Current issues in psychometric assessment of outcome measures

Aktualni problemi u psihometrijskoj analizi mjerenja rezultata

Franco Franchignoni[1*], Andrea Giordano[2], Lucia Marcantonio[3], Carlo Alberto Coccetta[4], Giorgio Ferriero[4]

Abstract. In recent years there has been an increasing use of outcome measures in clinical practice, audit procedures and quality control. The psychometric assessment of these measures is still largely based on classical test theory (CTT), including analysis of internal consistency, reproducibility, and criterion-related validity. But this approach neglects standard criteria and practical attributes that need to be considered when evaluating the fundamental properties of a measurement tool. Conversely, Rasch analysis (RA) is an original item-response theory analysis based on latent-trait modelling, and provides a statistical model that prescribes how data should be in order to comply with theoretical requirements of measurement. RA gives psychometric information not obtainable through CTT, namely: (i) the functioning of rating scale categories; (ii) the measure's validity, e.g. how well an item performs in terms of its relevance or usefulness for measuring the underlying construct, and the consistency of item difficulty compared with the expectations of the construct; (iii) the reliability, in terms of 'separation'; and (iv) the dimensionality of the scale and analysis of local independence of items.
For these reasons, RA is increasingly used and it was recently recommended as a method for assessing scale properties in addition to classical psychometric criteria for reviewing and assessing surveys and questionnaires for disability outcomes research.
The purpose of this paper is to provide some insights regarding the use of modern psychometric approaches such as RA for selecting and/or revising outcome measures in Physical and Rehabilitation Medicine.

Key words: Outcome assessment, Physical and Rehabilitation Medicine, psychometrics, Rasch analysis

[1]Department of Physical and Rehabilitation Medicine, Salvatore Maugeri Foundation, Clinica del Lavoro e della Riabilitazione, IRCCS, Genova-Nervi, Italy

[2]Unit of Bioengineering, Salvatore Maugeri Foundation, Clinica del Lavoro e della Riabilitazione, IRCCS, Veruno (NO), Italy

[3]Trainee in Physical Medicine and Rehabilitation, University of Turin, Turin, Italy

[4]Unit of Occupational Rehabilitation and Ergonomics, Salvatore Maugeri Foundation, Clinica del Lavoro e della Riabilitazione, IRCCS, Veruno (NO), Italy

Prispjelo: 2. 4. 2012.
Prihvaćeno: 18. 4. 2012.

Corresponding Author:
[*]**Franco Franchignoni, MD**
Fondazione Salvatore Maugeri, Clinica del Lavoro e della Riabilitazione, IRCCS
Via Missolungi 14, I-16167 Nervi (GE), Italy
e-mail: franco.franchignoni@fsm.it

http://hrcak.srce.hr/medicina

INTRODUCTION

Clinical practice, audit procedures, and quality control call for an increasing use of outcome measures. Physiatrists need to acquire specific expertise to be able to select the appropriate assessment tools, administer them thoughtfully, and interpret correctly the results[1].

Basically, an outcome measure is a tool to assess the magnitude of some longitudinal change (e.g. in impairment, functioning, activities, participation) in an individual or group[2]; in Physical and Rehabilitation Medicine (PRM) what is subject to change often is a 'latent trait', 'trait' meaning a hypothetical construct, domain, ability or other (e.g. functional independence, manual dexterity, locomotor capability) and 'latent' meaning that it cannot be measured directly but is 'hidden' within the person, who may manifest it through a set of behaviours indirectly assessed by a series of observations or questions (items)[3]. In order to be useful for their intended purposes, the rating scales and questionnaires measuring 'latent traits' must provide information that allows valid inferences and decisions to be made.

Classical test theory (CTT) is still widely used for validating these tools, in both their original and translated versions, as reports in peer-reviewed indexed journals show. These papers are based mainly on analysis of internal consistency, reproducibility, and criterion-related validity (usually the demonstration of a moderate to good correlation with some other measure of the trait under study). This approach, however, neglects standard criteria and practical attributes that need to be considered when evaluating the fundamental properties of a measurement tool[4], hence caution is needed when using ordinal scales and dealing with raw scores[3]. In fact, the numerical codes associated with each rating scale category ('0', '1', '2',...) do not necessarily imply proportionality among the measures (e.g. a subject with score '2' does not necessarily possess twice the amount of the latent trait with respect to a subject with score '1'). Moreover, using CTT item weighting is the same regardless of the difficulties or complexity inherent in the items (e.g. certain items are more difficult than others). Conversely, Rasch analysis (RA) – named after the Danish mathematician Georg Rasch – is an original item-response theory (IRT) analysis based on latent-trait modelling. RA provides a statistical model that prescribes how data should be in order to comply with theoretical requirements of measurement and it estimates, amongst other things, how much the modelled measure is supported by the actual observed scores (the so-called ''data-model fit''). Briefly, the model postulates that the probability of a person's response to each category of a rating scale item is governed only by the difference between two factors, which are calibrated simultaneously through an iterative process: the amount of latent trait possessed by the person (e.g. 'functional independence'), conventionally referred to as 'subject ability', and the amount of that trait analyzed by a given item, referred to as 'item difficulty'[5]. Thus, it is expected that a person with high levels of latent trait (e.g. more functional independence) will consistently use higher scoring response options than one with less functional independence. The Rasch model conceptualizes the hierarchy of 'item difficulty' and 'subject ability' resulting from the analysis of the ordinal response of each subject to each item like a ruler. If data fit the model, this ruler has the properties of an interval scale (i.e. it is linear and quantitative, which is particularly important when measuring change and responsiveness to treatment). RA provides also powerful diagnostic tools for scale functioning, item and person fit and dimensionality assessment.

For these reasons, RA is increasingly used and has been recently recommended as a method for

A series of validation methods (using both CTT and RA) must be applied to analyze the validity evidence of an outcome measure.

Unidimensionality of outcome measures is a core assumption of item response theory models and a prerequisite for any ensuing psychometric analysis.

Raw scores can be misleading and there is a potential for misinference when ordinal scales are used. If data fit the model, Rasch-transformed scores are at interval-scale level.

assessing scale properties in addition to classical psychometric criteria in reviewing and assessing surveys and questionnaires for disability outcomes research[6].

The purpose of this paper is to provide some insights to help in selecting and/or revising outcome measures in PRM, using modern psychometric approaches.

PSYCHOMETRIC SCALE ASSESSMENT USING RASCH ANALYSIS

From a practical point of view, RA gives psychometric information that is not obtainable through CTT and includes: (*i*) the functioning of rating scale categories; (*ii*) the validity of a measure, by evaluating how well an item performs in terms of its relevance or usefulness for measuring the underlying construct, and comparing the consistency of item difficulties with the expectations of the construct; (*iii*) the reliability, in terms of 'separation' (i.e. the ratio of the true spread of the measures with their measurement error); and (*iv*) the dimensionality of the scale and the analysis of local independence of items (as performed by principal component analysis of the standardized residuals). In this paper, we briefly review the above points. Once all these steps have been successfully verified, it is eventually possible to convert (via tables or nomograms) raw ordinal scores into Rasch interval scores, fully compliant with measurement theory.

Functioning of rating scale categories

In order to investigate whether a rating scale is being used in the intended manner, usually a procedure of 'rating scale diagnostics' based on RA is applied. The *rating scale model* is used to model responses for multiple-category rating scales that have the same number of choices for all questions. It specifies that a set of items shares the same rating scale structure and provides average measures and thresholds for response categories for the entire instrument. Conversely, the *partial credit model* is used if the pattern of threshold difficulties changes across items.

The performance of the response categories can be evaluated according to a set of common sense criteria that have been formalized statistically in the framework of Rasch models by Linacre[7]: (a) at least 10 observations per category; (b) even distribution of category use; (c) monotonic increase in both average measures across rating scale categories and thresholds. The average measure for a category is the average ability of the people who respond in that category. Thresholds (sometimes also called step calibrations) are the points at which the probability of a response in one or other of 2 adjacent categories is equally likely, i.e. thresholds represent the transition from one category to the next. Additional criteria are: (d) category outfit mean square values less than 2, and (e) threshold differences higher than 1 and lower than 5 logit units.

Where necessary, redundant categories are collapsed to optimize the rating scale. As an example, table 1 shows the main results of recent Rasch studies suggesting the reduction of response categories in some outcome measures[8-11] (Table 1).

As it is often possible to use different collapsing schemes (for instance, category 1 could be collapsed with category 0 or 2), several different categorizations are compared, keeping track of the reliability indices since the more you collapse categories, the more statistical and diagnostic information you lose. The aim is to select the solution that maximizes statistical performance and clinical meaningfulness[5].

The number of categories in a rating scale should be selected with parsimony. When the available categories exceed the number of levels of a construct that participants can discriminate, one begins introducing error variance rather than information into the ratings[3]. Category probability curves for a hypothetical 5-category scale are reported in figure 1. The ideal plot should look like an ordered even succession of hills, with an 'emerging' crest where each category is modal over a certain range. The "0" curve declines as the subject's ability increases; the crossing point (where 0 and 1 are equally probable) is the first "threshold". The same applies for the other curves. In figure 1A the graph shows that the probability of using category 1 is never higher than that of adjacent ratings, and that of using category 3 is quite slim. Conversely, in figure 1B the graph shows that the probability of selecting each of

Table 1 Main results regarding category collapsing in different outcome measures. Italics and bracket show the categories that RA indicates as redundant and may be collapsed.

Questionnaire	Question	Ordinal levels
Locomotor Capability Index[8]	ABILITY Whether or not you wear your prosthesis at the present time, would you say that you are able to do the following activities with your prosthesis on?	0 = no *1 = yes, if someone helps me* *2 = yes, if someone is near me* 3 = yes, alone, with ambulation aids 4 = yes, alone, without ambulation aids
ABILHAND[9]	EASINESS The patient is asked to evaluate the ease of performing 46 common manual activities of daily living	0 = not able to do *1 = very difficult* *2 = slightly difficult* 3 = easy 4 = very easy
Orthotics & Prosthetics User Survey[10]	EASINESS Please indicate how easily you perform the following activities	0 = cannot perform activity *1 = very difficult* *2 = slightly difficult* 3 = easy 4 = very easy
Parkinson's Disease Questionnaire -8[11]	FREQUENCY How often have you experienced difficulties due to Parkinson's disease in the preceding month?	0 = never *1 = occasionally/rarely* *2 = sometimes* *3 = often* 4 = always.

Table 2 Mean difficulty estimates for each of the 14 items of the Mini-BESTest with standard errors (S.E.) and infit and outfit mean-square statistics (MnSq). The higher the item difficulty estimate, the less likely it is for any subject to gain a high score.

ITEM	Mean difficulty	S.E.	Infit MnSq	Outfit MnSq
3 – Stand on Left/Right (L/R) leg	2.43	0.25	0.90	1.07
6 – Step lateral (L/R)	1.10	0.22	0.84	0.76
11 – Head turns	1.00	0.19	0.91	0.83
5 – Step backward	0.93	0.22	0.97	1.08
14 – Timed "Get Up and Go" with dual task	0.77	0.24	1.07	1.08
2 – Rise to toes	0.65	0.20	0.94	1.11
8 – Foam surface, eyes closed	0.54	0.20	1.04	1.12
13 – Step over obstacles	0.10	0.21	0.75	0.73
4 – Step forward	-0.03	0.21	1.14	1.23
9 – Incline, eyes closed	-0.64	0.21	1.12	1.00
12 – Pivot turns	-0.85	0.21	0.99	1.32
10 – Change speed	-1.00	0.20	0.89	0.78
1 – Sit to stand	-1.78	0.24	1.30	1.32
7 – Stance, eyes open	-2.51	0.39	1.12	0.66

the 3 revised rating categories (according to the scheme 01122) is a clear function of the level of ability shown by the subject in the x-axis. Correspondingly, the "thresholds" are ordered: i.e. a greater ability is required when the most likely response is 1 rather than 0, and 2 rather than 1 (Figure 1).

In general, using three to five well-selected categories improves the measurement qualities of the scale (without decreasing its reliability indexes), minimizing irrelevant construct variance and ensuring that each rating category represents a clearly distinct level of 'ability', level of agreement or similar.

Moreover, the art of asking questions is a crucial point for an outcome measure. Both Wolfe & Smith[12] and McColl et al.[13] suggest detailed guidelines for writing rating scale items, in order to maximize the measurement validity.

Item validity

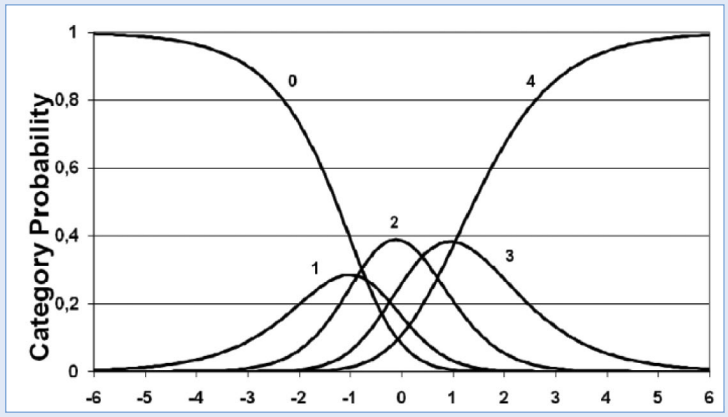
Depending on the string of ordinal raw scores, RA also assesses the extent to which the observed responses to the items accord with the responses predicted by the mathematical model. This is obtained by estimating goodness-of-fit (or simply fit) of the real data to the modelled data using particular expressions of the chi-square statistic (outfit = outlier-sensitive fit statistic, and infit = inlier-pattern-sensitive fit statistic) divided by its degrees of freedom [mean-square (MnSq)]. In accordance with the literature, with a sample size of about 100 persons MnSq values in the range of 0.7 to 1.3 indicate an acceptable fit (e.g., a value of 1.3 indicates 30% more variation in the observed data than the Rasch model predicted). If the differences between observed and expected scores are in the acceptable range, the data are said to "fit the model", and this is seen as equivalent to proving the theoretical construct validity and adequacy of the scale. Items outside this range are considered underfitting (MnSq > 1.3 suggesting the presence of unexpectedly high variability) or overfitting (MnSq < 0.7, indicating a too predictable pattern).

As already stated, RA provides estimates of the level of difficulty achieved by each item ('item difficulty') and of the location of each individual subject along the continuum ('subject ability' representing the global amount of trait in the individual). Item difficulty and subject ability are expressed – on a common interval scale – in logit units, a logit being the natural logarithm of the ratio (odds) of mutually exclusive alternatives (e.g. pass *vs.* fail or higher response *vs.* lower response). Logit-transformed measures represent linear measures (i.e. the intended amount of the trait). Conventionally, 0 logit is ascribed to the mean item difficulty. For RA it is reported that a sample size of about 100 people will estimate item difficulty with an alpha of 0.05 to within ± 0.5 logits.

As an example, table 2 shows the main results of a fit statistic of the Mini-BESTest, a Rasch-based balance measure.

During all the above procedures, the validity of the test items for their intended application and population is the most important aspect to consider. Thus, one needs to be careful about deleting items from an outcome measure based on statistical results only. Data analysis is an aid to thought, not a substitute[6]. The items to consider for deletion are those that[5,15]: 1) do not fit the Rasch model; 2) show redundancy, i.e. share the same span of item difficulty, thus introducing a risk of inflation of the cumulative raw score when the

a)



b)

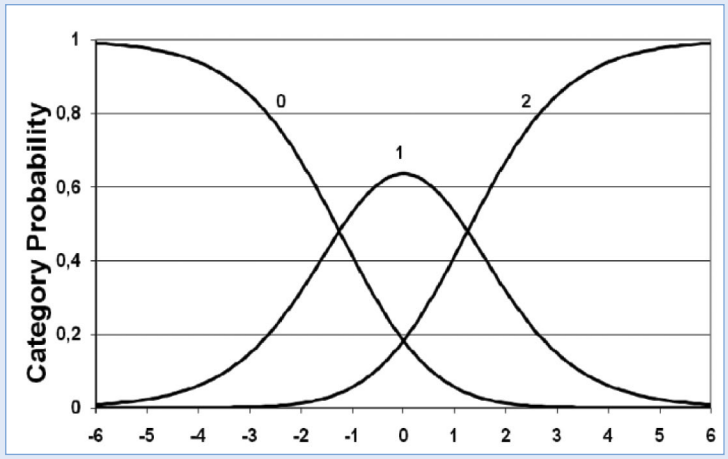
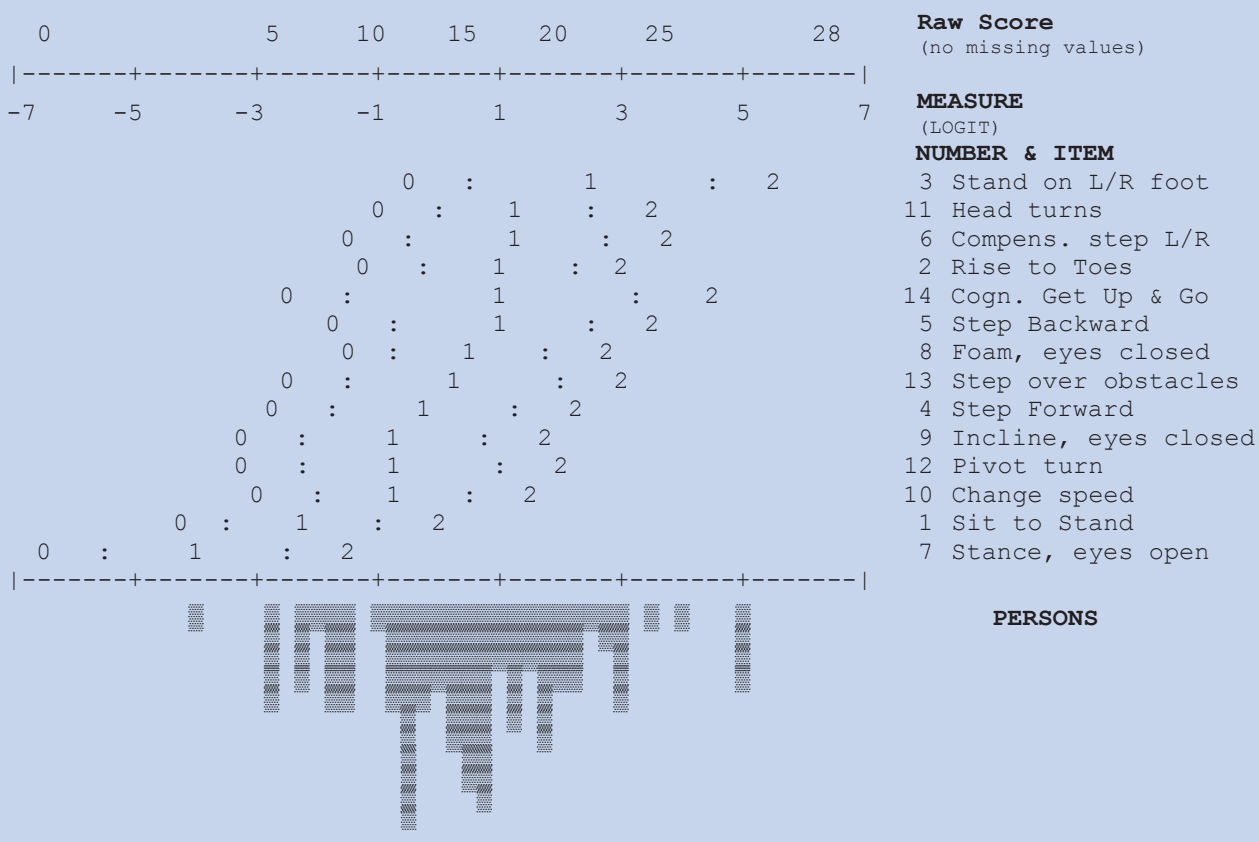


Figure 1 Category probability curves.

a) original scale with 5 categories (0-4)

b) revised scale after collapsing category 1 with 2, and 3 with 4 and renumbering (01122). The *y*-axis represents the probability (0 to 1) of responding to one of the rating categories and the *x*-axis represents the different performance values (patient ability minus the item difficulty) in logits (theta).

Table 3 Expected scores for a Rasch-based balance scale: Mini-BESTest. Distance between points is equal-interval. Logit measure at top of key, centered at the mean item difficulty. The rating scale is based on 3 categories (0= severely impaired; 1= moderately impaired; 2= normal). The threshold between adjacent categories is marked by ':'. At the bottom is the distribution of the person measures (subject ability) in the study sample: each marker ▒ is a single person.



scores of individual items reflecting the same level of ability are summed; 3) present local dependence (i.e. a large positive correlation at principal component analysis of the standardized residuals after Rasch modelling). For example, two items with a correlation > 0.7 share more than half their "random" variance, suggesting that just one of the two items is sufficient for measurement; 4) show differential item functioning, i.e. the probability of responding in different categories varies across subgroups (given an equivalent level of the underlying attribute). This means instability of item hierarchy across different samples and reduces the validity of between-group comparison, since the scores indicate additional attributes to the one the scale is intended to measure; 5) are judged by expert review as not very relevant for measuring the construct in question.

At the end of these analyses, in most cases 10 to 15 well-chosen items [i.e. with 'expert-certificated' validity (after evaluation of both the construct being measured and the conceptual model underlying the measurement of that construct), fitting the model, making an independent contribution to the construct and uniformly spaced in terms of difficulty over the measurement range] turn out to be suitable for a correct measurement.

A key form for clinical use for a given patient is shown in table 3. Clinicians can circle responses to the 14 items and then mark a vertical line that passes through the mid-point of the ratings; the point where this line intersects the horizontal axis (measure in logits) is the estimated measure for that person. If this line intersects the horizontal line at zero, the patient has a moderate level of balance function – these activities are not very difficult or very easy. Negative values reflect a lower level of functional ability, while positive values reflect a higher level of functional ability.

Reliability

In RA, reliability is evaluated in terms of separation (G), defined as the ratio of the true standard deviation of the measures to their standard deviation measurement error. Along the measurement construct the item separation index gives an estimate, in standard error units, of the spread or separation of items along the measurement construct, whereas the person separation index gives an estimate of the spread or separation of respondents. This index reflects the number of strata of measures that are statistically discernible. A separation of 2.0 is considered good and enables the distinction of three groups or strata, defined as segments whose centers are separated by distances greater than can be accounted for by measurement error alone [number of distinct strata = (4G + 1)/3]. A related index is the reliability of these separation indices, providing the degree of confidence that can be placed in the consistency of the estimates. Coefficients range from 0 to 1: coefficients of > 0.80 are considered good, and coefficients of > 0.90 are considered excellent.

Scale unidimensionality and local independence of items

In applying RA, it is important to evaluate the core assumptions of the model, first of all unidimensionality, because one critical point of these statistical models is that the person's response to an item that measures a construct is accounted for by his/her amount of that trait, and not by other factors.

Usually, dimensionality is preliminarily analyzed by factor analysis (for categorical data), but in RA a principal component analysis on the standardized residuals can be performed as a test of the unidimensionality of the scale (proportion of variance attributable to the first residual factor compared with that attributable to Rasch measures) and of the local independence of each item (i.e. the independence of item measures from extraneous variables, once their belonging to the shared construct has been ascertained).

After the removal of the trait/construct that the scale intends to measure (the so-called Rasch factor), the residuals for items should be uncorrelated and normally distributed (i.e. there are no principal components). The following are the main criteria used to determine whether additional factors are likely to be present in the residuals: 1) a cut-off of 50% of the variance explained by the measures; 2) an eigenvalue of the first residual factor smaller than 3 and b) a percentage variance explained by the first contrast of 5%.

Rating scale structure should be as simple as possible: in most cases three to five well-selected categories are enough. The wording of questions has a major impact on validity and reliability of an instrument.

Many parameters should be considered to select the set of items with best coverage and technical quality. Data analysis is an aid to thought, not a substitute for clinical reasoning.

As for the local independence between items, a high correlation (> 0.30) of residuals for 2 items indicates that they may not be locally independent, either because they duplicate some feature of each other or because they both incorporate some other shared dimension.

CONCLUSIONS

In this paper we have discussed just a few practical issues related to assessment of outcome measures by RA, underlining the complexity of this field. At present, RA represents one of the best methods for studying several key methodological aspects associated with scale development and construct validation that cannot be analysed by traditional techniques[5,6].

Outcome measures are an important aspect of clinical practice, audit and research. Considerable care needs to be taken to ensure that the best possible measure for the task in hand is selected, and that, wherever possible, the selected measure conforms to modern quality standards for measurement. We think that the awareness of this kind of validation can by itself help the final users to critically inspect each outcome measure and the related literature before deciding to use one in clinical practice, decision making or policy development.

Unfortunately, little attention in general is paid to the theoretical framework of health outcome measures and to the large variation in the methodological development and validation of commonly used tools[16]. Future research in PRM should address both methodological and applied issues, e.g. more use of modern psychometric methods for measurement validation, better calibration and responsiveness of the instruments, studies on comparability across different populations, more projects on item banks and computerized adaptive testing[17,18].

LITERATURE

1. Franchignoni F, Michail X. Selecting an outcome measure in Rehabilitation Medicine. Eura Medicophys 2003; 39:67-8.
2. Franchignoni F, Ring H. Measuring change in rehabilitation medicine. Eura Medicophys 2006;42:1-3.
3. Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med 2003;35:105-15.
4. Frost MH, Reeve BB, Liepa AM, Stauffer JW, Hays RD; Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value Health 2007;10 Suppl 2:S94-105.
5. Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. 2nd ed. Mahwah: Lawrence Erlbaum Associates; 2007.
6. McHorney CA, Monahan PO. Applications of Rasch analysis in health care. Med Care 2004;42(1 Suppl): I73-8.
7. Linacre JM. Optimizing rating scale category effectiveness. J Appl Meas 2002;3:85-106.
8. Franchignoni F, Giordano A, Ferriero G, Muñoz S, Orlandini D, Amoresano A. Rasch analysis of the Locomotor Capabilities Index-5 in people with lower-limb amputation. Prosthet Orthot Int 2007;31:394-404.
9. Burger H, Franchignoni F, Kotnik S, Giordano A. A Rasch-based validation of a short version of ABILHAND as a measure of manual ability in adults with unilateral upper limb amputation. Disabil Rehabil 2009;31:2023-30.
10. Burger H, Franchignoni F, Heinemann AW, Kotnik S, Giordano A. Validation of the orthotics and prosthetics user survey upper extremity functional status module in people with unilateral upper limb amputation. J Rehabil Med 2008;40:393-9.
11. Franchignoni F, Giordano A, Ferriero G. Rasch analysis of the short form 8-item Parkinson's Disease Questionnaire (PDQ-8). Qual Life Res 2008;17:541-8.
12. Wolfe EW, Smith EV Jr. Instrument development tools and activities for measure validation using Rasch models: part I – instrument development tools. J Appl Meas 2007;8:97-123.
13. McColl E, Jacoby A, Thomas L, Soutter J, Bamford C, Steen N et al. Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. Health Technol Assess 2001;5: 1-256.
14. Franchignoni F, Horak F, Godi M, Nardone A, Giordano A. Using the psychometric techniques to improve the Balance Evaluation Systems Test: the mini-BESTest. J Rehabil Med 2010;42:323-31.
15. Wolfe EW, Smith EV Jr. Instrument development tools and activities for measure validation using Rasch models: part II – validation activities. J Appl Meas 2007;8:204-34.
16. Franchignoni F, Giordano A, Ferriero G. Considerations about the use and misuse of Rasch analysis in rehabilitation outcome studies. Eur J Phys Rehabil Med 2009;45:289-92.
17. Franchignoni F, Giordano A, Michail X, Christodoulou N. Practical lessons learned from use of Rasch analysis in the assessment of outcome measures. Port J PRM 2010;19:5-12.
18. Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: Time to end malpractice? J Rehabil Med 2012;44:97-8.