Original article

Measuring through questionnaires: Rasch Analysis as a tool for keeping the same meaning across different patients

Misurare con i questionari: l'analisi di Rasch come strumento per mantenere lo stesso significato in pazienti diversi

Luigi Tesio^*, Laura Perucca^, Anna Simone^

Clinical Unit and Laboratory of Research of Neuromotor Rehabilitation, Istituto Auxologico Italiano, IRCCS
e *Università degli Studi, Chair of Rehabilitation Medicine, Milano, Italy

Abstract – Whole-person variables may describe behaviours, perceptions, knowledges, attitudes. These are the outcomes of the most various health-care interventions. Such variables can only be observed through samples of representatives behaviours (items in questionnaires). The amount of the variables is usually inferred through counts of the scores arbitrarily assigned to the various items. Rasch Analysis (RA) is a novel statistical method allowing the estimate of true linear measures of item difficulty and subject ability subtended by raw counts. This also allows the estimate of stability of the items hierarchy (which item is more difficult, which is less) across time, raters and diagnostic or cultural subgroups. If this hierarchy is unstable (Differential Item Functioning – DIF) the same questionnaire actually depicts qualitatively distinct, non comparable, conditions. RA may help to either detect the problem, or re-calibrate the subject measures by accounting for DIF.

This may warrant real metric equivalence across medical questionnaires applied to different diagnostic groups and/or different linguistic/cultural contests, thus fostering multicentric, international trials.

Key words: measurement, questionnaire, Rasch Analysis
Journal of Medicine and The Person, 2006; 4(2): 54-61
Received April 10th 2006, Accepted April 11th 2006

Measuring the person: the gradient from the parts to the whole.

The dominant medical model is biological[1]. Biology takes its research paradigms from "hard" sciences such as physics and chemistry. A cornerstone of the model is the philosophical approach called reductionism (as opposed to holism). Reductionism postulates that the whole (any "black-box", or system) can be understood through its parts, and that a cause-effect relationship can be established, provided that relationships across the internal parts are understood. Anatomy is perhaps the simplest example of the strength of this approach. For instance, how and why circulation and joint movement can happen could be understood only after body dissections could be rigorously practiced. Infectious diseases could be only understood only after micro-organisms could be extracted from the body and observed etc. The astonishing successes obtained by this paradigm might favour an extremist interpretation seeing reality as pure appearance: "true" reality always lies behind. In this view, "I" as a person am not different from an ideal robot replicating my internal parts and their relationships (the myth of a "human machine", either mechanically or genetically built, is as old as man[2]).

From a physician's perspective, it seems convenient to accept two pragmatic view points:

1. Both a holistic and a reductionistic approaches can be useful

Like any other scientific paradigm, reductionism also has its limits. Some phenomena can be better understood if they are treated as a whole, renouncing the reductionist approach. After all, what is it a part and what is it a whole is a conventional matter. Is a salmon a whole, as a fish, or is it part of a salmon shoal? Zoologists and statisticians know that it is easy to predict precisely the shoal position next year, while it is impossible to predict where the individual salmon will be within the shoal. Is a liver cell a part or a whole? One should say both (liver is the whole, cell organs are the parts?). Is the person a part or a whole? Leave aside ethical and religious concerns and stick to the salmon example. Which is the most appropriate scientific standpoint? Both of them are appropriate, of course, depending on the goal.

2. Holism and reductionism are aligned along a continuum

There is no reason to postulate discontinuity between holism and reductionism. While reading these lines, the reader actually "sees" concepts. If he/she is focusing on the typed characters, letters will be seen. Through a magnifying lens, ink dots will be perceived. Seen from a 6-m distance, the reader will only perceive a journal, etc. To sum up, it is not convenient thinking in antagonistic terms of whole/parts, forest/tree etc., but thinking in terms of different "zooming" onto a unique reality. How much one will zoom will depend on the issue at hand.

Measuring behaviours, perceptions, attitudes and knowledges: this is best accomplished under the latent variable theory.

If one has to measure behaviours and any kind of "mental/cognitive" properties (heretofore, person's variables), then a far perspective is needed (do not zoom too much). Behaviours can be thought of as the exchange of energy and/or information between the person and the environment (inclusive of other persons[3]). Locomotion, verbal communication, working ability cannot be ascribed to any "part" of the person. The same holds for depression, intelligence, knowledge of mathematics, political preferences. "Mental/cognitive" properties, too, need a motor behaviour (e.g., ticking a questionnaire) to be observed.

a. Measure can only be inferred from behaviours

Can these "person's variables" be measured, or can they be qualitatively described at best[4]?

Here, measurement can be said as the conceptual alignment of object along the conceptual continuum of "less-to-more". The issue of person's measurement was first approached scientifically in the early 19th century, and it is a very intricate an sophisticate one[5].

The state-of- the-art answer is: person's variables can be measured, provided the "latent theory" approach is properly applied.

Briefly, person's variables are postulated as inaccessible to direct observation. In the meanwhile, behaviours can be observed, deemed to be representative of the "latent", underlying variable. For instance, how "able to walk" is a person? This ability cannot be directly observed in its entirety: it is a hidden property. In the meanwhile, one can directly observe that today a person can walk for no more than 100 meters, that he/she needs a cane, and that he/she cannot walk uphill. The observer can infer that this person is in general "less able to walk", compared to a person that can walk 5 kilometres with no aid, and who walk uphill with only minimal fatigue. Observations indeed allow inferences. The other side of the coin is: knowledge is only probabilistic.

b. Behaviours are not measures in themselves, they represent amount of the latent variables

A key point is that behaviours are not measures in themselves, they are just indexes of the latent variable. Walking distance is a measure of distance, but not one of "walking ability". Walking 300 meters with another one's support may mean "less walking ability" than walking 100 meters alone.

The same reasoning applies, of course, to mostly "mental" variables such as knowledge.

Suppose a teacher wants to measure the math knowledge of a 9-yr old student. Ten questions are asked. Why ten, and why *those* ten? This is conventional. The teacher example helps understanding two features intrinsic to the "latent theory approach". First, behaviours (here: the elicited answers to questions) are only a random set coming from a potentially infinite pool of observations. Second, each observation represents a given "amount" of the latent variable: here lies its information content.

Everyone should agree that the right answer to the question "3-2=?" means "less" math knowledge compared to the right answer to the question "square root of 64=?". Asking also "4-3=?" would probably add little information. The two questions on subtraction are different, yet they likely represent the same amount of "math knowledge".

c. Counting behaviours is a first approximation to measures, and the basis of questionnaires

By counting the number of behaviours, one can start building an approximate measure. Being able to either walk alone on flat ground or climb uphill means "more walking ability" (or "less disability", this a matter of convention), compared to walking on flat ground only. Answering to either the subtraction or the square root question means "more math knowledge" than answering the former question only. This is the rationale underlying cumulative questionnaires, so widely applied in Medicine[6] (to say nothing of Education and Social Sciences).

The questionnaire boom: a risk for a measurement flop.

Questionnaires are facing a growing popularity, since "outcome" measures are increasingly required by leading medical journals. An "outcome" is any event relating to the person as a whole: mortality, disability, satis-

© Punto Effe, 2006

faction, return to work, compliance with the therapy, etc.

In recent years, the medical community (mostly coming from a "bio"-medical background, and hence used to apply physical measures), enthusiastically adopted questionnaires, often resuscitating old instruments developed well before the latent theory approach was born. Questionnaires provide numbers, hence they appeared measures ready for use. Why not? For several good reasons.

Counting is not measuring

Buying six oranges rather than four does not guarantee more orange juice: buying 2 Kilograms rather than 1 Kilogram probably does. Individual oranges may be different, Kilograms are not.

This implies aligning concrete oranges along the abstract continuum of "weight" (here, both concrete and abstract objects are considered to be "real"). When the scores on a questionnaire are counted, one should know "how big is each orange" (or how difficult is an item) added to the basket. Also, the consistency of responses should be considered[7].

Take the example provided by Table 1.

The hierarchy of "math knowledge" is not the one suggested by total scores. Question 5, presumably, has to do with knowledge of history not less than knowledge of math: it is wiser to neglect the score achieved on this question. Questions 1 and 2 indicate the same amount of knowledge: missing either of these two questions is probably a fortuitous event (answering neglected?). Answering question 4 while missing question 3 is more puzzling, given that the latter is more difficult than the former. Was answering question 3 neglected? Or: was the answer to question 4 copied from the next-bench student? Did subject D tried to guess in either item 3 or 4? Subject C probably knows as much math as subject B, while Subject D is suspect for knowing math less than subject A.

It is clear that raw scores can be highly misleading, despite their numeric appearance. A theory is required to transform raw counts (scores) into valid, linear, continuous measures. The state-of-the-art theory is Rasch modelling.

Rasch modelling as the shuttle from counts to measures.

The Danish mathematician Georg Rasch (who died in 1980) provided the scientific community with the solution to the long recognized problem of transforming counts into measures[8],[9]. He established the "rules" for such transformations (heretofore, the "Rasch model") in 1960. His work was largely ignored. After 1980 the model gained increasing interest mostly in the psychometric and the educational fields[10]. Since the late '80s, it gained increasing popularity across the medical community, starting with Rehabilitation medicine[11]. Rasch-based papers are now booming in referenced Journals (with some risk for an overly a simplistic application).

For the technicalities, the reader can refer to entire dedicated books and journal reviews[12,13,14]. Nowadays, these are easily accessible to bio-medical researchers provided with only basic statistical knowledge.

The point here is *what* the Rasch model does, not *how* does it.

Suppose you have a series of patients each getting a score, representing an alternative choice across 2 or more options (yes/no; pass/fail; 0/1; no/mild/moderate/severe; 0/1/2/3, etc.) on a series of items.

Table 2 focuses on item difficulties (now, items are on the abscissa). The table gives a schematic example of the simplest "dichotomous" case, where answers can be either 0 or 1, "1" meaning "more" of the variable (e.g. correct answer, versus wrong answer; walking alone versus walking with crutches, etc.). M in the table stands for "missing": the subject simply omitted the answer. The sum of the scores achieved by each subject is a first approximation to the amount of variable "latent" within the subject (conventionally defined as subject's "ability"). The sum of the scores achieved by each item across subjects is a first approximation to the "easiness" of the items. It is more convenient and intuitive to man-

Tab. 1 – Math knowledge questionnaire; 0= wrong or missing; 1=correct

Questions / Items	*Subject A*	*Subject B*	*Subject C*	*Subject D*
1 3-2=?	1	1	1	1
2 4-3=?	1	1	0	1
3 Sqrt 64?	1	1	1	0
4 Log 10 1627=?	0	1	1	1
5 Was Décartes born before Newton?	1	0	0	1
Total score (amount of knowledge)	4	4	3	4

Tab. 2 – Questionnaire: "dichotomous" case; 0= wrong; M=missing; 1=correct

Subjects	*Item a*	*Item b*	*Item c*	*Item d*	*Subject's "ability" score*
Subject 1	1	1	1	1	4
Subject 2	1	1	M	1	3
Subject 3	1	1	0	0	2
Subject 4	1	0	0	0	1
Item "difficulty" score (max-score)	0	1	3	2	

age the item "difficulty", so that the cumulative score of item difficulty is achieved by subtracting the observed score from the highest possible score. Item b is the most difficult, because it got the fewest count of "1" answers.

Rasch modelling starts from raw counts and does (among many other things) the following:

1. It jointly estimates subject's ability and item difficulty, providing linear measures. Getting "1" does not mean the same increase in ability whatever the item answered (some are easy, some are difficult). Mirror reasoning applies to items: getting "1" means a different difficulty, depending on the ability of the answering subject. Rasch measures provide true interval units. For both items and subjects, advancing 1 unit means the same increment, e.g. 1-0= 2-1=3-2 etc.
2. It estimates what the missing response would likely have been. Having passed the more difficult item d, Subject 2 can be confidently ascribed a "1" answer (hence, the linear measure corresponding to ability score 4, not 3). Item c can be confidently ascribed a lower difficulty measure (coming from a difficulty score of 2, not 3).
3. It measures the consistency of the scoring pattern[15]. Subject 2 passed the most difficult item d, yet he/she failed the easier item c. Did the subject luckily guess? A measure of "fit" of the subject's measure with the model-expected response profile is made available to the analyst.

An important point is that the Rasch model is probabilistic in nature. It "estimates" ability, difficulty and fit, together with confidence limits. No answer is deemed to be impossible; all answers are assigned a probability. Hence, like in conventional statistics, "significance" levels can be set for Rasch measures.

The so called "Rasch separability theorem" demonstrate that only the Rasch model provides true linear measures[16].

Rasch Analysis in practice

Rasch Analysis has three main applications in the medical field: a) it can be used to create new questionnaires from scratch[17], thus facilitating the achievement of good metric properties; b) it can be used to validate, refine or reject existing questionnaires[18]; c) once the questionnaire is deemed to be valid, it can be used to validly measure subjects. This paper is focussed on a particular aspect of the questionnaire validity, e.g. its conceptual "stability" across sub-groups of subjects.

Focus on sub-groups analysis: differential item functioning (DIF).

a) Difficulty of one item can be unstable across classes of patients (properly said DIF)

A key property of a questionnaire, so often neglected in the literature, is its "stability" across subgroups of subjects ("classes"). By "stability", here, it is not meant the overall score (of course, this may change, for instance from before to after treatment). Rather, it is meant the stability of the *relative* difficulty across items. If item b) is more difficult than item a) by a given amount, the same should hold whatever the sample of subjects tested, the raters, the time of testing, etc. In conventional psychometrics, such invariance is only inferred through repeated measurements. These are just samples of infinite potential replications. Rasch Analysis often shows that the amount of the difference in item difficulty levels can change so much, that the order of difficulty (the item "hierarchy") is even reversed. A clear example is provided by questionnaires provided with items culture or disease-dependent. For instance, "Eating" results as easier than "Upper body dressing" in most questionnaires of disability and/or upper limb mobility. Yet, "Eating" may become more difficult than "Upper body dressing" for Japanese subjects adopting chopsticks and/or in patients suffering from diseases specifically affecting the mobility of the finger joints (e.g. rheumatoid arthritis). The same cumulative scores may thus depict conditions that are *qualitatively* different and thus incommensurable in different classes of subjects. Paradoxically, cumulative scores may tell us "how much", but they may not tell "how much of what" (10 Kg for a class, 10 litres for another class?).

Rasch modelling represents an enlightening approach to this issue, which is named "differential item functioning", or DIF.

Figure 1 shows a scale developed under guidance from Rasch Analysis. It is called LAPMER[19], after "Level of Activity in Profound/Severe Mental Retardation". The

© Punto Effe, 2006

Fig. 1 – The LAPMER (Level of Activity in Profound/Severe Mental Retardation) scale: items and score

ITEMS	*Categories*	*Score*
FEEDING	Fed, modified food	0
	Fed, ordinary food	1
	Brings food to mouth (either with or without help or supervision)	2
SPHINCTERS	Does not signal need or leakage (bladder or bowel)	0
	Does signal, need or leakage	1
COMMUNICATION	Signals some need, unspecific stereotyped behavior	0
	Signals some need, identifiable through behavior	1
	Communicate needs verbally	2
MANIPULATION	Absent or grasping reaction	0
	Spontaneous palmar grasp	1
	Index-thumb pinch	2
DRESSING	Stays passive	0
	Strives to give some collaboration	1
LOCOMOTION	Stationary, chair/wheelchair	0
	Moves around	1
SPATIAL ORIENTATION	No spatial orientation	0
	Orients him/herself in customary environment	1
	Orients outside his/her ward	2
PRAXIAE	None, or aimless and stereotyped	0
	Makes plastic or graphic products (embedding, chaining, moulding-colouring) or drives wheelchair manually	1
	Makes drawings or drives electric wheelchairs	2

goal was to create a measure of disability in mentally retarded people whose IQ was much below 25%, so that conventional psychometric tests could not apply. "Ability" had thus to be inferred from very primitive motor activities, taken as representative of different amounts of the latent "mental ability" (of course, assumed to be the higher, the lower the "retardation"). Just a note on item "sphincters", for example. Originally the authors assigned scores 0,1,2 to subjects showing decreasing frequency of incontinence. Rasch Analysis showed that scores on this item were inconsistent with the overall subjects' ability level. This was ascribed to the presence of epilepsy and/or spasticity in several patients. Incontinence, therefore, could come from neurological, not mental, impairments. Signalling or not the need for bladder/bowel voidance was deemed more representative of "mental" ability: this was confirmed by Rasch Analysis.

Figure 2 gives an example of the basic graphic output from a Rasch Analysis, i.e. the so-called "Rasch ruler" (Winsteps 3.2 software).

The difficulty levels of the items become the "ticks" of a familiar metric ruler. The vertical dashed line represents the "amount of the variable". From bottom to top, increasing levels of item difficulty (right) or subject's ability (left; X=2 persons, *=1 person) are aligned along a linear continuum. For dichotomous items (e.g. sphincters 0/1) the average difficulty is given. For items graded on more than 2 levels (e.g. "feeding" 0/1/2) the so called "thresholds" between adjacent scores are given. For instance "Feeding.1" is the 1st "threshold", which can be thought of as the mean difficulty level between score 0 and 1; "Feeding.2" is the 2nd "threshold", i.e. the mean difficulty level between score 1 and 2, etc. The so-called "logit" units (left column) are linear: the difference between 0 and -1 is the same as the difference between 3 and 2, etc. A "0" value is conventionally assigned to the mean difficulty of the items. The higher the logit value, the higher the subject's ability or the item difficulty. A

Fig. 2 – The "Rasch ruler" (output from Winsteps, 3.2 software)

Location (logits)	*Persons (X=2 persons *=1 person)*	*Item Thresholds*
5.0		
4.0	*X	
		SpaceOrientation.2
		Praxiae.2
3.0	*XXXX	Communication.2
	*XXXXXXX	
2.0		
	*XXXXXXXXXX	
		Praxiae.1
1.0		
	XXXXXX	Sphincters.1
0.0	*XXX	
		Dressing.1
	*XXX	Communication.1
		Manipulation.2
-1.0	*XXXX	
		Feeding.2
	*XXX	
	SpaceOrientation.1	
-2.0	XXX	
		Locomotion.1 Feeding.1
	XXXX	
-3.0	Manipulation.1	
	XXXXXXX	
-4.0		

Fig. 3 – "Differential Item Functioning-DIF": Rasch analysis across three items.

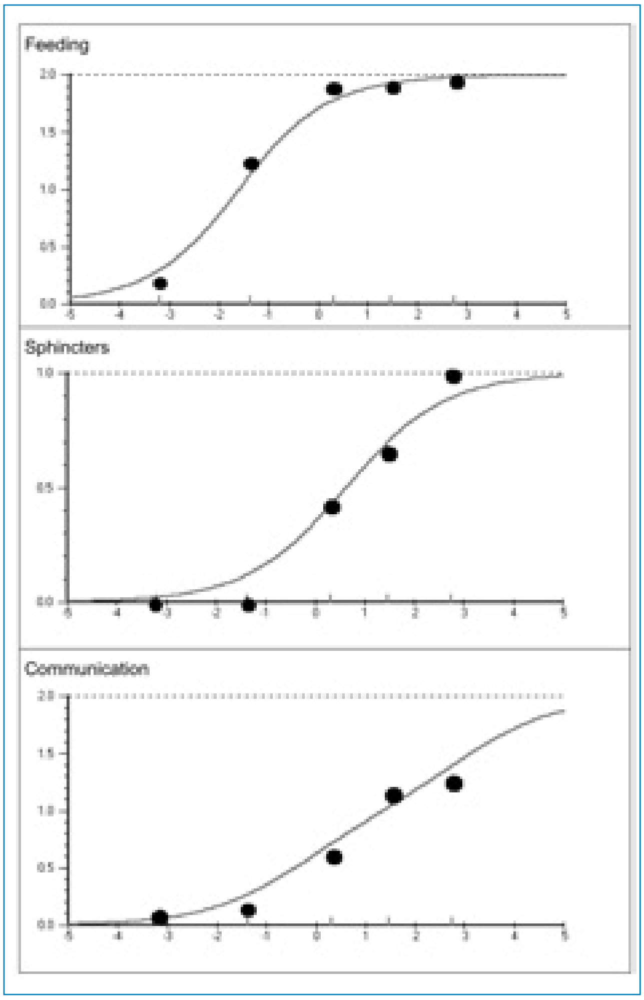
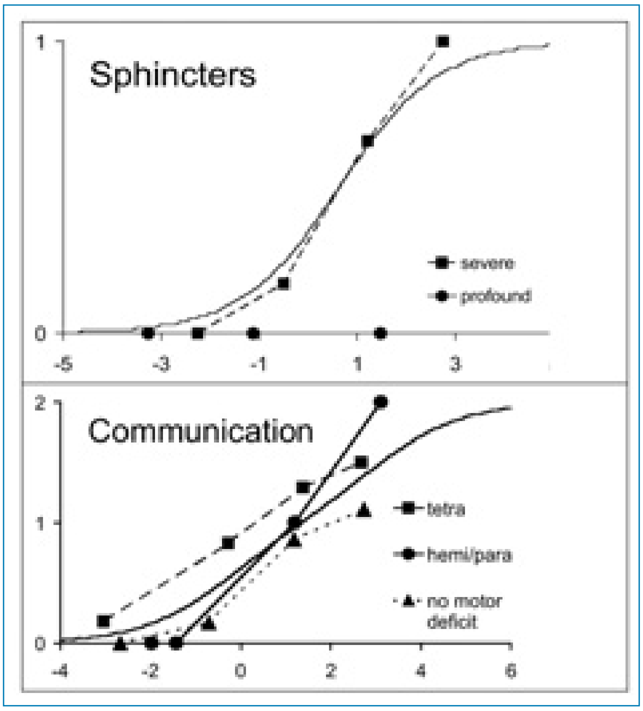


Fig. 4 – "Differential Item Functioning-DIF": Rasch analysis across two items and clinical subgroups.



technicality to remember: a "0" difference means 50% pass probability. The higher the difference between subject's ability and item threshold, the higher the probability that the subject will pass the threshold.

At a glance one can perceive various properties of the scale:

a) targeting (items are easy enough to be passed by even the most compromised patients);

b) spread (there are "ticks" for measuring either more or less affected patients);

c) density (ticks are dense enough to allow precision in measurement along various levels of abilities perhaps with a "gap" around 2 logit).

The LAPMER paper contains an Appendix introducing the reader to the various technicalities of the analysis. Here, we focus on the DIF issue alone.

Figure 3 refers to three LAPMER items, i.e. feeding (top panel), sphincters (middle panel) and communication (bottom panel). The abscissa gives, in logit units, the ability level of the patients (one can also imagine the total score, for simplicity), while the ordinate gives the Rasch-expected score in each particular item. The model prediction is given by the S-shaped curves. Intuitively enough, as long as the overall ability increases, the score in each particular item also increases. Dots give the average expected scores in classes of patients, i.e. across "quintiles" of ability levels (0-20%, 21-40% etc. of the ability distribution). The ideal location of the dots is on the model curve. Setting aside any formal statistical testing (strongly implemented in Rasch softwares), looking at the middle panel one can perceive visually that the "sphincter" ability of low-performers is underestimated. The 2nd quintile class gets, on average, a score much lower than that predicted by its overall ability level: here is a relevant DIF. Apparently, no DIF emerges from the top and the bottom panel. Now look at Figure 4. In the upper panel, the subjects are further sub-classified depending on the clinical assignment to severe or profound cases (the latter representing the worst condition). The DIF already visible in Figure 3 can now be confidently ascribed to a rough underestimation of ability in the "profound" class only, whatever the ability level.

The interpretation is that the raters did not properly apply the questionnaire. Incontinence may come from epilepsy or spasticity in these patients, not only from "mental retardation", as it was described above. Raters declared that they were biased, nonetheless, by the finding of wet diapers or linen. The item "communication " also shows a worrying DIF, once the subjects are further sub-classified depending on their motor comorbidity (in-

© Punto Effe, 2006

creasing severity from hemi- to para-, to tetra-plegia). The raters admitted that they tended to be much more lenient towards "mental retardation" when facing severe motor syndromes. They had a propensity to "credit" the patient with higher communication skills, somehow "discounting" a portion of the communication deficit that they ascribed (arbitrarily) to the motor comorbidity (e.g. slurred speech, shortness of breadth etc.).

The various "DIF" affecting LAPMER does not prevent it from being useful, and it remains a valid instrument for measuring "disability" patients so severely affected. Yet, the DIF findings provided powerful clues to the improvement of raters' training.

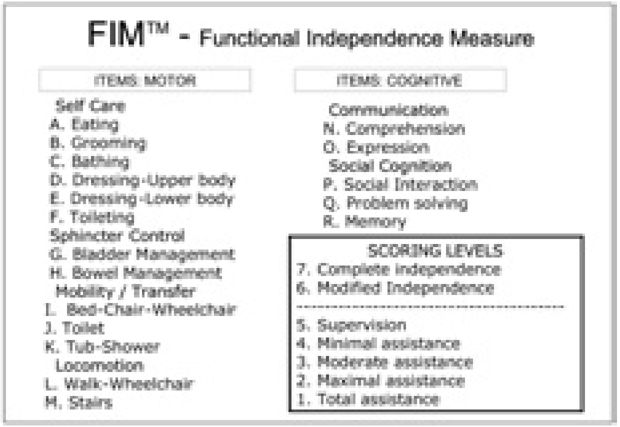
In general, "how much DIF is too much" is not an all-or-none issue. Interpreting Rasch output requires a strong interaction between statistical skills, clinical knowledge and common sense.

b) Item hierarchy may be unstable across classes of subjects (differential test functioning, DTF).

A DIF like the one depicted in Figures 3 and 4 does not give an insight on the overall hierarchy of the items. The subtle DIF shown by "Sphincters" for the few "profound" patients may or may not distort the hierarchy of average item difficulties. If only average item difficulty is considered, one can "zoom-out" and see whether the overall ruler stays the same across classes of patients.

Figure 5 shows the FIM™ questionnaire, the most famous questionnaire for measurement of independence in daily activities. The FIM™ is adopted world-wide. National data-banks exist both in USA and Italy[20]. The FIM™ score is a valid index of both appropriateness and effectiveness of the inpatient stay in hospital rehabilitation units. For this reason, it is used for quality control of the care processes[21] and, in USA, even for federal payment of MEDICARE-covered discharges (see www.udsmr.org).

Fig. 5 – The FIMTM (Functional Independence Measure) scale: items and score.



FIM™ - Functional Independence Measure

ITEMS: MOTOR	ITEMS: COGNITIVE
Self Care	Communication
A. Eating	N. Comprehension
B. Grooming	O. Expression
C. Bathing	Social Cognition
D. Dressing-Upper body	P. Social Interaction
E. Dressing-Lower body	Q. Problem solving
F. Toileting	R. Memory
Sphincter Control	SCORING LEVELS
G. Bladder Management	7. Complete independence
H. Bowel Management	6. Modified Independence
Mobility / Transfer	
I. Bed-Chair-Wheelchair	5. Supervision
J. Toilet	4. Minimal assistance
K. Tub-Shower	3. Moderate assistance
Locomotion	2. Maximal assistance
L. Walk-Wheelchair	1. Total assistance
M. Stairs	

Fig. 6 – "Differential Test Functioning-DTF": Rasch analysis on stability of item difficulty levels across two times of measurement. Difficulty of FIMTM items (see Figure 5) at admission (abscissa) and discharge. Two-hundred cases from an inpatient rehab unit within a general hospital.

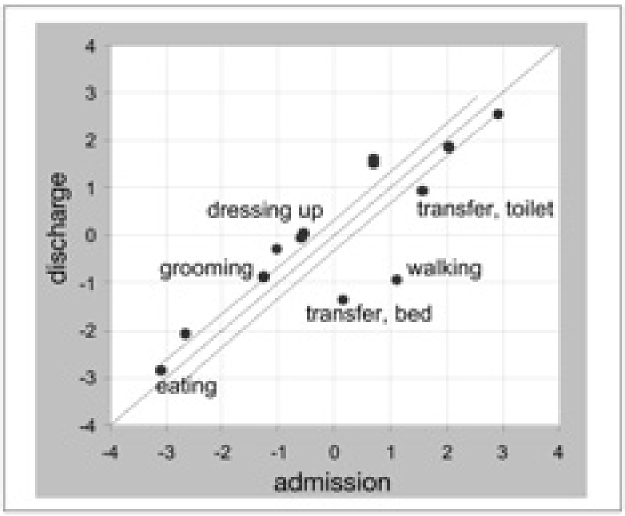


Figure 6 refers to 200 disabled patients (about 65% orthopaedic, 35% neurologic) discharged from a rehabilitation unit within a general hospital in Italy (courtesy of So.ge.com srl, Milan, www.so-ge-com.it).

The axes give the difficulty of the "motor" FIM items (1 to 13, see Figure 5) in Rasch linear logit units (the higher, the higher the item difficulty). The abscissa and the ordinate refer to admission and discharge, respectively. Patients do find all items easier at discharge, compared to admission (they get higher scores in all items). This notwithstanding, the intrinsic difficulty of the items (filled dots), relatively to each other, should stay the same. Hence, one should expect that their relative position along the difficulty continuum stay unchanged. For instance, "Eating" must remain easier than "Walking", and by the same amount, either at admission or at discharge. The identity line (and the 95% confidence limits) is drawn. It can be seen that "Transfer, bed", "Walking" and "Transfer, toilet" are unstable, in that they are relatively more difficult at admission, compared to discharge. For instance, "reading" the FIM items on the abscissa shows that "Transfer, bed" is more difficult than "Grooming", while the opposite comes out when reading the same items on the ordinate. Looking at DIF on the whole set of items simultaneously (so called Differential "Test" Functioning, DTF) provides a profile that allows useful interpretations. In this case, the DTF flags the behaviour of rehabilitation units forced to admit patients from acute care units within the same facilities, to allow them to accelerate their turn-over.

Such a DTF is much less pronounced in free-standing facilities who can better "negotiate" their admissions. Often, in the hospital units patients are transferred when they are not yet ready for rehabilitation exercises, and thus are prevented from leaving their bed, not because of severe disability but because of clinically unstable conditions (e.g. fever, post-surgical anaemia) and/or because of ongoing diagnostic procedures (e.g. X-ray controls on fractures, CT scans for previous brain bleeding etc.). At discharge, of course, the difficulty profile of the items changes.

DTF thus gives a "fingerprint" of the care process and signals that the low admission score, in this case, might give rise to FIM increments that do not only reflect effectiveness of the rehabilitation process, but a different mission of the unit at admission, compared to discharge.

An emerging application of DIF analysis: cross cultural studies.

DIF is a sort of measure of the stability of "meaning "of person measures across classes defined according to the most various criteria. An emerging application is the so-called cross-cultural validation. International multicentric studies are often based on questionnaires originally developed in one language and/or in one Country. Refined translation protocols[22] may warrant semantic, not metric equivalence.

In the former example of the Japanese patients, the FIM item "Eating", as well as the scoring procedures, may well be perfectly translated into their Japanese counterparts, yet the intrinsic difficulty of the item remains distinct in either context.

Sophisticated statistical techniques now exist, allowing to assign distinct difficulty values to the "diffing" items, depending on the classes to which they are applied, provided that a robust core set of shared items remains free from "diffing". This paves the way for building centralized "item banks" with items calibrated according to the context/class of application[23]. The construction of valid questionnaires may require year-long efforts and costly resources. Questionnaires are more and more required in clinical research, as long as the whole-person outcomes need to be measured. Therefore Rasch Analysis is a very promising frontier of medical research, fostering communication between bio-medicine and clinical medicine.

Addresses for correspondence:
prof. Luigi Tesio
Istituto Auxologico Italiano, IRCCS
Unità Clinica e Laboratorio di Ricerche di Riabilitazione Neuromotoria
via Mercalli, 32 - 20122 Milano, Italy
e-mail: l.tesio@auxologico.it

© Punto Effe, 2006

References

1. Tesio L. Bio-medicine between science and assistance. Rehabilitation Medicine: science of assistance (in Italian). Il nuovo Areopago 1995; 2:80-105
2. Israel G. The living machine. Bollati Boringhieri, Torino 2004 (in Italian)
3. Tesio L. Functional assessment in Rehabilitation Medicine (in Italian). Encycl. Méd. Chirurg. Italian Edition, Elsevier Italia, 2005 Issue 26; 030; B-10, 6pp
4. Thurstone LL. Attitudes can be measured. Am J Sociol 1928;33;529-554
5. Tesio L. Special report. Measuring person's behaviours and perceptions: Rasch Analysis as a tool for rehabilitation research. J Rehabil Med 2003;35:1-11
6. Haigh R, Tennant A, Biering-Sørensen F, Grimby G, Marincek C, Phillips S, Ring H, Tesio L, Thonnard JL. The use of outcome measures in physical medicine and rehabilitation within Europe. J Rehabil Med 2001; 33,6:273-278
7. Tesio L Bridging the gap between Biology and Clinical Medicine. Some help from Rasch measurement theory. J Appl Meas 2004; 5,4:362-366
8. Rasch G. Probabilistic models for some intelligence and attainment tests.1960. Danish Institute for Educational research. Expanded edition with foreword and afterforeword by Wright BD: 1980. The University of Chicago Press. Reprinted in 1993 by MESA Press, Chicago
9. Andrich D. Rasch models for measurement. 1998. Sage Publications, Newbury Park-CA
10. Linacre JM. Understanding Rasch measurement: estimation methods for Rasch measures. J Outcome Meas 1999; 3,4:382-405
11. Wright BD, Linacre JM Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil 1989; 70:857-860.
12. Libro bond and fox
13. libro penta
14. Linacre JM, Wright BD. Construction of measures from many-facet data. J Appl Meas 2002; 3,4:486-512
15. Smith RM. Fit analysis in latent trait measurement models. J Appl Meas 2000; 1,2:199-218
16. see ref. 8 and 9
17. Tesio L, Granger CV, Fiedler R. A unidimensional pain-disability scale for low back pain syndromes. Pain 1997; 69:269-278
18. 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
19. Tesio L, Valsecchi MR, Sala M, Guzzon P, Battaglia MA. Level of activity in profound/severe mental retardation (LAPMER): a Rasch-derived scale of disability. J Appl Meas 2002; 3,1:50-84
20. Tesio L, Granger CV, Perucca L, Franchingoni FP, Battaglia MA, Russell C. The FIMTM Instrument in the United States and Italy: a comparative study. Am J Phys Med Rehabil 2002, 81:168-176
21. Tesio L, Franchignoni FP, Battaglia MA, Perucca L. Quality assessment of FIM (Functional Independence Measure) ratings through Rasch Analysis. Eur Med Phys 1997; 33:69-78
22. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 2000; 25:3186-91
23. Tennant A, Penta M, Tesio L, Grimby G, Thonnard J-L, Slade A, Lawton G, Simone A, Carter J, Lundgren-Nilsson A, Tripolski M, Ring H, Biering-Sørensen F, Marincek C, Burger H, Phillips S. 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(1 Suppl):I37-48