J Rehabil Med 2003; 35: 105–115

Taylor & Francis
healthsciences

SPECIAL REPORT

MEASURING BEHAVIOURS AND PERCEPTIONS: RASCH ANALYSIS AS A TOOL FOR REHABILITATION RESEARCH

Luigi Tesio[1]

From the Department of Rehabilitation, Salvatore Maugeri Foundation, IRCCS, Pavia, Italy

Variables present in an individual, for example, independence, pain, balance, fatigue, depression and knowledge, cannot be measured directly (hence the term "latent" variables). They are usually assessed by measuring related behaviours, defined by sets of standardized items. The homogeneity of the different items, and proportionality of raw counts to measure, can only be postulated. In 1960 Georg Rasch proposed a statistical model that complied with the fundamental assumptions made in measurements in physical sciences. It allowed for the transformation of the cumulative raw scores (achieved by a subject across items, or by an item across subjects) into linear continuous measures of ability (for subjects) and difficulty (for items). These 2 parameters, only, govern the probability that "pass" rather than "fail" occurs. The discrepancies between model-expected scores (continuous between 0 and 1) and observed scores (discrete, either 0 or 1) provide indexes of inconsistency of individual subjects, items and classes of subjects. In subsequent years the same principles were extended to rating scales, with items graded on more than 2 levels, and to "many-facet" contexts where, beyond items and subjects, multiple raters, times of administration, etc. converge in determining the observed scores. Rasch modelling has increasing application in rehabilitation medicine. New scales with unprecedented metric validity (including internal consistency and reliability) can be built. Existing scales can be improved or rejected on a sound theoretical basis. In clinical trials the consistency and the linearity of measures of either subjects or raters can be validly matched with those of physical and chemical measures. The stability of the item difficulties across time, cultures, diagnostic groups and time of administration can be estimated, thus making it possible to compare homogeneous measures or foster diagnostic procedures on the reasons for differential item functioning.

Key words: Rasch analysis, rehabilitation, disability evaluation, measure, outcome assessment.

J Rehabil Med 2003; 35: 105–115

Correspondence address: Luigi Tesio, Department of Rehabilitation, Salvatore Maugeri Foundation, IRCCS, via A. Ferrata 8, IT-27100 Pavia, Italy. E-mail: ltesio@fsm.it

Submitted September 9, 2002; accepted January 3, 2003

INTRODUCTION

Rasch analysis is a statistical approach to the measure of human performance, attitudes and perceptions. It is named after its inventor, the Danish mathematician Georg Rasch. He published his theory in 1960 (1) and died in 1980. Rasch analysis was conceived as a psychometric tool for use in the social sciences. In the last 10 years it has become increasingly applied to rehabilitation research. Yet, perhaps due to its originality and specific terminology, this approach to measurement is still felt to be quite cryptic and abstruse by the rehabilitation community. In order to understand the general principles of Rasch analysis, asking *what* to measure should precede asking *how* to measure. This article is focused accordingly.

MEASUREMENT IS NOT COUNTING, BUT IS AN ABSTRACT CONTINUUM

In real life we often observe phenomena related to individual objects or persons. These phenomena appear to us as discrete: they either happen or they do not. Measurement begins by counting these discrete observations, but in order to have "quantity" we need continuous linear measures. These allow us to understand what is happening and to predict what will happen (2).

To give an example, we can pick up a series of individual, discrete oranges from a market stall. We count the oranges, but we actually aim to achieve an abstract continuous entity, i.e. "weight". Weight, not numbers, is a true equal interval measure. We are confident that the step from 2 to 3 kg means as much increase in "weight" as the step from 3 to 4 kg (while adding 3 oranges each time does not warrant this uniformity). The curious thing is that weight does not exist as a tangible entity, whereas number of oranges does. Weight is a mental "construct". All kinds of measures are abstract continuous gradients that can all be represented by an infinite straight line along which the "quantity" of the variable grows from "less" to "more". Albeit invented, the measure is "objective". It remains constant (1 kg is 1 kg, 1 metre is 1 metre) across raters, subjects, time, etc., as long as it can be compared with reference to physical objects (e.g. the standard platinum metre for length) or events (e.g. the freezing and boiling points of water at sea level for temperature).

 ISSN 1650–1977

Table I. *A schematic questionnaire of "mobility". Items are aligned from easiest to most difficult in rightward direction. A "mobility" scale: no = 0; yes = 1*

(a) Expected pattern of responses across subjects and items

	Turns in bed	Sits	Stands	Walks	Climbs upstairs	Total
A	1	1	1	1	1	5
B	1	1	1	1	0	4
C	1	1	1	0	0	3
D	1	1	0	0	0	2
E	1	0	0	0	0	1
F	0	0	0	0	0	0

(b) Unexpected string of responses, due to an extraneous item

	Turns in bed	Sits	Stands	Speaks Italian	Walks	Climb upstairs	Total
A	1	1	1	0	1	1	5
B	1	1	1	1	1	0	5
C	1	1	1	0	0	0	3
D	1	1	0	1	0	0	3
E	1	0	0	0	0	0	1
F	0	0	0	1	0	0	1

(c) Unexpected string of responses due to an inconsistent subject's record

	Turns in bed	Sits	Stands	Walks	Climbs upstairs	Total
A	1	1	1	1	1	5
B	1	1	1	1	0	4
C	1	1	1	0	0	3
D	1	1	0	0	0	2
E	1	0	0	0	0	1
F	0	0	0	0	1	1

APPROACHING ABSTRACT CONSTRUCTS: ASSUMING THEIR RELATIONSHIP WITH PHYSICAL EVENTS

In human history measuring was first applied to either concrete or abstract objects (e.g. cattle or Euclidean geometric figures) placed or imagined outside the person. It was not until the 19th century (3, 4) that "psychophysics" dared to measure abstract human sensations. These, however, could still be related to quantifiable physical external stimuli (e.g. touch pressure, sound pitch). Anchoring to the external, physical world still warranted objectivity. Later on, the question became how could objectivity be achieved for abstract mental variables which cannot simply be assumed to be the consequence of known external physical stimuli. Such variables are, for example, intelligence, depression, suffering, attitudes and knowledge about various topics. A critical step was the acknowledgement that mental constructs can indeed be accessible to measurement, if they manifest themselves through external physical events (e.g. depression can become manifest through crying). Contrary to psychophysics, these events are assumed to be the consequences, not the causes, of their mental counterparts. The latter remain hidden inside the individual. This perspective gave rise in the early 20th century to modern "psychometrics" (5): questions replaced stimuli. The physical event (an observed behaviour, as well as a tick on a questionnaire) only allows one to infer the existence of these variables: hence the name "latent traits" or "latent variables". Their quantity can also be inferred from counts of physical events. For instance, ticking "yes" (a discrete event) to the question "does your back hurt sometimes?" can be assumed to reveal some "pain". Ticking "yes" to the question "do you feel excruciating pain in your back?" can be assumed to reveal yet "more" pain, but to an unknown extent (we only know that 2 "yes" responses indicate *more* pain than 1 "yes" response). Psychophysics is deterministic; a cause-effect relationship is assumed between stimulus and response. The latent trait approach is probabilistic, in that it implies inferences. The following pivotal consequences should be highlighted:

- The particular behaviours we observe are only a sample coming from a universe of infinite alternative behaviours manifesting the same construct. These represent quantifiable levels of the shared underlying construct.
- Cumulative scores are only a count of discrete observations/items, no matter which numbers we assign to them (0/1, 0/10, etc.). Any intermediate levels assigned to observations such as *pain while sitting = no/mild/moderate/severe = 0/1/2/3* also give rise to mere counting of alternative events (2 happens rather than 1, etc.), separated by unknown quantities. In concept, a no/yes decision is made whenever one category is selected rather than an adjacent one. The interval between 0 and 1 is not necessarily equal to the interval between 1 and 2, etc.
- The "latent variable" approach implicitly removes the distinction between "psychological" and "physical" variables, as long as the person as a whole, a "self", an "I" is postulated.

Physical behaviours are movements credited with a meaning. They always reflect something from within the individual (for instance, walking reflects the intention to walk, the capacity for space-time orientation, etc.). Therefore, in the author's opinion "psycho-metrics" should be more properly termed "person-metrics".

APPROACHING ABSTRACT CONSTRUCTS: ESTIMATING THEIR QUANTITY FROM THE COUNTS OF RELATED PHYSICAL BEHAVIOURS

The latent trait paradigm, in itself, does not provide cues to valid quantitative measurement. A further step is needed, i.e. a model relating counts of observations to an abstract linear continuum of "less to more". Table Ia shows a simplistic questionnaire purported to measure "mobility".

Each item is scored "no" = 0, "yes" = 1. Higher scores mean "more" mobility. More difficult items are aligned to the right of the easier ones. More able subjects are aligned above the less able ones. The matrix of "0" and "1" responses shows that difficult items are only passed by high-scoring subjects. Less able subjects only pass easier items. This "diagonal" pattern makes sense. This is because the properties of the questionnaire comply with at least some of the *theoretical requirements* of measurement: unidimensionality and additivity.

In Table Ib a new item ("speaking Italian") was added to the scale in an attempt to widen the measure. Speech implies some "mobility", but few would deny that it predominantly represents another construct, i.e. knowledge of that language. It may well happen that a patient is paralysed, yet he/she can speak the native language fluently. In the simplistic example of Table Ib the unexpectedness of the scores of patients B, D and F is easily ascribed to a lack of unidimensionality of the scale. Summing "speaking Italian" with the other scores does not imply adding mobility. As a result, another requirement of measurement, i.e. additivity, is also lost. In Table Ic "Climbing stairs" gets 2 points = difficult item; person F, who cannot do easy items ("Turn in bed", "Sit") was able to pass a difficult item. This is unexpected. Was this patient presenting with a mixture of interfering "constructs"? (e.g. was he/she hysterical, or malingering, beyond presenting with mobility deficits? Were the scores miscoded when entered into a computer?). Again, unidimensionality was lost and additivity was severely challenged. We know that care must be taken in ascribing to this particular subject the amount of overall "mobility" usually implied by a cumulative score of 1.

FILLING THE GAP BETWEEN DISCRETE COUNTS AND CONTINUOUS MEASURES

The need for a continuous ruler

The ideal scale given in Table Ia is not a measure yet. First, scores are bounded between 0 and 5, but there may be subjects whose mobility would be better represented by even easier or more difficult items. Second, we do not know "how much more" mobility is revealed by the increasing scores, always advancing in arbitrary units of "1" (see the initial section, above). What we need is a ruler-like instrument like the one represented in Fig. 1.

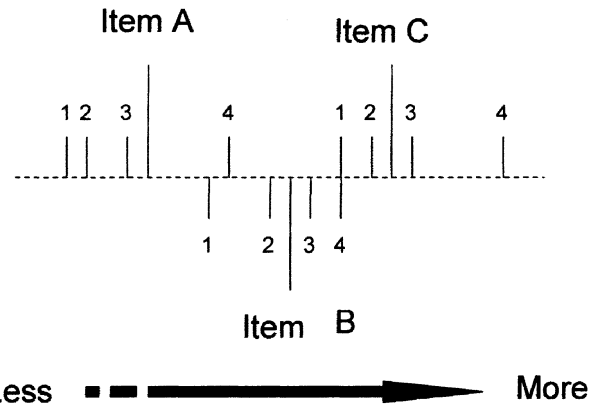

Fig. 1. The measure given by an item-response scale should resemble, in concept, a metric ruler. Items (A, B, C major ticks) represent quantitative levels along the same construct. Ordinal categories within each item (1, 2, 3, 4, minor ticks) represent more precise levels. Unlike in conventional metric rulers, raw integer scores can be spaced irregularly, because the measure they represent is not necessarily proportional. Also, "ticks" from different items may overlap. When the corresponding level of ability is reached several raw scores can be gained. In this case, raw scores overestimate the actual ability of the subject.

Now the item scores represent ticks along a continuous gradient, like the 1-cm ticks along a 1-metre ruler.

Stable ordering does not require that subsequent ticks be spaced evenly. For instance, going from "Turn-in-bed" to "Sitting" may mean advancing more, along the construct of "mobility", than going from "Sitting" to "Standing". Why are there minor ticks in the figure? The Rasch model can be extended to "rating scales" where each item is graded through ordinal categories. Take, for instance, a questionnaire of independence in daily activities (dressing, walking, etc.). Each item might be rated (again, through no/yes decisions) "with someone helping/with assistance/independent = 0/1/2", and the like. This justifies the analogy between mean item difficulty and item grades, on the one hand, and the cm and mm ticks in a ruler, on the other. Unlike for conventional length units, however, there are no reasons to assume that the minor ticks in an item-response scale are evenly spaced. Nor must it be assumed that the spacing pattern (whatever it is) is replicated across items.

Raw scores are misleading: more ticks do not always mean more information

In general, using more items and a rating scale (vs merely a pass/fail one) adds latitude (i.e. width), precision and sensitivity by increasing the range of available measures and the density of the marks. This is not always the case, however. Figure 2 comes from a study of the familiar Barthel Index of disability (7). Note that different items (or different grades from different items) may represent the same amount of the variable. This makes the

raw scores misleading. Depending on the overall ability of the subject, the same improvement of ability can lead to a very different raw score change, simply due to a local overlap between ruler "ticks".

Building up a continuous ruler from discrete counts: frequencies are not enough

Returning to the oranges as a representation of weight, a true measure occurs when the market scales transform the raw count of oranges into the measure of weight (kg). Unlike the count, the output of the market scales is continuous, so that "quantity" can be assessed. How can a continuous transformation of the raw scores from an item-response scale be achieved? The basic idea is to estimate the "difficulty" of an item from the frequency (i.e. the count) of people able to pass that item and the ability of a subject from the frequency of items he/she can pass. Let us consider the difficulty parameter (reasoning is symmetrical for the ability parameter). A given sample of 0/1, pass/fail items is administered to a sample of subjects. Item scores (the individual count of "fail") will show a given frequency distribution. For each item the "fail" frequency can be converted into a fail/pass proportion. Proportions are a continuous entity. The inference can be made that proportions represent the general probability that an item is failed. Another strategy may be one of turning frequencies into z scores (the number of standard deviations a given frequency is away from the mean frequency) (5). The z scores are also a continuous, generalizable entity. These solutions are not satisfactory, however.

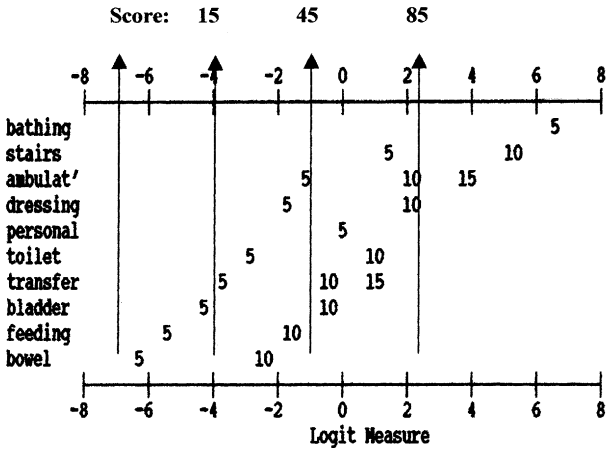

Fig. 2. The sum of raw scores (ticks in Fig. 1) can be a misleading pseudo-measure. This graph refers to 192 patients admitted to a rehabilitation unit (7). Items of the Barthel Index of disability are aligned from bottom to top in order of increasing difficulty. The ordinate gives the overall ability required to reach the ordinal levels (0/10, or 0/10/15) foreseen for each item. The 4 vertical lines encase 3 equal intervals of ability, yet corresponding to unequal cumulated raw scores. For instance, advancing from score 0 to score 15 requires the same overall improvement in ability required to advance from score 45 to score 85. Wrong conclusions about patients' improvements may be attained by looking at raw scores.

A CONTINUOUS RULER IS NOT NECESSARILY LINEAR, NOR OBJECTIVE

Unfortunately, obtaining a continuous ruler is not enough. The linearity of the measures must be achieved. Proportions are continuous, but they are bound between 0 and 1. At the extremes, subjects with different abilities tend to be ascribed the same probability for "fail" or "pass". The z scores are not linear, either (they are not proportional to the frequencies observed). As will be shown later, "logit" transformation provides the solution to linearization of proportions. Let us now focus on the "objectivity" issue. An item measure is "objective" if it is independent of the particular sample of subjects tested. A subject's measure is objective if it is independent of the particular set of items administered (so-called "separation" of items and subjects). If, say, the distance (i.e. the difference in difficulty) between "Climb upstairs" and "Walking" is 4 times the distance between "Walking" and "Standing", this must hold true for any subject of whatever ability. Of course, an able subject will have a higher probability of climbing upstairs compared with a less able subject. But if the ruler remains the same, both of them will find the step from "Walk" to "Stairs" 4 times "longer" than the step from "Stand" to "Walk". The relative difficulty of the items must not change. Items must be difficult *per se*, they must be separated from the sample of examinees. This property was never attained before Rasch. In any other model (including those relying on z scores) the particular distribution of people's abilities does influence the estimate of relative difficulty of the items.

THE RASCH SOLUTION: PROBABILITIES ARE DICTATED BY A MODEL IMPOSED ON THE DATA

The Rasch model reverses the traditional view of data-model relationship. Data must conform to the model, it is not the model that strives to "explain" the data (this is often referred to as a *prescriptive* vs a *descriptive* approach). The Rasch model is a theory of how probabilities of response *should* be, in order to comply with fundamental requirements of measurement. Then, the observed frequencies of response are compared with the expectations. Of course, differences between observed and expected scores ("residuals") will be found. If these are not too large, it is said that "the data fit the model", and the estimate of item difficulties and subjects abilities are said to be "likely". The Rasch model is a rule, formalized in a simple equation, stating how the probability of passing a no/yes item should change as a function of 2 "parameters" most often called subject ability (β) and item difficulty (δ). It is defined as a *1-parameter logistic (1-pl)* model, because conventionally item difficulty is not counted.

For the Rasch's *original dichotomous model* (no/yes, 0/1 items) the equation is

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

It is read as: the probability (P) that response (X) is observed to

be 1, given that (|) it may be either 0 or 1 (0, 1), depends on subject's ability and item difficulty, according to the relationship… (see the right side of the equation). The term "e" is the base of natural logarithms (2.718…). The exponentiation may look abstruse. In fact, the graphic representation (Fig. 3a) gives an intuitive S-shaped relationship between probability to pass rather than fail (on the ordinate) and ability (on the abscissa).

This shape recalls that probabilities cannot go either below 0 or above 1, whereas ability can be infinitely higher or lower compared with the difficulty of an item.

Furthermore, the elegance and strength of the model become evident once the equation is rewritten, after simple algebraic manipulation, as:

$$\ln\,(\mathrm{P}/(1-\mathrm{P})) = \beta - \delta \qquad [2]$$

where (1– P) is obviously the probability for fail.

Let us now define:

$$\ln\,(\mathrm{P}/(1-\mathrm{P})) = \mathrm{logit}$$

Logit stands for *logarithm of the odds (log-odds*) where odd = P/(1–P).

In the particular case where $\delta = 0$, Eq. [2] simplifies to

$$\mathrm{logit} = \beta \qquad [3]$$

When exponentiation disappeared on the right side of the equation, a linear measure of the variable arose. The greater the subject's ability, the greater the measure. The term "logit" is derived from the root of the adjective "logistic" and is given to the S-shaped function of Fig. 3a. In Rasch models the logit actually measures a difference, a local distance (e.g. between subjects, between items or between ability and difficulty, as in Eq. 1). Zero is conventionally assigned to the average difficulty of the items, so that one number only is sufficient to represent a measure. Abilities of 0, 1, 2 and 3 logits correspond to 50%, 73%, 88% and 95% probabilities of passing an item with 0 logit difficulty. To capture the clinical meaning of this unit, one may consider that, at discharge from post-acute inpatient rehabilitation, stroke patients show average improvements in independence, compared with their condition at admission, as measured by the FIM scale (see below, Table IV), in the order of magnitude of 1 logit (unpublished observation).

The *rating scale model* (6) applies to items scored on more than 2 levels (e.g. 0/1/2/3). Increasing categories must imply "more" of the variable. The more a subject is able, compared with an item, the more any higher category must become probable, compared with the previous one. The ability level at which 2 adjacent categories are equally probable are called "thresholds" or "steps", depending on the different authors. The model equation becomes:

$$\ln\,(\mathrm{P_{nik}}/(1-\mathrm{P_{nik}})) = \beta_\mathrm{n} - (\delta_\mathrm{i} + \tau_\mathrm{k}) \qquad [4]$$

where β_n indicates the ability of person n, δ_i indicates the average difficulty of item i, and τ_k indicate the difficulty of the k_{th} threshold (same for all items). If the pattern of threshold difficulties changes across items (so-called *partial credit model*) the equation becomes:

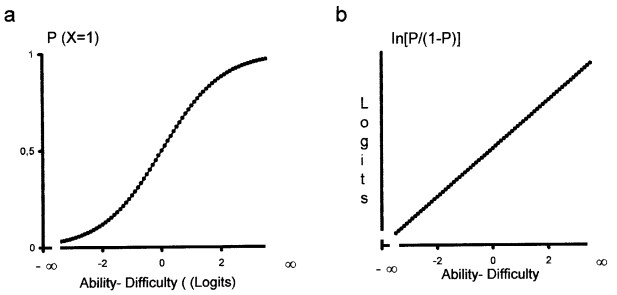

Fig. 3. (a) The "logistic" relationship between the probability that an item is passed rather than failed, foreseen by the Rasch model. The probability (on the ordinate) is only a function of the difference between subject's ability and item difficulty. (b) The relationship becomes linear if the probabilities are turned into logits (the natural log of the odd, i.e. of the pass/fail ratio).

$$\ln\,(\mathrm{P}_{nik}/(1-\mathrm{P_{nik}})) = \beta_\mathrm{n} - \delta_\mathrm{ik} \qquad [5]$$

Note that δ_ik replaced $(\delta_\mathrm{i} + \tau_\mathrm{k})$ from Eq. 4, because each pattern of threshold estimates is unique to the corresponding item and it is not estimated as a separate set of threshold applying to all items.

To sum up, what should be expected from data in the Rasch perspective? They should comply with the model. The observed frequencies of response should give rise, through dedicated computations, to the ideal β and δ values, for each subject and each item. "Ideal" here means that, reasoning the other way round, β and δ values are such that, once their relationship is the one prescribed by Eq. 1 (or Eq. 4 or Eq. 5, depending on the model adopted) they give rise for each response to the observed score, i.e. nearly 1 when 1 was observed, nearly 0 when 0 was observed. Why "nearly"? The model is probabilistic.

The pattern of responses depicted in Fig. 1a (so called Guttman-deterministic, "too good to be true") is not "ideal" in this context. First, certainty for either "pass" or "fail" is never expected (1 and 0 are asymptotes). Second, each prediction is an estimate, surrounded by uncertainty (the Bernoulli variance for dichotomous independent events).

THE RASCH SOLUTION TO OBJECTIVITY: ONLY SUBJECT'S ABILITY AND ITEM DIFFICULTY MUST GOVERN THE RESPONSE

The question remains, as to why the Rasch model claims to be the only one generating objective measures, so that the data *must* comply with its prescriptions? This is because in his "separability theorem" (1) Rasch demonstrated that Eq. 1 is the only formulation allowing the item difficulties to be independent from the particular set of respondents, and the subjects' abilities to be independent from the particular set of items administered. This item-person separation is a fundamental requirement of measurement. The separation is also demonstrated for Eq. 4 and Eq. 5, which therefore belong to the family of Rasch models. The logistic formulation is essential. In fact, it gives lineariza-

tion of the probabilities (see legend to Fig. 3), yet it does not imply any assumptions about the sample distribution of either abilities or difficulties (unlike, for instance, a *probit* formulation based on normal distribution). However, perhaps an even stronger characteristic is that the parameters of ability and difficulty, *only*, through their linear difference, are assumed to cause the expected response. Why are 1-parameter logistic Rasch models the only ones that are objective? In Fig. 4 the ordinate gives the probability that a dichotomous 0/1 item is passed. The abscissa gives the item difficulty. Each curve (called Item Characteristic Curve) refers to a dichotomous item with a different difficulty (more difficult items to the right).

Note that the curves do not cross; they run parallel. This prevents the hierarchy of difficulty of the items from changing depending on the ability of the subjects. The instrument remains conceptually the same for any subject. Other logistic models can include 2 or 3 parameters (e.g. item-specific slope and intercept of the Item Characteristic Curves). For instance, a 20% minimum probability to pass (an intercept) can be ascribed to a given 5-choice item, to account for guessing. This will makes the curves cross, and the hierarchy of difficulty of the items change, depending on the overall ability of the subjects. Perhaps the scores of the individual people belonging to the sample being examined will be more thoroughly "predicted" by the model. In the mean time, however, the requirements of separation and objectivity will be violated and comparability of measures across persons will be challenged.

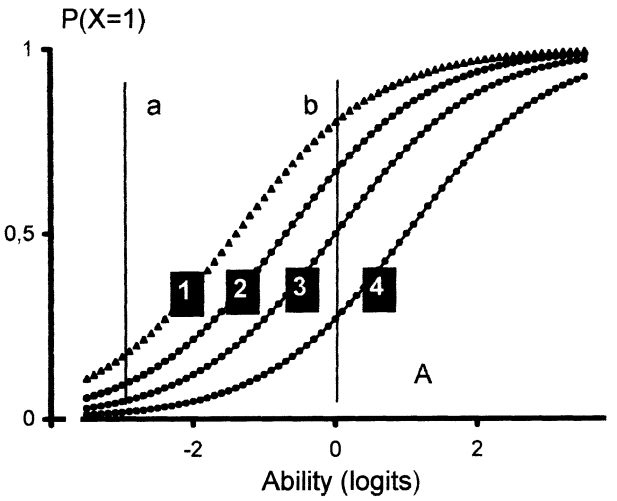

Fig. 4. In this example, the simple Rasch model was applied to an imaginary scale composed of 4 dichotomous items. The "Item Characteristic Curves" give, for each item, the probability for pass rather than fail (on the ordinate) as a function of the subject's ability (on the abscissa). More difficult items (1 to 4, in order of increasing difficulty) are shifted to the right. In the meanwhile, the Item Characteristic Curves always remain parallel. The vertical lines represent the ability of 2 subjects (a: less able; and b: more able). For both a and b the item hierarchy remains stable so that the construct, the very nature of the measuring instrument, is independent of the ability of the particular subject measured.

DISCOVERING THE MEASURE CONCEALED BY THE DATA

How can one "discover" in practice the ability and difficulty parameters from the data matrix? One should look again at Eq. 1 (but the principles are the same for Eq. 4 and Eq. 5). From the matrix (items × persons) of raw scores the logit values for β_n and δ_i should be extracted. For each person-item interaction, β_n and δ_i give rise to an "expected" score between 0 and 1. The mathematical algorithms for the "extraction" may be various, complex and computer-intensive. They belong to the family of "maximum likelihood estimation" (8). The parameters that one estimates are those that result in the smallest difference (becoming a "residual" after algebraic manipulations) (9) between the expected and the observed scores, and are thus said to be the "most likely" ones. Minimization of the residuals does not apply independently to any single response, but it is a compromise across responses, because the matrix as a whole must be the "most likely" one. One should remember that "most" likely does not necessarily mean "very" likely. The problem is how much the actual matrix of responses supports the inference that (in the universe) the extracted parameters hold. If high residuals remain after estimation, this indicates that real data do not fit well with the "most likely" model the data themselves suggest.

DO YOUR DATA JUSTIFY THE MODEL ESTIMATES? THE ISSUE OF "FIT"

How much do the observed scores justify the inference that the scale is a Rasch-consistent measure?

The accumulation of residuals gives rise to sophisticated summary indexes for individual subjects and individual items (9). These quantify the consistency (fit) of a whole string (across subjects and across items) of observed scores with the string expected by the model. The most common index in the literature is the "fit mean-square". A value of 1 indicates that the subject (or the item) overall has an expected pattern of responses; values between 0 and 1 indicate too a predictable pattern; values above 1 indicate too an unpredictable pattern. Note that some residuals are expected by this probabilistic model. Therefore, a value of 1 does not indicate absence of residuals, but that the amount of residuals is the one expected. When is a given "fit mean square" too large? There are no mandatory limits. Fit mean-square indexes can be assigned probabilities, so that significance levels (and thus, unexpectedness of the unexpectedness) can be adopted as an acceptance criterion. For instance, mean-squares in the range 0.7–1.3 are acceptable at $p < 0.05$, for sample sizes of about 100 records. Significance is highly dependent on sample-size, however, so such criteria should not be taken dogmatically. Given the acceptable range, "fit" indexes either above (misfit, too unexpected scores) or below (overfit, too predictable scores) should trigger diagnostic investigations. For instance, in case of misfit, through successive software runs one

Table II. *Fit of individual records to the Rasch model. Five typical patterns of response to an imaginary 18-item 0/1 questionnaire are depicted, with self-explanatory labels. The more difficult (farther right) is the item, the more "0" is the expected response. Within a probabilistic framework, some 0/1/0/1 sequence is "ideal" near the estimated ability of the subject ("pass" and "fail" are equally expected when an item captures the limit of subject's ability). A fit mean square value of 1 indicates that the expected (Bernoulli) variance was accumulated across the items. Values below and above 1, respectively, indicate lack of variance or excessive randomness, to be diagnosed*

Pattern	Items Easy to Difficult	Fit mean square
Modelled-Ideal	111101101101000000	1
Deterministic	111111111000000000	<1
Carelessness	011111111110100000	>1
Lucky guessing	111111110101000001	>1
Miscoding (reverse deterministic)	000000000111111111	$\gg 1$

0 = fail; 1 = pass.

Table III. *Rasch measure (logit transformed into 0–100 units) and fit mean square of the BACKILL scale for low-back pain syndromes, derived from existing scales (14). McGILL:McGill Pain Questionnaire—short form; OSWESTRY: Oswestry Low Back Pain Disability Questionnaire; FASQ: Functional Assessment Screening Questionnaire. Items made up a linear measure of "back illness". Fit indexes revealed good fit to the Rasch model*

Source scale	Item	Logit measure (transformed 0–100)	Fit mean square
McGill	Aching	60	0. 71
Oswestry	Lifting	57	1. 13
McGill	Tiring	56	0. 93
FASQ	Sitting	55	1. 07
FASQ	Standing	53	1. 30
Oswestry	Traveling	48	0. 79
FASQ	Low sit (raising from)	47	0. 92
Oswestry	Walking	42	0. 83
Oswestry	Personal care	34	0. 86

(Modified after Tesio *et al*. (14), Table 2).

can easily test the effect of removing items and/or subjects (does the subject's — or item's — fit to the model increase? Do the changes make clinical sense?). In rating scales with items graded along many ordinal categories (such as 0/1/2/3) one can also detect misfit of categories. Suppose category 2 = *with assistance* and 3 = *with help* are selected randomly by the patients, who do not perceive any substantive difference: these categories will accumulate large "residuals" across many subjects. By contrast, suppose that 0 = *can't do* and 4 = *independent* are clearly selected as a function of subject's overall ability. These categories will show responses very close to the ones expected. Perhaps the 4 categories can be collapsed by rescoring 0/1/2/3 into 0/1/1/2, without any substantive loss of precision in the estimations of ability and difficulty. In short, knowledge of the specific field and Rasch statistics can be used together very effectively in clarifying the "construct". Unexpected records also convey a lot of information. The experienced analyst knows that specific alterations of the fit indexes may flag specific respondent's behaviours (Table II).

THE GROWING FAMILY OF RASCH MODELS

Georg Rasch proposed his model for dichotomous scales (no/yes, 0/1). Since then, research has evolved, and is still in progress. The principles of the model were applied, as anticipated above, to "rating" scales with graded items (6). The model remains a 1-parameter logistic one. A "many-facet" model was also developed (10). For instance, when several raters apply the same test to a given sample of subjects, raters' severity and consistency may influence the observed response. If a rating scale model is adopted, the "many-facet" version takes the form:

$$\ln\left(P_{nikj}/(1-P_{nikj})\right) = \beta_n - (\delta_i + \tau_k + \gamma_j) \qquad [6]$$

where all symbols are as given in Eq. 4, and γ_j is the severity of rater j.

The model is still a linear, Rasch one. Raters' parameters are estimated independently from persons' and items' parameters, and represent an additional separate "facet" along the shared continuum. Raters are thus provided with "ability" and "fit" estimates. In fact, their "ability" is their "leniency": more "able" raters are those assigning, rather than receiving, higher scores to the same subjects on the same items. The analysis of fit may help to detect any rater-dependent bias in examinee's scores (10). Perhaps the most advanced example of "many-facet" application to rehabilitation is the Assessment of Motor and Process Skills (AMPS, 12, 13). This instrument provides a measure of quality of performances of the activities of daily living. People (facet 4) perform tasks (facet 1) scored on items (smaller units of activities of daily living; facet 2) rated by raters (facet 3). A large and ever-growing data bank provides stable estimates of item difficulty parameters. Against these anchor values, credentialed raters can be "calibrated" in severity, and their ratings adjusted correspondingly. This system allows the AMPS to provide ability measures generalizable world-wide (13).

RASCH MODELLING AND REHABILITATION RESEARCH: SOME EXAMPLES

Any item-response scale or questionnaire aiming at functional assessment can lend itself to Rasch analysis. In practice, the analysis has 2 main applications:

- the building and/or validation of scales
- the assessment of persons, either examinees or raters, once a scale is credited with established validity.

Scale building and validation

Table III is taken from the original publication of "BACKILL", a questionnaire for the assessment of back pain (14).

The study originated from the authors' experience that

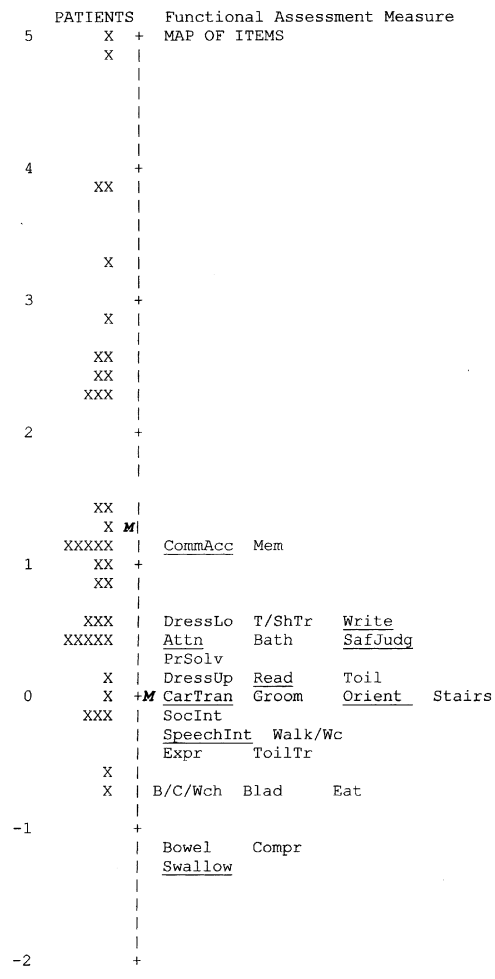

Fig. 5. A "Rasch ruler", also said an "item map". The figure refers to the application of the Functional Assessment Measure-FAM (Table IV) to 42 high-functioning brain-injured outpatients (16). The vertical line represents the measure of disability, with logit values given on the left. Patients' ability levels are represented as "X" symbols and aligned to the left of the corresponding measure. Items are aligned to the right of the corresponding values. Ten fitting FAM items added to the FIM are underlined. "M": mean value of subjects' abilities (left) and item difficulties (right).

satisfaction with treatments (reporting "I feel improved") cannot easily be confirmed ("objectified") on the basis of either pain-based or activity-based questionnaires. Are the patients unreliable? The authors hypothesized that "feeling better" is a latent variable, distinct from either pain or mobility, in which "feeling less pain" and "moving better" interact (for instance, you can feel less pain if you move less). A series of existing validated scales were collected and administered to patients with lumbar disc herniation who were undergoing a physical treatment (lumbar auto-traction). The most responsive items were collected from each of the source scales, and included in a "hybrid" scale, which was subjected to repeated runs of Rasch analysis. On the basis of the fit indexes, the records from misfitting items and subjects were recursively inspected, interpreted, and removed or retained. Nine items appeared to define a homogeneous construct of "back illness". Rasch analysis revealed that the scale was no longer "hybrid", although it included items of both pain or mobility. That is, a single latent trait emerged. This Rasch-built questionnaire worked well. In fact, an improvement in the score predicted the overall feeling of improvement with sensitivity and specificity above 0.85.

A promising field of application is the cross-cultural validation of scales. Linguistic equivalence does not warrant metric equivalence. For example, in a disability scale "Eating"—whatever its translation—can be much more difficult in east Asian countries where chopsticks are used, compared with western countries. A large European project was specifically dedicated to cross-cultural validation of questionnaires adopted in rehabilitation (15). Differences in the hierarchy of item difficulties were found across different countries even for instruments used worldwide such as the Mini-Mental State Examination and the Functional Independence Measure (FIM)[1]. This suggests that techniques for "equating" the

[1]Tennant A, Penta M, Tesio L, Grimby G, et al. A Rasch-modelling approach to cross-cultural validity of item-response scales through differential item-functioning: The Pro-ESOR project. Submitted manuscript.

Table IV. *The Functional Assessment Measure (FAM) (17) consists of the 18-item Functional Independence Measure (FIM™), with the addition of 12 items deemed to be more specific for brain injury outpatients. All items are scored 1–7, the higher the score, the higher the patient's independence or performance*

Motor FIM™ items	Cognitive FIM™ items	Newly designed FAM items
1. Eating	14. Comprehension	19. Swallowing
2. Grooming	15. Expression	20. Car transfer
3. Bathing	16. Social interaction	21. Community access
4. Dressing upper body	17. Problem solving	22. Reading
5. Dressing lower body	18. Memory	23. Writing
6. Toileting		24. Speech intelligibility
7. Bladder management		25. Emotional status
8. Bowel management		26. Adjustment to limitations
9. Bed, chair, wheelchair transfer		27. Employability
10. Toilet transfer		28. Orientation
11. Tub, shower transfer		29. Attention span
12. Walking, wheelchair		30. Safety judgement
13. Stairs		

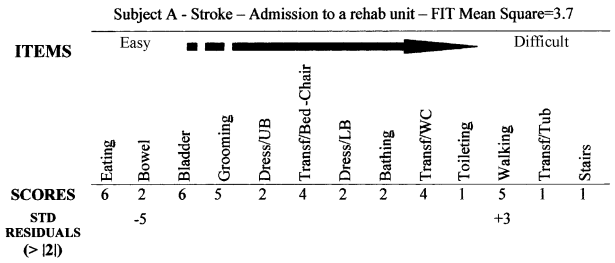

Fig. 6. The table refers to a person with stroke admitted to an inpatient rehabilitation unit. The 13 "motor" items of the FIM (Table IV) are aligned, left to right, in order of increasing difficulty. Scores may range from 1 to 7, the higher the score the greater the patient's independence, and are given below each item label. Scores are expected to decline from left to right. Below each score, the residuals (9) between the observed and the model-expected score (z-standardized and squared—STD) are given. Only residuals significant at $p < 0.05$ (i.e. > 2 or < -2) are represented. The most unexpected scores were observed in "Bowel" (too low) and "Walking" (too high). Across the various items, the patient mean square "fit" index is very high (3.7, expected 1), suggesting that the overall pattern of responses is unexpected. The reasons need to be diagnosed.

different measures (also available from the Rasch armamentarium) might be necessary in cross-cultural trials.

As another application, Rasch analysis allows one to re-test existing, validated instruments more strictly. This was the case for the Functional Assessment Measure (FAM) (16). The scale was designed specifically for measuring disability in high-functioning brain-injured outpatients, for whom the widely adopted 18-item FIM (Table IV) is too easy. In the attempt to raise the "ceiling" of the FIM and to capture more specific impairments of these patients, the authors added 12 new items (e.g. Entering a car; Speech intelligibility; Employability) and proposed that disability should be measured by cumulating the scores achieved in the 30 items.

The Rasch analysis demonstrated that the attempt was not successful. Figure 5 shows the classic "Rasch ruler" (also termed the "item map") obtained through the analysis (17).

Although a distinct analysis of the 13-item "motor" and the 5-item "cognitive" FIM sub-scales is recommended, the FAM authors' approach was followed (16). The analysis was thus conducted on the whole 30-item set. The vertical line represents the variable (disability in brain injured outpatients). The logit measure is given on the left of the figure. Items are aligned on the right of the line, the more difficult on top. Subjects are represented by crossed symbols on the left of the line, more able subjects on top. Underlined, 10 of the new items are shown (2 were deleted because of excessive misfit). It is clear that the new items share the same span of difficulty as the FIM items, which indicates potential item redundancy and the risk for inflation of the cumulative raw score when the scores of individual items reflecting the same level of ability are summated (see Fig. 2). Despite the greater number of items, the reliability indexes of the 28-item FAM were superimposable on those recorded from the 13-item "motor" FIM (not shown), thus confirming the redundancy of the added items.

Testing the quality of data from item-response scales

Individual records. A useful application of Rasch analysis is the control of data-model fit, for the string of responses given by each subject.

Figure 6 is taken, modified, from a software output after Rasch analysis on FIM data from 100 stroke patients at admission to a rehabilitation unit (unpublished observations).

Misfit indexes warned that this particular stroke patient gave an unexpected string of responses (fit mean square = 3.7, expected value 1). Given his/her overall ability revealed by the cumulative score of 41 (available range 13–91), the score of the item "Bowel" was unexpectedly low, while the score of the item "Walking" was unexpectedly high. Was this subject atypical for clinical reasons? Or, were the scores taken carelessly? Inspection of the individual clinical record and of other FIM records taken by the same rater can clarify these issues. Is bowel care systematically overlooked by the nursing staff? Does the "misfit" of this response string reflect an intrinsic flaw of the FIM scale, where the item "Bowel" represents a construct extraneous to the one represented by the other items? Inspection of large record sets from several different sources might clarify these issues.

Patterns of caring procedures within the hospital unit. In Rasch language the term "differential item functioning" (DIF) indicates the instability of the hierarchy of item difficulty levels. The same scale may not be measuring exactly the same variable across groups. Figure 7 is a DIF plot relating to the FIM, when applied to healthcare management.

The difficulty parameters of the 13 motor items of the FIM at discharge (on the ordinate) are plotted against the parameters of the same items scored at admission. Conventionally, "0" is assigned to the average difficulty of items in either set, so that an identity line is expected (the diagonal). The model also allows one to estimate confidence interval of the distances between the item position and the identity line (95% in this case, lines aside the diagonal, see figure legend). Records come from a mix of 200 orthopaedic and stroke patients admitted to an inpatient rehabilitation unit (unpublished observation). Some items lie outside the confidence "band". To the right of the identity line, one finds items that are more difficult at admission than at discharge, compared with the other items. These are "Upper body dressing", "Transfer to chair and WC", "Locomotion (by walking or on wheelchair)", and "Stairs". These items depict the capacity to move out of bed. Difficulty is a relative matter so that some other items (to the left of the identity line) appear easier (rather more difficult) at admission, compared with at discharge. In the author's experience this pattern flags a "vicarious nursing" typical for admissions made too early from acute wards (which was found to be the case). Patients are prevented from moving around as much as they could, because they are confined to bed for biomedical reasons (e.g. stabilization of fractures, optimization of coagulation parameters, waiting for radiological controls etc.). Once this restriction is removed, these items show a dramatic improvement and resume their proper hierarchy. This

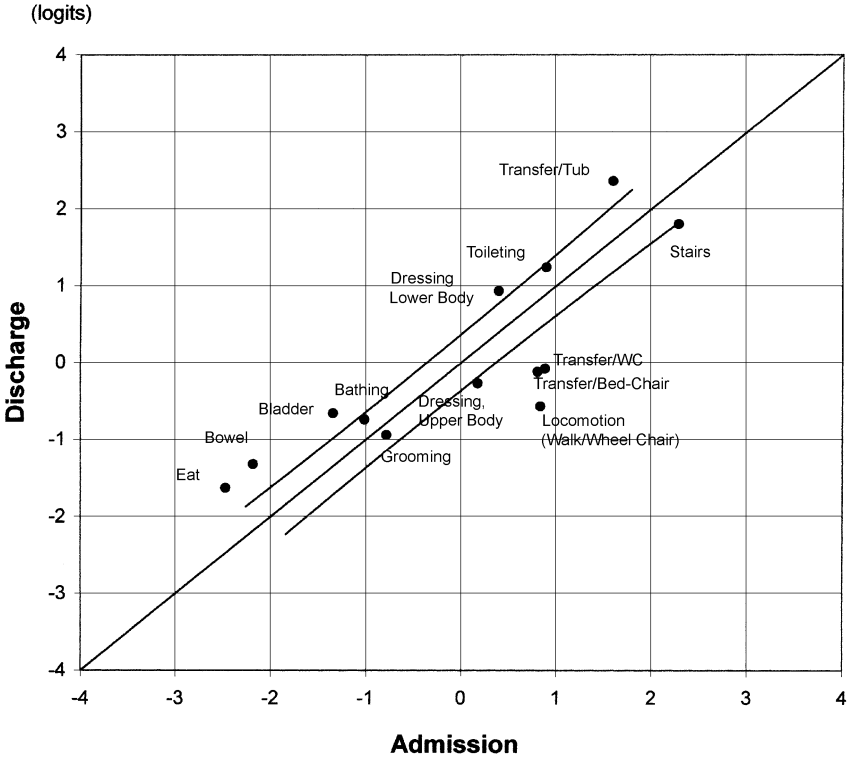

Fig. 7. Detection of "differential item functioning" in a data set. The figure refers to a mix of 200 orthopaedic and stroke inpatients admitted to an inpatient rehabilitation unit (personal data). The FIM scores were recorded at admission and discharge. The item difficulties at discharge are given on the ordinate, as a function of their values at admission, given on the abscissa. Conventionally, "0" is assigned to the mean item difficulty on both occasions. Items are represented through dots. The diagonal represents the identity line. This is paralleled by 95% confidence lines, computed according to Wright BD and Stone M, 1997, pp. 94–95 (see the "further suggested readings" section). The reasons need to be diagnosed.

DIF plot warns that gains in FIM scores between admission and discharge can be partly artefactual, due to the admission policy. The DIF analysis is a general approach which can help testing the stability of item hierarchy across classes of observations defined according to the most various criteria, e.g. time of administration, diagnostic category, age group, language, etc.

CONCLUSION

The principles, methods and applications of Rasch analysis go far beyond those depicted here (see references for more hints). Rasch analysis is not panacea (no model is). Its theoretical framework, however, is the first one removing the distinction between "psychological" and "physical" variables, in favour of a unitary approach to the study of person's behaviour. This is perhaps of minor relevance for statisticians, but it is of the utmost importance for the clinical community. Rehabilitation medicine is deeply rooted in bio-medicine, which derived its contemporary strength from physical sciences (18). In the meantime, rehabilitation medicine ultimately aims at restoring behaviours and perceptions (such as independence, balance, continence and fatigue), through behaviours (such as exercise, teaching, counselling and functional assessment). Once this amazing potential of the method is captured, it is difficult to renounce the attractive scenarios it discloses to the discipline. The same rigorousness and power belonging to the physical sciences and biostatistics are now within the reach of rehabilitation medicine, in a version tailored to the scientific needs of a person-oriented, holistic specialty. This may help to accelerate the relatively slow development of research claimed for this discipline, compared with organ-oriented ones (19).

ACKNOWLEDGEMENTS

Nowadays it is still difficult to find a written systematic explanation of Rasch theory applied to rehabilitation medicine. This was even more difficult when the author approached this exciting field, 10 years ago. Therefore, much of what this article strives to communicate comes from a countless series of informal discussions, personal communications, lectures, face-to-face teaching/learning meetings, and classroom lessons, which have no written citable counterparts. The author can only declare that he is indebted and grateful to David Andrich, Carl Granger, Mike Linacre, Gunnar Grimby, Massimo Penta, Richard Smith, Alan Tennant and Ben Wright.

REFERENCES

1. Rasch G. Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research; 1960. Expanded edition with foreword and afterforeword by Wright BD. The University of Chicago Press; 1980. Reprinted in 1993 by MESA Press, Chicago.
2. Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil 1989; 70: 857–860.
3. Weber EH. De pulsu, resorptione, auditu at tactu. Annotationes anatomicae et physiologicae. Koeler, Lipsiae. 1834.
4. Fechner GT. Elemente der Psycophysik. Leipzig: Breitkopf & Hartel; 1860.

5. Thurstone LL. Attitudes can be measured. Am J Sociol 1928; 33: 529–554.
6. Andrich D. Rasch models for measurement. Newbury Park, CA: Sage Publications; 1998.
7. Tennant A, Geddes JML, Chamberlain MA. The Barthel Index: an ordinal score or an interval level measure? Clin Rehabil 1996; 10: 301–308.
8. Linacre JM. Understanding Rasch measurement: estimation methods for Rasch measures. J Outcome Meas 1999; 4: 382–405.
9. Smith RM. Fit analysis in latent trait measurement models. J Appl Meas 2000; 2: 199–218.
10. Linacre JM, Wright BD. Construction of measures from many-facet data. J Appl Meas 2002; 4: 486–512.
11. Lunz ME, Stahl JA. The effect of rater severity on person ability measure: a Rasch model analysis. Am J Occup Ther 1993; 47: 3111–3317.
12. Fisher AG. The assessment of IADL motor skills: an application of many-faceted Rasch analysis. Am J Occup Ther 1993; 47: 319–329.
13. Fisher AG. Assessment of motor and process skills (4th edn). Ft Collins, CO: Three Star Press; 2001.
14. Tesio L, Granger CV, Fiedler R. A unidimensional pain-disability scale for low back pain syndromes. Pain 1997; 69: 269–278.
15. Haigh R, Tennant A, Biering-Sørensen F, Grimby G, Marincek C, Phillips S, et al. The use of outcome measures in physical medicine and rehabilitation within Europe. J Rehabil Med 2001; 33: 273–278.
16. Hall KM, Hamilton BB, Gordon WA, Zasler ND. Characteristics and comparisons of functional assessment indices: Disability Rating Scale, Functional Independence Measure, and Functional Assessment Measure. J Head Trauma Rehabil 1993; 8: 60–74.
17. 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: 79–96.
18. Tesio L. Bio-medicine between science and assistance. Rehabilitation medicine: the science of assistance. (In Italian). Il Nuovo Aeropago (Forli, Italy) 1995: 80–105.
19. Tesio L, Gamba C, Capelli A. Rehabilitation: Cinderella of neurologic research? A bibliometric study. Ital J Neurol Sci 1995; 16: 473–477.

FURTHER READING

Books

- Bond TG, Fox CM. Applying the Rasch model. Fundamental measurement in the human sciences. Mahwah, New Jersey: Erlbaum Publishers; 2001.
- Embretson SE, Hershberger SL (eds). The new rules of measurement. Mahwah, New Jersey: Erlbaum Publishers 1999.
- McNamara T. Concepts and procedures in Rasch measurement. Ch. 6, pp. 149–181. In: McNamara T. Measuring second language performance. Harlow: Longman Publishers 1996.
- Wright BD, Masters GN. Rating scale analysis. Rasch measurement. Chicago: MESA Press; 1982.
- Wright BD, Stone MH. Best test design. Rasch measurement. Chicago: MESA Press; 1979.

Articles

- Andrich D. Understanding resistance to the data-model relationship in Rasch's paradigm: a reflexion for the next generation. J Appl Meas 2002; 3: 325–359.
- Campbell SK, Kolobe TH, Wright BD, Linacre JM. Validity of the Test of Infant Motor Performance for prediction of 6-, 9- and 12-month scores on the Alberta Infant Motor Scale. Dev Med Child Neurol 2002; 44: 263–272.
- Chang W-C, Chang C. Rasch analysis for outcome measures: some methodological considerations. Arch Phys Med Rehabil 1995; 76: 934–939.
- Embretson SE. The new rules of measurement. Psychol Assess 1996; 8: 341–349.
- Fisher WP Jr., Harvey RF, Taylor P, Kilgore KM, Kelly CK. Rehabits: a common language of functional assessment. Arch Phys Med Rehabil 1995; 76: 113–122.
- Grimby G, Andren E, Holmgren E, Wright B, Linacre JM, Sundh V. Structure of a combination of Functional Independence Measure and Instrumental Activity Measure items in community-living persons: a study of individual with cerebral palsy and spina bifida. Arch Phys Med Rehabil 1996; 77: 1109–1114.
- Penta M, Tesio L, Arnould C, Zancan A, Thonnard JL. 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: 1627–1634.
- 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: 50–84 (with Appendix 1 on building scales through Rasch modelling).
- Wright BD, Mok M. Rasch models overview. J Appl Meas 2000; 1: 83–106.

Web site (including software information)

www.rasch.org.