model and not vice versa).

Data-model misfit as a matter of how much

The algorithms of RA give back the "most likely" unidimensional scale subtending the observed scores. Thanks to the Rasch measures of persons and items, the expected scores are those most complying with the axioms: they are said to be the "most likely." However, most likely does not necessarily mean very, nor sufficiently likely. How much "abstraction" from observed scores had to be applied to satisfy the axioms? Stated otherwise, how much do the observed scores differ from (i.e., they "misfit") the scores expected by the model? Once the data-model misfit is given a quantitative estimate, a rater's value judgement is unavoidable. Given the research purposes, is the data-model misfit acceptable? Too large a misfit suggests that the raw scores cannot be taken as decent proxies of measures and that, in turn, the model-expected (estimated) scores will hardly reflect the real world.

Residuals: the constituents of data-model misfit

"Residuals" come from the differences between observed and expected pass probabilities. Contrary to the deterministic Guttman model, the Rasch model always foresees some residuals because the "expected" score is a probability, never achieving 0 or 1. The probability error (i.e., its variance and standard deviation) can be estimated as per classical statistics.

Residuals: model-expected and empirically observed

The Rasch model does not predict only ideal measures, but, given its probabilistic nature, it also predicts a certain amount of variance (hence, some residuals) around each of the measures themselves [23,24]. In short, one part of these residuals is model-expected, and another is not. Misfit comes from the unexpected part. In most applications, residuals are standardised by their modelled standard deviation [25].

How much misfit is too much?

Indexes of "fit" (see below) summarise the accumulation of residuals (e.g., for an item, across responses from persons). This accumulation follows a chi-square distribution. As discussed above, a researcher's decision remains whether misfit is acceptable. Making this decision less arbitrary is advisable, yet it is still an open issue. In particular, it depends on the perspective adopted, i.e., one privileging the size of the chi-square (*how large this size is*) or one privileging statistical significance (*how unexpected this size is*) (see below). Different authors discuss and advocate different goodness of fit statistics (mostly chi-square and t-statistics [26–29]). In many articles, the squares of the observed-expected score difference are z-standardised (see below) [25].

Misfit indexes: far from a supreme judgement

As a rule, Rasch's articles prioritise the evidence for "misfit" indexes remaining within accepted statistical limits. However, it must be highlighted that "fit" does not have the final say. The model is axiomatic, whereas the error is empirical. The researcher applying the model requirements (e.g., unidimensionality, independence of person and item measures, categories' ordering) obtains the "most likely" parameters of ability and difficulty, which are objective in the sense that they are independent of each other. By contrast, the model-expected and the unexpected error surrounding these estimates depend on the data, particularly sample size. As a consequence, Rasch results do not necessarily provide "good enough" persons' measures or a "good enough" scale, no matter how fitting the observed data are (or are made after various manipulations). Regardless of the size of "misfit," the "most likely" parameters offered by RA must also make sense from an external (not only from a purely algebraic) perspective. For instance, does the order of items' difficulty levels and patients' ability levels match clinical expectations? Are the spread and the density of the thresholds of difficulty levels sufficient to cover with decent accuracy the range of ability levels encountered in clinical practice?

Forms of data-model fit

The fit of single responses

The example in Table 1(B) shows a representative series of response strings in which the standardised residuals are given for every response.

Subjects come from a sample of 300 stroke patients discharged from an inpatient rehabilitation unit. The FIMTM-Functional Independence Measure was administered. This is a well-known questionnaire scoring independence in 18 activities of daily living [30]. Each item can be scored 1 to 7: the higher, the greater the person's independence (Table 1(A)). The 13 items, labelled A to M, making the "motor" sub-scale, were only considered here.

Table 1. (A) The FIM™-Functional Independence Measure scale of independence in daily life and (B) the fit of individual responses to the Rasch model.

(A)						(B)													
FIM™- Functional Independence Measure						A	B	C	D	E	F	G	H	I	J	K	L	M	Total score
Self-care	Communication																		
A. Eating	N. Comprehension	prs 1	OUTFIT	5.8	Observed	6	6	3	3	3	3	1	1	5	5	5	5	5	51
B. Grooming	O. Expression		INFIT	4.2	Z-residual							-5	-5			2		2	
C. Bathing		prs 2	OUTFIT	4.3	Observed	6	5	5	5	5	5	1	6	5	5	5	5	5	63
D. Dressing-upper body	Social cognition		INFIT	3.0	Z-residual							-7							
E. Dressing-lower body	P. Social interaction	prs 3	OUTFIT	4.2	Observed	4	2	2	3	3	3	1	1	5	5	3	5	4	41
F. Toileting	Q. Problem solving		INFIT	3.6	Z-residual							-3	-4					2	
	R. Memory	prs 4	OUTFIT	4.1	Observed	4	2	2	2	2	1	1	1	4	4	4	1	1	29
			INFIT	2.6	Z-residual							-2	-2			6			
Sphincter control	Scoring levels	prs 5	OUTFIT	3.6	Observed	5	4	3	2	2	2	5	1	5	5	5	5	5	49
G. Bladder management	7. Complete independence		INFIT	3.1	Z-residual								-5			2		2	
H. Bowel management	6. Modified Independence																		
	5. Supervision																		
Mobility / transfer	4. Minimal assistance																		
I. Bed-chair-wheelchair	3. Moderate assistance																		
J. Toilet	2. Maximal assistance																		
K. Tub-shower	1. Total assistance																		
Locomotion																			
L. Walk-wheelchair																			
M. Stairs																			

Notes: The 13 items (from A to M) of the “motor” FIM sub-scale were analysed in 300 stroke patients at discharge from an inpatient rehabilitation unit (first author’s data). Two hundred sixty-nine non-extreme scores (i.e., scores between 14 and 90) entered the modelling algorithms. Results from the five persons (prs, from 1 to 5) with the worst fit to the model (OUTFIT decreasing from top to bottom) are reported. OUTFIT: outliers sensitive mean square fit. INFIT: inliers sensitive mean square fit. Observed: observed score on each of the 13 items. Z-residual: z-standardised residuals; negative residuals indicate that the observed score is smaller than expected. Z-residuals are truncated to their integer value for graphic reasons and Z-residuals > -2 and < 2 (i.e., non-significant at $p=0.05$) are not reported.

Notably, “extreme” scores (here, 13 or 91) are not considered in data modelling. In these cases, the person’s ability lies at an unknown distance outside the raw scores’ artefactual boundaries. One might always devise items easier or more difficult than the items with extreme scores in the questionnaire. A “zero” score in all items, including the easiest one, does not indicate the absence of the variable.

On non-extreme scores, “residuals” are standardised regarding the expected score standard deviation. The z-standardised residuals become significant (at $p < 0.05$) when they are lower than -1.96 or higher than 1.96 (usually rounded here to -2 and 2). It can be seen, for instance, that person 1 gave unexpected responses to four items. Given the overall ability level of that person, the score observed was significantly higher than expected in items K and M (residuals = 2) and significantly lower than expected in items G and H (residuals = -5). Person 2 gave only one unexpected response (score 1 to item G, while all other items are scored 5 or more). Diagnosing why single persons responded unexpectedly to those items may be interesting.

The fit of the string of responses (persons across items; items across persons)

For persons, the sum of residuals has a chi-square distribution, thus providing summary indexes of fit of the entire response string (outfit and infit, explained below). The same logic can be applied to items. For instance, in Table 1, some items received misfitting responses from four of the five persons, while others never evoked misfitting responses.

Interpreting person fit

RA is focussed on how the total score is achieved, not less than on the size of the total score. Examples of hypothetical individual response patterns, and a person’s fit indexes, are given in Table 2 [31].

A qualitative “diagnostic” interpretation of these patterns is possible. In theory, the same kind of representation could be done for items, but the high number of persons to be aligned

(from less to more able) makes this diagnosis unpractical. The outfit and infit indexes summarise the amount of unexpectedness of the response strings (see below and Table 1(B)).

Quantitative assessment of item fit

Table 3 replicates some of the information in Table 2 of the companion article.

The main Rasch parameters of both the 13 motor-FIM™ items [30] (Table 3(A)) and a sample of six persons of intermediate ability (Table 3(B)) are given. The fit indexes have been added.

In Table 3(A), the leftmost column gives the entry order of the 13 items of the motor-FIM scale (see Table 1(A)). The second column from the left provides the items with difficulty (in logit units) and their standard errors, corresponding to the CTT’s standard error of measurement (see the companion paper). Four fit indexes are provided: “Infit” and “Outfit,” each expressed as mean square (MNSQ) and z-standardised (ZSTD) units (see below). From top to bottom, items are listed in order of decreasing fit (ZSTD column).

Table 3(B) replicates the information provided by Table 3(A) for an illustrative sample of persons of intermediate ability within the sample at hand. The leftmost column gives the persons’ labels. Usually, the long person table is not provided in published results.

What MNSQ (mean-square) and ZSTD (z-standardised) indexes tell

One may wonder why four different indexes of data-model “fit” can be found in several Rasch publications. Let’s start with the difference between MNSQ and ZSTD. The debate on the pros and cons of either index is still open and rather complex [32].

A premise is worthwhile. Fit can be “acceptable” if observations are close enough to the model expectations. In most studies, the “enough” judgement is usually based on statistical significance. As already outlined, residuals should be neither larger nor smaller than expected. In the Rasch jargon, “underfit” and “overfit” refer to these opposite conditions. Underfitting is more troublesome than overfitting, which is usually not worrying,

Table 2. Representative individual response patterns to 11 dichotomous (fail/pass, 0/1) items.

Person responses:	Diagnosis	Outfit	Infit
easy – items – hard	pattern	Mean-square	Mean-square
111 0110110100 000	Modelled/ideal	1	1.1
111 111110000 000	Guttman/deterministic	0.3	0.5
000 000001111 111	Miscode	12.6	4.3
011 1111110000 000	Carelessness sleeping	3.8	1
111 111100000 001	Lucky guessing	3.8	1
111 01010101 000	Low discrimination	1.5	1.6

Notes: Items are listed in order of increasing difficulty, from left to right. Data should be considered anecdotal. The vertical segments delimit the items with scores encasing the person's ability, i.e., where the Rasch model expects some random alternation of "pass" and "fail" (top row of numbers). A Guttman-deterministic, over-fitting pattern (second row of numbers from top) is also dubbed "too good to be true" (fit indexes much lower than 1). All other patterns reveal "too unexpected" responses from the person and show fit indexes much greater than 1. Bold numbers highlight the unexpected sequences of scores (leftmost column) and the mean square values indicating "misfit" (two rightmost columns) (after [31], modified, with permission).

Table 3. Parameters of 13 motor-FIM items (A) and six representative persons (B) were obtained through Rasch analysis (Winsteps[®] software, Rating scale model).

A)							B)						
Item	Meas.	Model SE	Infit		Outfit		prs	Meas.	Model SE	Infit		Outfit	
			MNSQ	ZSTD	MNSQ	ZSTD				MNSQ	ZSTD	MNSQ	ZSTD
G	-1.23	0.08	2.06	8.29	1.72	5.29	153	-0.40	0.29	1.12	0.43	1.11	0.41
H	-1.36	0.08	2.02	8.05	1.73	5.23	186	-0.32	0.29	1.05	0.26	1.15	0.49
L	0.37	0.08	1.61	5.22	1.41	3.08	289	-0.24	0.29	0.53	-1.31	0.56	-1.22
K	1.43	0.08	1.17	1.63	1.20	1.22	199	-0.15	0.30	1.25	0.71	1.29	0.80
C	0.81	0.08	0.97	-0.29	0.97	-0.19	134	-0.06	0.30	0.53	-1.28	0.53	-1.29
A	-1.69	0.08	1.03	0.32	0.94	-0.44	112	0.03	0.31	1.86	1.81	2.11	2.24
M	1.34	0.08	0.93	-0.64	0.68	-2.21							
B	-0.42	0.08	0.84	-1.69	0.76	-2.42							
I	-0.87	0.08	0.65	-4.03	0.73	-2.73							
F	0.87	0.08	0.72	-3.13	0.60	-3.44							
D	0.21	0.08	0.58	-4.97	0.62	-3.79							
E	0.89	0.08	0.53	-5.68	0.52	-4.25							
J	-0.38	0.08	0.65	-3.92	0.55	-5.00							

Notes: Items are sorted in order of decreasing Outfit-ZSTD value from top to bottom. Item: item label; meas.: item's calibration (logit units); model SE: model standard error of measurement; MNSQ: mean square; ZSTD: z-standardised statistics; prs: persons' label; meas.: persons' measures (in logit). The output recalls the one provided by Winsteps[®] software (winsteps.com), but other software packages provide different outputs, including p values from chi-square.

because "too much expected" scores add little information but do not imply the suspect that the measurement process is invalid.

The MNSQ represents the "mean square" statistic (MNSQ). For an item, the squared standardised residuals are calculated for each of the N observations and averaged. The MNSQ is the chi-square statistic divided by the degrees of freedom. The ideal value is 1. The index summarises "how big" the accumulation of the (squared) residuals is. In most published studies, the accepted MNSQ indexes lie between 0.5 and 1.5, as suggested in an old article. Items showing an MNSQ within this range are commonly considered "productive for measurement" [33].

However, if a perspective based on "significance" rather than size is applied to the MNSQ range, the 0.5–1.5 choice may be misleading. First, unlike the z - or t -distributions, the distribution of the chi-square statistic is asymmetrical, with a positive tail. Hence, top and bottom limits sharing the same distance from 1 (such as 0.5 and 1.5) do not entail the same p -levels. Second, the significance of the chi-square test depends on the sample size. An MNSQ value of 1.5 can be non-significant (i.e., expected) for a small sample size and highly significant (i.e., unexpected) for a larger sample size. In short, according to the significance perspective, MNSQ values "significant at $p < 0.05$ " should be adapted to the sample size (and will be asymmetrically spaced around 1). This complication does not affect the ZSTD (z -standardised) index. The Wilson-Hilferty transformation [34] makes the MNSQ index approximate the standard, symmetrical normal distribution: values of ± 1.96 SD correspond to $p < 0.05$.

Whichever the preferred index, it remains true that, with sufficiently large sample sizes, any response set "significantly" misfits

the model. For instance, it has been shown that, for sample sizes greater than 500, data can be significantly different from the model's expectations even if the MNSQ is as small as 1.1, which is commonly considered to indicate optimal fit to the model [35].

For these reasons, the analysis should not be obsessed with the p -level, which remains an arbitrary cut-off. Not surprisingly, there is a recent surge against a blind acceptance of the concept of "significance" [36], and cut-off values should not be considered as a certification for or against "evidence" [37]. More generally, clinicians should not be frightened by the high uncertainty inherent to Medicine [38].

Outfit and infit

As described above, indexes of "fit" summarise the accumulation of residuals (e.g., for an item, across responses from persons). The Outfit is a synonym for "fit." The "out" prefix highlights that the index may be sensitive to even single outliers, i.e., by highly unexpected responses causing exceptionally high residuals. Infit stands for "information weighted-fit." The residuals are multiplied by their variance. Why? A problem arises with the most informative items, i.e., those showing a difficulty level close to the person's ability level. In such cases, the probability of a "pass" or a "fail" response are similar. Consistently with the response uncertainty, high residuals are also expected by the model, and a "mis"-fit would be harder to detect without inflating the index.

The concept that, in RA, information and uncertainty travel together is profound, albeit counterintuitive, but a simple example may clarify the issue. High-jump athletes' best-ever (therefore, exceptional) performance does not give their "ability."

Ability is provided by the bar height they can cross 50% of the time: the lower or higher the bar, compared to this ability level, the more predictable the outcome (pass or fail). In these cases, a smaller variance (flagging the lower uncertainty) surrounds the response.

A significantly high outfit encourages inspection of the data to detect severe outliers and decide on further analysis (see below). A significantly high Infit reflects more of a structural “misfit” of the item, less affected by incidental outliers. Here, items G and H (Bladder and Bowel management, see Table 1(A)) are misfitting, whichever index is adopted. This is not surprising. These items are highly influenced by visceral reflexes, which are somewhat independent of the overall behavioural adaptation of the person.

As anticipated above, fit indexes may also be significantly lower than expected. This finding does not distort the measurement process but suggests redundancy of the items eliciting too predictable responses.

Dealing with persons’ and items’ misfit

If an already validated scale is applied, one can assume that items’ misfit arises from peculiarities of the sample of persons. In this case, scores of misfitting persons should be investigated and, in some cases, deleted.

In building a new scale, misfit suggests that the questionnaire should be refined. Misfitting subjects and/or items can be iteratively deleted from the analysis; item categories might be re-scored, etc. (see the companion article on the “re-scoring” procedures). If residuals are correlated, thus reflecting an extraneous variable (see next paragraph), one solution frequently found in published articles is merging a set of dependent items (a “testlet”) [39] into a unique higher-order polytomous item (“super-item”) and re-analyse the data. Testlet items are transformed into levels of a shared item. In any case, manipulating the results should always be driven, whenever possible, by substantial reasons or, to the least, by reasonable hypotheses (see below).

Fit at the whole-scale level: do residuals flag multidimensionality?

Do the residuals reflect random errors or co-vary because of shared, extraneous influences? Once the effect of the shared variable is “conditioned out,” residuals should only reflect randomness and be independent of each other [40].

a) Violation of local independence: response dependence

As Marais and Andrich elegantly highlighted [40], independence may be violated when the response to an item is influenced by the response to another item or the same item in the previous test (“response dependence”). In this case, the “extraneous” variable at work might be the rater’s tendency to infer scores from other scores rather than relying upon direct observation. In the first Author’s experience, this might be the case for the sphincter control domain of the FIM scale. The “Bowel” score is more complicated to record than the “Bladder” score (unpublished observations). Overscoring the latter will drag overscoring of the former one (mirror reasoning for underscoring).

b) Violation of local independence: bias from a shared extraneous variable

A reason for spurious score co-variation, entrained by residuals co-variation, can be the influence of an extraneous variable shared by a subset of items (e.g., language knowledge in a math test with wordy items).

As long as the fundamental axiom of “local independence” is violated, items’ scores do not depend only on the quantity of the

target variable but also on the amount of additional, extraneous variables. Items’ scores are not independent of each other once the variable modelled by the Rasch model is “conditioned out.” Scores, and the derived parameters of ability and difficulty, are no longer “sample independent” since the extraneous variables could be different and act with varying strengths in different samples of respondents.

Also, when local independency is violated, classic reliability indexes (e.g., Cronbach alpha) might be inflated by the co-variation of item responses [41]. By contrast, change indexes might be deflated because the extraneous variable may be insensitive to the causes of change in the target variable [40, 42]. This subtle yet relevant issue can be approached by detecting structural “components” within the messy residuals (see Supplemental Material, Note 3).

Fit at the whole-data matrix level

Residuals in single responses can be cumulated across the whole data matrix. The data-model consistency can be computed as a Pearson global chi-squared index. For the matrix of FIM scores analysed above, the chi-square was 6974.3, with 7093 degrees of freedom and $p=0.841$. Results highlight here a non-significant data-model misfit of the whole questionnaire. However, in the Authors’ experience, this index is often “significant” (meaning that the observed data matrix misfits the model). This is not surprising and should not raise too much concern. As reported above, any fit index may become statistically significant with a large enough sample size.

Discriminative capacity of the scale

The capacity of the modelled scale to discriminate across persons is reflected by the index of “Separation” (G = the spread of the participants’ measures expressed in standard error units), “Strata” (the number of “layers” of persons’ measures significantly different; at $p=0.05$, $\text{Strata} = (4G + 1)/3$), and “Separation Reliability” (the ratio of true measure variance to observed measure variance, ranging from 0 to 1) [43]. A Separation Reliability of at least 0.7 is required to distinguish at least two “Strata” within the measures, i.e., to reject the hypothesis that all measures reflect random error.

As a rule of thumb, person reliability > 0.7 is considered good enough when groups of persons are measured and group means are compared. However, for comparing measures of single persons (e.g., measures of the same patient before and after treatment), higher person reliability is recommended (> 0.85) [44].

Unlike CTT, in RA, the Reliability (as well as Separation and Strata) can also be computed for the items and not just persons. However, the items’ Reliability is usually satisfactory. In fact, as a rule, more persons are available to estimate item difficulty levels than items to estimate persons’ ability levels. These points are given a more detailed explanation in the companion article.

The quest for invariance of the ruler: a matter of quality, not only quantity

Measures may differ across persons and within persons across time points: but the ruler should stay the same. Stated otherwise, the “difficulty” parameters of the scale, i.e., the ticks of the Rasch “ruler,” should be stable across classes of persons and time. There should not be “differential item functioning” (DIF). More precisely, at least in the Authors’ experience, the acronym DIF refers to a

single item. The item may show observed scores constantly higher or lower than expected across the range of persons' ability (uniform DIF). This may hold for the whole sample of persons or sub-groups (genders, ethnic or diagnostic groups, etc.). The pattern of score "unexpectedness" may also interact with the ability class of the respondents (so-called non-uniform DIF).

For greater clarity, another form of DIF can be dubbed "differential test functioning" (DTF) [45]. DTF gives an overview of the scale as a whole. The mean difficulty level of all items is contrasted, through an x-y plot, through separate analyses between two classes of respondents. DIF and DTF are detected based on the statistical significance of the differences found (computed through ANOVA for DIF and regression-like testing for DTF [7]).

It must be pointed out here that, for strict statistical reasons, a real DIF in one item entails an artificial DIF or DTF in one or more other "pure" items [46] to complicate matters. Caution is needed, therefore, in deciding that malfunctioning primarily affects one or more of the statistically "diffing" items. The DTF will be treated here. For the DIF, see [Supplemental Material, Note 4](#).

Differential "test" functioning (DTF) and the item-split technique

Figure 1 gives some DTF analyses of the motor-FIM measures estimated in an inpatient rehabilitation unit at admission and discharge.

Three hundred patients (142 women) were scored (the same sample analysed in [Tables 1](#) and [3](#)). Two hundred and sixty-nine non-extreme scores were retained for the analysis. In each panel, the difficulty estimate of each item is contrasted between two sub-groups of observations (e.g., admission vs. discharge). It is

expected that values lie on the identity line and within 95% confidence limits. The figure shows that both expectations are met by the left vs. right brain damage only (upper right panel). For example, item M (Stairs) looks more difficult at admission than discharge (upper left panel). DTF emerges for various other items in other contrasts. Reasons for DTF should be investigated, not simply described. Then, a decision should be taken. Depending on the study goals, various bivariate contrasts of this kind can be applied to any other grouping criteria [47].

Understanding why the items' difficulty may be different across subgroups of respondents: two examples

a) DTF within a sample of disabled persons

In [Figure 1](#), the "independence in stairs" item applied to stroke patients looks more difficult at admission to an inpatient rehabilitation unit than at discharge. The analyst should strive to interpret this finding and give an explanation to the reader. In this case, this is a frequent bias due to excess prudence in testing patients on stairs in the early stage of recovery or, due to transient clinical instability imposing bed rest regardless of the disability level.

DTF also appears in other contrasts. For instance, several items show DTF associated with the severity of disability ([Figure 1](#), lower right corner of the plot). Items G and H (bladder and bowel management) are more difficult than the model expects in more disabled, compared to less disabled patients. A reasonable hypothesis is that the more disabled patients are bedridden and treated with enemas, catheters, and diapers and thus suffer specific limitations to the self-management of continence.

b) Cross-cultural validation of questionnaires

Motor FIM™

Self Care

- A. Eating
- B. Grooming
- C. Bathing
- D. Dressing-Upper body
- E. Dressing-Lower body
- F. Toileting

Sphincter Control

- G. Bladder Management
- H. Bowel Management

Mobility / Transfer

- I. Bed-Chair-Wheelchair
- J. Toilet
- K. Tub-Shower

Locomotion

- L. Walk-Wheelchair
- M. Stairs

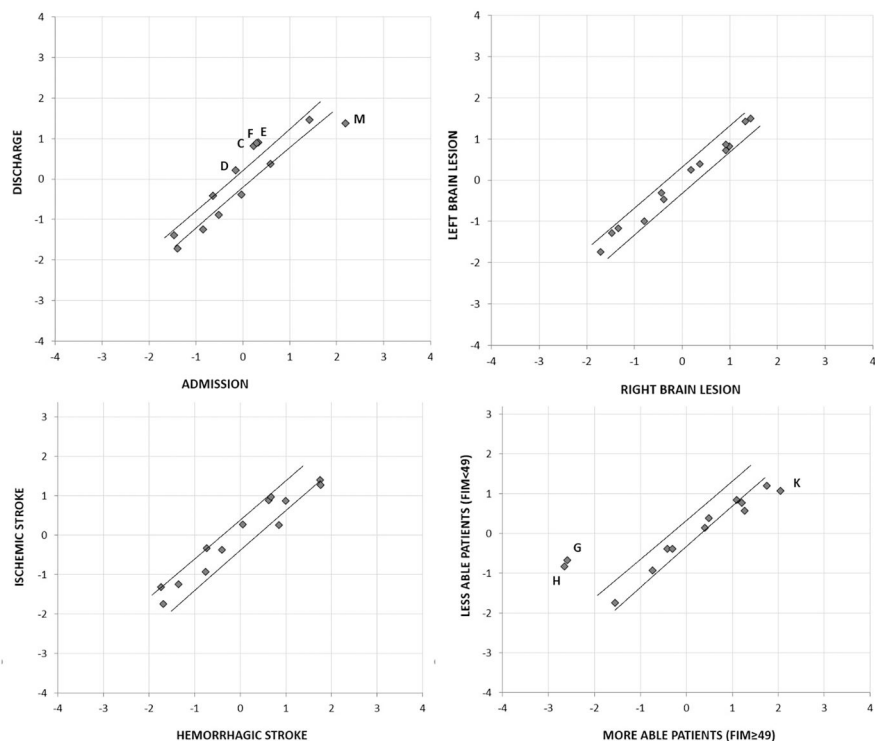


Figure 1. Differential Test Functioning (DTF) across the items of the motor-FIM scale. Data refer to hemiplegic patients admitted to and discharged from an inpatient rehabilitation unit (the same sample analysed in [Tables 1](#) and [3](#), $n = 269$). The 13 motor-FIM items are listed (for the scoring criteria, see [Table 1](#)). In each panel, item difficulty estimates are contrasted between two sub-groups, divided as per the ordinate and the abscissa criteria. These were: admission vs. discharge; for discharge only: left vs. right brain damage; ischaemic vs. haemorrhagic stroke; FIM scores below vs. at or above the median. The motor-FIM score may range from 13 to 91; the higher, the greater the patient's independence. The median score recorded in this sample at admission was 49. Diamond symbols refer to the X-Y coordinates of the items (labels are defined in the list of items on the left). The continuous lines encase the 95% confidence limits around the estimates of single items [7].

Linguistic translations of questionnaires, however elaborate, cannot warrant the conservation of the same semantic meaning nor the same metric meaning. Cultural contexts may differ even if people adopt the same language. As for the semantics, see the example provided by the attempts to translate the “Aching” descriptor of pain in the famous McGill pain questionnaire [48], initially written in English. This descriptor highlights the pervasive nature of pain, being a person’s “illness” rather than a focal “disease” [49]. It has no clearcut equivalents, for instance, in Latin nor German languages [9,50]. Even in the lucky case of semantic equivalence, the items’ “metric” meaning may remain different.

For example, consider independent “eating” and “dressing” in the Barthel scale of “independence in daily life.” The same score may indicate a different overall “independence” level depending on the cultural context. Consider using chopsticks rather than cutlery, complex vs. simple dress needs or codes, etc. The hierarchy of difficulty of the items (which means “less” or “more” difficulty within the item set) may be very different.

Another example is the “transfer to tub or shower” item of the FIM scale, which also showed variation in difficulty between Countries [51]. DIF should be expected. The size and shape of tubs and showers may differ across Countries. In addition, transferring to the tub can be more difficult than moving to the shower. Tubs may be more or less common than showers, depending on the Country.

A DTF analysis and the item-split remedy (see below) may counteract misfit [10].

DTF (or DIF) is there: what to do?

a) Dropping misfitting items (or persons): two approaches [52]

Deleting items or subjects should follow substantive, not only statistical, reasons. The most straightforward (and radical) solution is to drop the items with significant DTF from the questionnaire (the same reasoning applies to DIF). However, this approach suffers from two apparent limitations. First, if several items are removed because they show DTF or DIF, the questionnaire can become too short, and thus its persons’ Reliability would be too low. Second, eliminating items may weaken the face validity of the questionnaire (e.g., if items with established clinical relevance are deleted). Deleting a class of persons may also attenuate the DTF and the DIF. Still, it may hide diagnostic reasons for this kind of “misfit” (e.g., clinical peculiarities of the group) and reduce the generalisability to future samples.

A pragmatic approach consists of evaluating whether DTF causes harm to measures, with a procedure similar to that used to assess the consequences of multidimensionality [52,53]. Data are analysed with a set of “pure” items, which are the items showing little or no DTF concerning relevant variables. Patients’ measures (and the corresponding standard errors) from the analysis with the set of “pure” items are compared with those from the complete questionnaire (i.e., all items, including those showing DTF). Then, t-tests are calculated to compare the patients’ measures at a single subject level. DIF can be ignored if the proportion of patients measuring significantly different between the two analyses is < 5% (see Supplemental Materials 1 in [54] for additional details on this procedure).

b) The item-split technique

In the case of established DTF, the “item-split” technique can compensate for this bias. For instance, an item showing DTF, say, across men and women, is duplicated. In the “male” items, women’s responses are given as missing; in the women’s items, men’s responses are given as missing. RA manages missing responses

very efficiently so that unbiased estimates of the ability of persons of either gender can be computed. Items can be split across more than two classes. This technique may be fundamental in studies when the same questionnaire is adopted [10] in several cultural and linguistic contexts. Again, it is advisable to find a reasonable interpretation for the DTF, although it can be neutralised algebraically.

Is item difficulty different between time points? Items “anchoring” and the “rack and stack” techniques

Test-retest Reliability is conventionally tested by repeating the measurement at two-time points. This procedure assumes that the distance pattern across ticks of the ruler does not change. This requisite is fundamental when treatment outcomes have to be assessed. It has been nicely said that in persons’ measurement,

the intervention does not affect the responses to all items equally, but it more strongly influences the items it relates to. The response categories might differ across the two-time points due to the different health statuses of the patients before and after the intervention. These changes would make the meaning of change uncertain [55].

A DIF(or DTF)-by-time, which should always be suspected (e.g., [56]), poses some methodological issues given that repeated measures are collected from the same items and individuals. For example, observation independence is often required in significance testing. Items anchoring with data stacking (i.e., each person can contribute to the dataset with more than one row) or data racking (i.e., the same item contributes to the dataset with multiple columns, one for each repeated assessment) provides a solution to the repeated measurement problem (see [55] for details).

Once thresholds are estimated in a reference circumstance (e.g., at baseline before treatments within a “rack” or “stack” paradigm), these can be “imposed” to the following occasions (e.g., after treatments). Depending on the study goals, “anchor” thresholds may be computed at time 1 or time 2. To keep independence between measurements, a random selection of two half-samples can provide persons to each analysis (of course, at the cost of a lower Reliability).

Finally, the Many-Facet model (see next section) also provides a suitable solution for the time series analysis and repeated measures data.

A family of Rasch models

Andrich’s “rating scale” model

The original Rasch model, conceived for dichotomous items (no/yes; 0/1), evolved into a family of models that make it possible to analyse polytomous items (no/mild/moderate/severe, 3/2/1/0, and the like). The most famous is Andrich’s “rating scale” model [5]. This model assumes that thresholds share the same distance pattern across all items, although they need not be equispaced.

Masters’ “partial credit” model

The Masters’ “partial credit model” [57] generalises Andrich’s model. The thresholds’ distance pattern may change from item to item. In RA, this “credit” is the probability that a given response is selected. This model requires estimating more thresholds (a series for each item) than needed for the rating scale model (the same distance pattern for all items).

Deciding between the rating scale or partial credit model can be challenging, and both practical and theoretical aspects should be considered. The partial credit model has more parameters, thus returning a better fit to the model. However, given the larger number of parameters, larger datasets are needed for a partial credit analysis to have consistent parameter estimates.

The following procedure can assess if the richer partial credit model returns a better data-model fit than the simpler rating scale model. Pearson's chi-square (see above) is calculated for each item from both the partial credit and the rating scale analyses. Next, this chi-square from the two competing models is compared by computing a chi-squared difference test for each item [58]. Suppose the chi-squared difference test is not significant for each item, despite the larger number of parameters. In that case, the data-model fit of the partial credit model is not better than that of the rating scale model, and the simpler rating scale model can be preferred.

The reader can find an application of the chi-squared difference procedure detailed above in an article considering the Falls Efficacy Scale International (a questionnaire measuring concern about falling) [54]. This study highlights relevant features of either model. First, in partial credit analysis, the categories' distance pattern can change from item to item (Figure 1S in Supplementary Material 2 from [54]). Second, in rating scale analysis, items with ordered thresholds can coexist with items with disordered thresholds. Third, a questionnaire can show ordered thresholds according to the rating scale model and a mixture of items with ordered and disordered thresholds according to the partial credit model (compare Figure 2 in the main text and Figure 1S in Supplementary Material 2 from [54]).

Andrich's and Masters' models fully respect the fundamental Rasch requirement of linearity of the measure. In the Authors' opinion, adding categories to a Rasch-compliant model somehow changes the variable, which becomes "thinner" (less general) than the one tackled by dichotomous items [59]. For instance, on a pain scale, items that scored *no/yes* provide a measure of "pain in general," items scored *always/sometimes/never* measure "frequency of pain," and items scored *no pain/mild/moderate/intense* target "intensity of pain," and the like (see also the next paragraph).

Linacre's "many-facets" model

In the "many-facets" model, one or more other "facets" are added to predict the response probability. In this model, the mean item difficulty or (for polytomous items) the thresholds' difficulty and the rater's severity (or other parameters) are all subtracted from the person's ability to estimate the probability for success [60,61]. Considering the rater's severity makes the comparison between examinees fair. Also, the rater's fit can be assessed (should we trust raters assigning low scores to able subjects passing difficult items?). Another common situation calling for many-facets modelling is a multiple-rater panel evaluating research projects (rather than persons). A concrete example [15] is illustrated in the [Supplemental Online Material, Note 5](#). "Facets" may be seen as variables influencing the response probability as long as they affect the estimate of both the item's difficulty and the subject's ability. This apparently contradicts the original Rasch model (see [Equation \(2\)](#)). This is not so. The many-facets model respects the fundamental Rasch's requirement for linearity of the measure. In the model, the rater's severity is conceived as an "offset" adjusting (up or down) the measures.

A Rasch paradox: the greater the sample size, the more likely the data-model misfit

Usually, RA in clinical studies does not require large sample sizes since stable parameter estimates are obtained with a few hundred persons. As a rough rule of thumb, there should be at least ten observations per category for estimating the items category's structure [62,63]. Most published papers are based on short questionnaires (say, 5 to 30 items) applied to sample sizes ranging from 30 to 300 cases.

The greater the sample size, the more precise the model estimates are, and the more probable it is to reach a significant data-model misfit (see above the section entitled "What MNSQ (mean-square) and ZSTD (z-standardised) indexes tell").

The "size" paradox can be faced with different algebraic strategies. These solutions include analysis of small random sub-samples [28,64], attenuation of the fit indexes [65], and setting a pragmatic threshold for "misfit" at 0.5 logit difference between observed and expected parameters, neglecting the significance issue [66]. The ongoing debate will probably lead to a consensus on new fit indexes suitable for large samples.

Of course, the precision of the estimates depends on many other parameters beyond sample size: e.g., targeting and spread of the "ruler," the density of the ruler "ticks," and the fit of responses.

A philosophical hint. Rasch analysis can not demonstrate the existence of the measured variable

When applied to persons' latent traits, the mental process called measurement relies on intangible objects and intangible "units" of measurement. But: "*How can an attribute constructed by humans be a quantity or a real property?*" [67].

A "realist" perspective innervates latent traits (including Rasch's) theory [68]. The idea of "being latent" assumes existence, and the idea of "error" implies truth [69-71]. Dropped into person-metric practice, "realism" means that observed items should be conceived as the "reflection" of a latent variable existing independently from its measures and provided with a quantitative structure (one that weight has, while nationality and gender, for example, have not): variables of this kind are said "reflective." At the other extreme, items can "invent" (construct, form) the variable, which is said to be "formative" [72,73]. In this case, the variable is "that thing" measured by "that questionnaire," as per an "operationalist" view applied to psychometrics. Psychometric measurement becomes self-referenced and, according to some, a form of "pathological science" [74].

Unfortunately, a "formative" questionnaire can be Rasch-consistent even if the measured variable does not exist outside the questionnaire itself (i.e., outside the Author's mind) [75].

Different items are chosen to measure the variable of interest when a questionnaire is created. However, the same item (i.e., the same person's behaviour) can be an indicator of different variables (e.g., "crying" can be an indicator of "pain" but also of "happiness" or "rage"), and items stemming from different variables can be inadvertently brewed in the same questionnaire.

Which is then the variable measured by the questionnaire? In health care studies, "inventing" rather than "discovering" a variable is more likely when a questionnaire of, say, pain or independence in daily life is developed by including signs, symptoms, and behaviours specific to a given clinical condition. The questionnaire can provide measures useful in the building sample but lacking invariance across other clinical states (for a practical

debate on the topic, compare ref. [76] and ref. [77]). Peculiar clinical characteristics may open the door to multidimensionality [78].

The proponents of a new questionnaire are necessarily “constructing” a questionnaire of items taken from their own experience. Still, they should speculate first on reasons to suspect the real, independent existence of the variable. This is a difficult task.

Items, in themselves, may be either “formative” or “reflective” concerning a latent variable [79], depending on the order of complexity of the variable itself [59]. For instance, “joint pain” may be seen as an item “formative” concerning “independence in daily life” as well as “reflective” of the “perceived effectiveness of an anti-inflammatory drug” [80].

Formative items drag the questionnaire towards a formative nature (although no questionnaire can be 100% reflective). A good question in selecting or assessing items, based on common sense, is: do the variable’s quantity and the item’s score change together? Co-variation is more likely with reflective than formative items. If independence changes, does knee pain change? Probably not. If the dosage of the painkiller changes, does the pain perception change? Probably yes.

Potential misuse of Rasch analysis

Some hints are needed to stimulate a critical standpoint on published results. In the healthcare field, Rasch’s articles tend to be too “algebraic”: they skip over the philosophical implications of the model on one side and the practical application of results on the other. Many more articles seem to address scale construction or validation than those aiming at measuring persons [81].

Different variables, same name?

Healthcare professionals may be disconcerted by the flood of scales (including Rasch-compliant ones) available in Physical and Rehabilitation Medicine (for an excellent directory, see www.sralab.org). To add confusion, many scales are allegedly addressing the same variable. The problem is twofold. In some cases, items are different but reflect the same latent trait. This is the case, for instance, for the Barthel index and the motor-FIM subscale, reflecting the same kind of “independence in daily life.” Rasch modelling provides procedures for “equating” measures coming from “alternate form” questionnaires [82]. In essence, thresholds of different scales tackling the same variable are “anchored” to shared logit values. “Traceability” to a shared metrological standard [83] is the same requirement of physical and chemical measures taken with a different instrument or with different units (e.g., metres and feet for length, kilograms and pounds for weight).

Sometimes, however, variables with the same name are substantially different. This ambiguity may affect “formative” scales (see the countless scales claiming to measure “Quality of Life” [73]).

Rasch measures and raw scores

It is often sustained that, for most scales, raw scores provide linear measures with good approximation in the mid-range of the questionnaire scores, far enough from the floor and the ceiling of the potential questionnaire scores. This would make it useless to resort to the complexities of the Rasch analysis [84]. This argument sounds circular. The claim for a quasi-linear relationship for the middle portion of the score-to-measure curve (see Figure 5 of the companion paper) holds only once the proportionality of

scores to linear Rasch measures has been demonstrated. When a questionnaire is Rasch consistent, ordinal scores can be used as far as they benefit from their Rasch consistency. Nevertheless, as a rule, the challenging data-model misfit remains hidden under the carpet, and diagnostic chances are lost (see above). Does the theoretical superiority of Rasch measures have an experimental demonstration in terms of accuracy, precision and reliability? For many scales, this superiority has been demonstrated [85–89]. Unfortunately, most articles proposing a new instrument do not provide such evidence, perhaps valuable for convincing sceptical readers.

Does the adopted Rasch software make a difference?

Depending on the Rasch software, the reader can find different graphic and numeric outputs in different articles. One should consider that leading general software packages such as STATA[®] and SAS[®] also include macros and add-ons for Rasch analysis.

The Authors of the present paper are experienced users of the WINSTEPS[®] software. They have a smaller experience with RUMM[®] and Rasch packages within the R[®] software (the three most popular packages, at the moment, at least within the biomedical literature). Hence, the examples reported here inevitably reflect this experience. [Supplemental Material, Note 6](#) is dedicated to a brief overview of similarities and differences across leading software packages.

For the Rasch analyst, learning how to use any specific software is demanding, and a high statistical competence is required to select any package depending on the subtleties of the analysis needed. For the lay reader, the take-home message is that at least the two most popular Rasch software (i.e., WINSTEPS and RUMM – see [Supplemental Material, Note 6](#)) seem to provide similar estimates of measures, at least for clinical questionnaires. Nevertheless, numeric and graphic outputs from different software packages follow different “perspectives” on how the models should be applied, the data should be explored, and the results should be interpreted, thus leading to distinct conclusions [90,91]. “Perspectives” may well diverge within a shared theory of measurement. In a penetrating Letter, Linacre contrasted the “hypothesis testing” perspective of RUMM with the “utility perspective” of WINSTEPS. Linacre writes:

One of the two perspectives accords broadly with social-science descriptive statistics with its focus on hypothesis testing. The other perspective accords broadly with industrial quality control (W.E. Deming and Genichi Tamaguchi) with its focus on utility... the text of the paper quickly reveals if the Author is concerned about reporting findings with strong statistical properties (the hypothesis-testing perspective) or findings which maximise the usefulness of the Rasch measures for the end (the utility perspective) [92].

Rasch analysis is operator-dependent. Hints to a critical assessment

In most published articles, readers cannot see precisely how results were obtained. Researchers usually strive to achieve the highest data-model fit in the scale construction (or validation) phase. In so doing, they manipulate and re-analyse the data iteratively (e.g., [91]). Here are the most frequent procedures:

1. collapsing disordered categories (somehow a mandatory requirement);
2. deleting misfitting items;
3. deleting misfitting persons;
4. changing to “missing” some unexpected responses;

5. changing the Rasch model, e.g., from “rating scale” to “partial credit” or vice versa.

The Rasch-based refinement process is a trial-and-error one, iterative and operator-dependent. The process, however, is usually kept in the background.

At least two Guidelines for Authors publishing results obtained through RA are available. The former [93] contains 23 recommendations; the second one [94] includes 59 recommendations. Recommendations are welcome but should not be taken as procrustean constraints. Performing Rasch analysis remains intrinsically an operator-dependent method.

Lay readers, too, can appreciate the quality of a Rasch study

To accommodate the perspective of the lay readers, an attempt will be made below to summarise 12 points (labeled A to L) towards which the reader’s attention should be directed for a critical appraisal of the published results. All these points are not conceived to detract from the usefulness of Rasch-compliant scales (which the Authors of the current work consider the best psychometric options available). Instead, all the 12 points insist on a shared concept, i.e., that Rasch modelling must not be conceived as a magic wand but as a tool for stimulating diagnostic reasoning.

- A. Routine use of a Rasch-built scale on new subjects requires published item and threshold calibrations and dedicated software, which are rarely available to the readers. In this lucky case, the scores of one or more subjects are loaded, and results pop up. An excellent example is provided, for a few rehabilitation scales, by a free website making Rasch software run in the background with calibrated items and providing individual online results (person’s measure and SE, and fit indexes), also in cases with missing scores (<http://rssandbox.iescagilly.be/> (accessed January 21st, 2023)). In each study, calibrations come from the specific persons’ sample, of course, but they are relatively stable if further research, extracting the parameters “de novo,” is well conducted. Online services with “anchored” item parameters would greatly improve the usability of Rasch-made scales: but this is not the present situation. For the time going, as far as results with non-missing scores are considered, tables converting raw scores to Rasch measures and “nomograms,” allowing to estimate a person’s fit, are recommended and inserted in a growing number of Rasch articles (see the companion paper [3]).
- B. Readers should be prudent in adopting the latest published “Rasch validated” scale, even when it is a modified version of an existing scale. If data are changed during the analysis, the published “final” scale (e.g., with collapsed categories, deleted items, etc.) is not precisely the scale administered to the original sample of persons. Albeit more compliant with Rasch axioms, the new scale should be tested on the field, which the proponents rarely do.
- C. Readers should not be fascinated by the “good fit” to the Rasch model requirements. The readers would check whether, in modifying the original data and/or shifting from one model to another, the Authors displayed substantive reasons, not only algebraic arguments.
- D. Readers should reason about how and why the Authors collapsed item categories. For instance, collapsing “sometimes” and “frequently” (i.e., assigning the same raw score to both responses) looks sensible because the distinction seems quite

blurred conceptually. By contrast, collapsing “frequently” and “always” should raise concern.

- E. Authors frequently delete misfitting subjects or items to improve the overall data-model fit. Again, the readers should look for reasons motivating misfit, although these are rarely published. Also, they should look for explanations of large misfit indexes (e.g., clinical peculiarities of a misfitting subject): unfortunately, this is a rare practice from Rasch authors. The search for substantive (e.g., clinical) reasons underlying the Authors’ choices would reinforce their proposal much more than purely statistical reasons.
- F. The readers should be critical in assessing the Authors’ choice for the Rasch model adopted. In general, Andrich’s “rating scale” model requires smaller sample sizes than the partial credit models. Is this a sufficient reason to prefer the former to the latter? No, although it can be necessary due to the small sample size. There should be *a priori* reasons to assume that a change from, say, “autonomous with aids = 6” to “supervision needed = 5” (as in the FIMTM scale) requires the same change in “independence” in whichever item, be it motor, sphincteric, or cognitive. This is a heavy assumption, not needed for the partial credit model (and rarely discussed in Rasch papers).
- G. The DIF and DTF tests are necessarily limited to a few classifications (e.g., age groups, diagnostic categories, linguistic groups, etc.). The reader should ask: are these grouping criteria sufficient and relevant enough in my field of application of the scale? For instance, what about the instrument’s stability when translated from -say- English into my own (and my patients’) language? What if I’m using the scale for studying change over time? It must be admitted that the scale’s stability across time points is rarely tested (it requires double testing, often a demanding enterprise). The problem also arises in the cross-sectional use of Rasch measures, e.g., across genders or diagnostic groups. Adjusting the items affected by DTF through the item-split technique “resolves” (as per Andrich’s terminology) [46] the lack of invariance (see above). Reasons for being cautious in removing suspect “guilty” items based only on statistical significance have been given above. Detecting substantial item DTF requires a refined statistical procedure which is exceptionally found in published articles (see above and [95]). Non-statistical considerations are also important, as rightly stressed by Andrich [46,96]. Suppose the source of DTF is irrelevant to the variable at hand (e.g., a linguistic bias in a mobility test). In that case, the item-split approach will correctly attenuate unjustified differences in ability between the two classes. If the reason is relevant (e.g., on a mobility scale, more active lifestyle habits in one class), cancelling DTF will mask a substantial difference in ability levels between classes [46,96,97], which could affect the questionnaire’s validity.
- H. The fascinating x-y plot showing DTF highlights the desirable invariance of the item hierarchy (consistency). However, this does not demonstrate the full metric equivalence (agreement) of the items between the contrasted groups. For instance, is the distance from the identity line larger than the (conventional) cut-off value of 0.5 logits? The intercept may be non-zero or, stated otherwise, several items in a class may be more (or less) difficult than items in the contrasted class. However, they may still share the hierarchy of difficulty. The reasons for the systematic difference in difficulty should be investigated.

- I. About “testlets.” This issue is considered in the paragraph entitled “Fit at the whole-scale level: do residuals flag multidimensionality?.” A “super-item” trapping items with correlated residuals may appear as a magic wand to “resolve” local dependency and multidimensionality. The readers should be suspicious: “response dependence” justifies the testlet, whereas “trait dependence” suggests that distinct scales tackling different traits should be honestly considered. Substantive, extraneous reasons should always be scouted before manipulating observed data.
- J. The readers must be aware that Rasch’s modelling is (correctly) focussed on the unidimensionality of the measure. Considering the “external” validity of the scale is not its job, but this form of validity should be tested anyway. A nicely fitting scale might be useless in practice. The capacity of any measure, either Rasch or non-Rasch compliant, to predict meaningful events should be tested on the field (e.g., for the power to predict independence in daily life [98], length of inpatient stay [99,100], mortality risk [101], return to work [102], etc.).
- K. Fit and unidimensionality are not on-off concepts. On closer inspection, every object may appear as an assembly of distinct “components.” Therefore, how much fit or unidimensionality is enough is an empirical, not a theoretical, issue. The reader should check whether the analysts looked for “meaning” only from inside numbers, relying mechanistically on arbitrary cut-off values (the very concept of statistical “significance”), or tried to anchor their decisions to substantive, not only statistical, criteria.
- L. To make things even more troublesome, a Rasch-compliant scale does not warrant that the supposed variable generating the observations exists outside the analyst’s mind (as already discussed above).

The above 12 points should help the reader to interpret the researcher’s choices. In scale construction or validation, Rasch modelling helps the idea of “less to more of what” to take shape. It does not transform messy data into ideas. It forces the analyst to reflect on the reasons for the misfit. Misfit must be treated as information, not simply as noise. This conceptual effort should emerge from the published paper more, not less than glowing Rasch statistics and graphs.

Persons’ measurements are at the heart of clinical medicine, including Physical and Rehabilitation Medicine [103]. Yet, they raise more philosophical and statistical problems than those presently raised, at least in Medicine, by chemical and physical measurements. This challenge does not seem a good reason to abandon this scientific enterprise.

Acknowledgements

The authors are indebted to dr. Franco P. Franchignoni for helpful criticisms on the manuscript.

Disclosure statement

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

Funding

The Italian Ministry of Health, Ricerca Corrente IRCSS, RESET project, Fondazione Italiana Sclerosi Multipla; Ministero della Salute;

and FISM-Italian Foundation for Multiple Sclerosis (FISM), CORE-SETS project, supported this study.

ORCID

Luigi Tesio  <http://orcid.org/0000-0003-0980-1587>
 Antonio Caronni  <http://orcid.org/0000-0003-3051-1031>
 Anna Simone  <http://orcid.org/0000-0003-1537-3187>
 Dinesh Kumbhare  <http://orcid.org/0000-0003-3889-7557>
 Stefano Scarano  <http://orcid.org/0000-0002-9217-4117>

References

- [1] Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *J Rehabil Med.* 2003;35(3):105–115.
- [2] 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.
- [3] Tesio L, Caronni A, Kumbhare D, et al. Interpreting results from Rasch analysis 1. The ‘most likely’ measures coming from the model. *Disabil Rehabil.* DOI: [10.1080/09638288.2023.2169771](https://doi.org/10.1080/09638288.2023.2169771)
- [4] Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago (IL): University of Chicago Press; 1980.
- [5] Andrich D. Rasch models for measurement. Newbury Park (CA): Sage Publications; 1988.
- [6] 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.
- [7] Wright B, Stone M. Best test design. Rasch measurement. Chicago (IL): MESA Press; 1979.
- [8] Penta M, Thonnard JL, Tesio L. ABILHAND: a Rasch-built measure of manual ability. *Arch Phys Med Rehabil.* 1998; 79(9):1038–1042.
- [9] Tesio L, Granger C V, Fiedler RC. A unidimensional pain/disability measure for low-back pain syndromes. *Pain.* 1997;69(3):269–278.
- [10] 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;42:137–148.
- [11] 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.
- [12] Fisher AG. The assessment of IADL motor skills: an application of many-faceted Rasch analysis. *Am J Occup Ther.* 1993;47(4):319–329.
- [13] Tesio L, Perucca L, Battaglia MA, et al. Quality assessment of the FIM (functional independence measure) ratings through Rasch analysis. *Eur Medicophys.* 1997;33:69–78.
- [14] Cella D, Gershon R, Lai JS, et al. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Qual Life Res.* 2007;16:133–141.
- [15] Tesio L, Simone A, Grzeda MT, et al. Funding Medical research projects: taking into account referees’ severity and consistency through many-faceted Rasch modeling of projects’ scores. *J Appl Meas.* 2015;16(2):129–152.
- [16] Gantschnig BE, Fisher AG, Page J, et al. Differences in activities of daily living (ADL) abilities of children across world

- regions: a validity study of the assessment of motor and process skills. *Child Care Health Dev.* 2015;41(2):230–238.
- [17] 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.
- [18] Linacre JM. Rasch dichotomous model vs. one-parameter logistic model. *Rasch Meas Trans.* 2005;19:1032.
- [19] WPJr F. IRT and confusion about Rasch measurement. *Rasch Meas Trans.* 2010;24:1288.
- [20] Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? *Med Care.* 2004;42: i6–17.
- [21] Andrich D. Advances in social measurement: a Rasch measurement theory. In: Guillemin F, Lepège A, Briçon S, Spitz E, J. Coste J, editors. *Perceived health and adaptation in chronic disease.* Routledge:Taylor & Francis Group; 2018. p. 66–91.
- [22] Andrich D. Rating scales and Rasch measurement. *Expert Rev Pharmacoecon Outcomes Res.* 2011;11(5):571–585.
- [23] Linacre JM. PCA: data variance: explained, modeled and empirical. *Rasch Meas Trans.* 2003;17:942–943.
- [24] Linacre JM, Tennant A. More about critical eigenvalue sizes (variances) in standardized-residual principal components analysis (PCA). *Rasch Meas Trans.* 2009;23:1228.
- [25] Linacre JM. Detecting multidimensionality: which residual data-type works best? *J Outcome Meas.* 1998;2:266–283.
- [26] Smith RM. Detecting item bias with the Rasch model. *J Appl Meas.* 2004;5:430–449.
- [27] Bernstein I, Samuels E, Woo A, et al. Assessing DIF among small samples with separate calibration t and Mantel-Haenszel χ^2 statistics in the Rasch model. *J Appl Meas.* 2013;14:389–399.
- [28] Müller M. Item fit statistics for Rasch analysis: can we trust them? *J Stat Distrib Appl.* 2020;7:5 (2020).
- [29] Christensen KB, Thorborg K, Hölmich P, et al. Rasch validation of the danish version of the shoulder pain and disability index (SPADI) in patients with rotator cuff-related disorders. *Qual Life Res.* 2019;28(3):795–800.
- [30] 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.
- [31] Linacre JM, Wright BD. (Dichotomous mean-square) chi-square fit statistics. *Rasch Meas Trans.* 1994;8:360.
- [32] Smith R, Suh K. Rasch fit statistics as a test of the invariance of item parameter estimates. *J Appl Meas.* 2003;4(2): 153–163.
- [33] Wright B, Linacre M. Reasonable mean-square fit values. *Rasch Meas Trans.* 1994;8:370.
- [34] Schulz M. The standardization of mean-squares. *Rasch Meas Trans.* 2002;16:879.
- [35] Smith RM, Schumacker RE, Bush MJ. Using item mean squares to evaluate fit to the Rasch model. *J Outcome Meas.* 1998;2:66–78.
- [36] Amrhein V, Greenland S, Mcshane B. Retire statistical significance. *Nature.* 2019;567:305–307.
- [37] Berkson J. Tests of significance considered as evidence. *J Am Stat Assoc.* 1942;37:325–335.
- [38] Tonelli MR, Upshur REG. A philosophical approach to addressing uncertainty in medical education. *Acad Med.* 2019;94(4):507–511.
- [39] Wainer H, Bradlow ET, Wang X. What's a testlet and why do we need them? In: *Testlet response theory and its applications.* Cambridge: Cambridge University Press; 2007. p. 44–59.
- [40] Marais I, Andrich D. Formalizing dimension and response violation of local independence in the unidimensional Rasch model. *J Appl Meas.* 2008;9:200–215.
- [41] Christensen KB, Makransky G, Horton M. Critical values for yen's Q3: identification of local dependence in the Rasch model using residual correlations. *Appl Psychol Meas.* 2017;41(3):178–194.
- [42] Wright BD. Rack and stack: time 1 vs. time 2 or pre-test vs. post-test. *Rasch Meas Trans.* 2003;17:905–906.
- [43] Wright B, Masters G. Number of person or item strata: $(4 \times \text{separation} + 1)/3$. *Rasch Meas Trans.* 2002;16:888.
- [44] 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.
- [45] Badia X, Prieto L, Linacre JM. Differential Item and test functioning (DIF & DTF). *Rasch Meas Trans.* 2002;16:889.
- [46] Andrich D, Hagquist C. Real and artificial differential item functioning in polytomous items. *Educ Psychol Meas.* 2015;75(2):185–207.
- [47] Gel K, 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.
- [48] Melzack R. The short-form McGill pain questionnaire. *Pain.* 1987;30(2):191–197.
- [49] Tesio L, Buzzoni M. The illness-disease dichotomy and the biological-clinical splitting of medicine. *Med Humanit.* 2021;47(4):507–512.
- [50] Maiani G, Sanavio E. Semantics of pain in Italy: the italian version of the McGill pain questionnaire. *Pain.* 1985;22(4): 399–405.
- [51] Lundgren-Nilsson Å, Grimby G, Ring H, et al. Cross-cultural validity of functional independence measure items in stroke: a study using Rasch analysis. *J Rehabil Med.* 2005; 37(1):23–31.
- [52] Tennant A, Pallant JF. DIF matters: a practical approach to test if differential item functioning makes a difference. *Rasch Meas Trans.* 2007;20:1082–1084.
- [53] Mallinson T. Rasch analysis of repeated measures. *Rasch Meas Trans.* 2011;25(1):1317.
- [54] Caronni A, Picardi M, Redaelli V, et al. The Falls Efficacy Scale International is a valid measure to assess the concern about falling and its changes induced by treatments. *Clin Rehabil.* 2022;36(4):558–570.
- [55] Anselmi P, Vidotto G, Bettinardi O, et al. Measurement of change in health status with Rasch models. *Health Qual Life Outcomes.* 2015;13:1–7.
- [56] Tesio L, Perucca L, Franchignoni FP, et al. A short measure of balance in multiple sclerosis: validation through Rasch analysis. *Funct Neurol.* 1997;12:255–265.
- [57] Wright B, Masters G. *Rating scale analysis.* Chicago (IL): MESA Press, The University of Chicago; 1982.
- [58] Werner C, Schermelleh-Engel K. Deciding between competing models: chi-square difference tests; 2010; [accessed 2022 Oct 29]. Available from: <http://www.dgps.de/>
- [59] Tesio L. Items and variables, thinner and thicker variables: gradients, not dichotomies. *Rasch Meas Trans.* 2014;28: 1477–1479.
- [60] Linacre JM. *Many-Facets Rasch measurement.* 2nd ed. Chicago (IL): MESA Press; 1994.
- [61] Eckes T. *Introduction to Many-Facet Rasch measurement: analyzing and evaluating rater-mediated assessments.* 2nd ed. Frankfurt am Mein: Peter Lang GmbH; 2015.

- [62] Linacre JM. Optimizing rating scale category effectiveness. *J Appl Meas*. 2002;3(1):85–106.
- [63] Linacre JM. Sample Size and item calibration [or person measure] stability. *Rasch Meas Trans*. 1994;7:328.
- [64] Bergh D. Chi-squared test of fit and sample size—a comparison between a random sample approach and a chi-square value adjustment method. *J Appl Meas*. 2015;16:204–217.
- [65] Tristan. An adjustment for sample size in DIF analysis. *Rasch Meas Trans*. 2006;20:1070–1071.
- [66] Wang WC, Yao G, Tsai YJ, et al. Validating, improving reliability, and estimating correlation of the four subscales in the WHOQOL-BREF using multidimensional Rasch analysis. *Qual Life Res*. 2006;15(4):607–620.
- [67] Maul A, Torres Iribarra D, Wilson M. On the philosophical foundations of psychological measurement. *Measurement*. 2016;79:311–320.
- [68] Michell J. The logic of measurement: a realist overview. *Measurement*. 2005;38:285–294.
- [69] Borsboom D, Mellenbergh GJ, van Heerden J. The theoretical status of latent variables. *Psychol Rev*. 2003;110(2):203–219.
- [70] Buzzoni M. Robustness, intersubjective reproducibility, and scientific realism. In: Agazzi E, editor. *Varieties of scientific realism*. Heidelberg (NY): Springer International Publishing; 2017. p. 133–150.
- [71] Agazzi E, editor. *Scientific truth revisited*. In: *Scientific objectivity and its contexts*. Cham: Springer International Publishing; 2014. p. 387–411.
- [72] Edwards JR, Bagozzi RP. On the nature and direction of relationships between constructs and measures. *Psychol Methods*. 2000;5(2):155–174.
- [73] Tesio L. Quality of life measurement: one size fits all. *Rehabilitation medicine makes no exception*. *J Med Person*. 2009;7:5–9.
- [74] Michell J. Normal science, pathological science and psychometrics. *Theory Psychol*. 2000;10:639–667.
- [75] Stenner AJ, Fisher WP, Stone MH, et al. Causal Rasch models. *Front Psychol*. 2013;4:1–14.
- [76] 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(2):131–140.
- [77] 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.
- [78] Franchignoni F, Mora G, Giordano A, et al. Evidence of multidimensionality in the ALSFRS-R scale: a critical appraisal on its measurement properties using Rasch analysis. *J Neurol Neurosurg Psychiatry*. 2013;84(12):1340–1345.
- [79] Andrich D. A structure of index and causal variables. *Rasch Meas Trans*. 2014;28:1475–1477.
- [80] Tesio L. Causing and being caused: items in a questionnaire may play a different role, depending on the complexity of the variable. *Rasch Meas Trans*. 2014;28:1–3.
- [81] Tesio L, Simone A, Bernardinello M. Rehabilitation and outcome measurement: where is Rasch analysis-going. *Eura MedicoPhys*. 2007;43:119–132.
- [82] Skaggs G, Wolfe EW. Equating designs and procedures used in Rasch scaling. *J Appl Meas*. 2010;11(2):182–195.
- [83] Pendrill L. Man as a measurement instrument. *J Meas Sci*. 2014;9:24–35.
- [84] Xu T, Stone CA. Using IRT trait estimates versus summated scores in predicting outcomes. *Educ Psychol Meas*. 2012;72:453–468.
- [85] 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.
- [86] 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.
- [87] O'Connor RJ, Cano SJ, Thompson AJ, et al. Exploring rating scale responsiveness: does the total score reflect the sum of its parts? *Neurology*. 2004;62(10):1842–1844.
- [88] Cano SJ, Barrett LE, Zajicek JP, et al. Beyond the reach of traditional analyses: using Rasch to evaluate the DASH in people with multiple sclerosis. *Mult Scler*. 2011;17(2):214–222.
- [89] Caronni A, Donzelli S, Zaina F, et al. The Italian Spine Youth Quality of Life questionnaire measures health-related quality of life of adolescents with spinal deformities better than the reference standard, the scoliosis research society 22 questionnaire. *Clin Rehabil*. 2019;33:1404–1415.
- [90] Linacre JM. Standardized mean-squares. *Rasch Meas Trans*. 2001;15:813.
- [91] 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.
- [92] Linacre JM. Two perspectives on the application of Rasch models. *Eur J Phys Rehabil Med*. 2010;46:301–302.
- [93] Smith RM, Linacre JM, Smith EVJ. Guidelines for manuscripts. *J Appl Meas*. 2003;4:198–204.
- [94] van de Winkel A, Kozlowski 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.
- [95] Hagquist C, Andrich D. Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health Qual Life Outcomes*. 2017;15:1–8.
- [96] El Masri YH, Andrich D. The Trade-off between model fit, invariance, and validity: the case of PISA science assessments. *Appl Meas Educ*. 2020;33:174–188.
- [97] Kopf J, Zeileis A, Strobl C. Anchor selection strategies for DIF analysis: review, assessment, and new approaches. *Educ Psychol Meas*. 2015;75(1):22–56.
- [98] Stineman MG, Goin JE, Granger C V, et al. Discharge motor FIM-function related groups. *Arch Phys Med Rehabil*. 1997;78(9):980–985.
- [99] Stineman MG, Granger CV. A modular case-mix classification system for medical rehabilitation illustrated. *Health Care Financ Rev*. 1997;19:87–103.
- [100] Franchignoni FP, Tesio L, Martino MT, et al. Length of stay of stroke rehabilitation inpatients: prediction through the functional independence measure. *Ann Ist Super Sanita*. 1998;34(4):463–467.
- [101] Crimaldi S, Porta A, Vaccari A, et al. A comparison between extensive and intensive rehabilitation: FIM measurements as indicators of appropriateness and efficiency. *Eura MedicoPhys*. 1999;35:177–183.
- [102] Cantagallo A, Carli S, Simone A, et al. MINDFIM: a measure of disability in high-functioning traumatic brain injury outpatients. *Brain Inj*. 2006;20(9):913–925.
- [103] Tesio L. 6.3B Scientific background of physical and rehabilitation medicine: specificity of a clinical science. *J Int Soc Phys Rehabil Med*. 2019;2: S113–S121.