*Pregledni rad*
*Review*

*ISSN 1846-1867*

# New psychometric strategies for an appropriate selection and use of outcome measures in physical and rehabilitation medicine

**Franco FRANCHIGNONI[1], Giorgio FERRIERO[2], Marcella OTTONELLO[3,4]**

*[1]School of Specialization in Physical and Rehabilitation Medicine,*

*Tor Vergata University, Rome, Italy*

*[2]Department of Physical Medicine & Rehabilitation,*

*Hamad Medical Corporation, Doha, Qatar*

*[3]Department of Physical & Rehabilitation Medicine, Salvatore Maugeri Foundation,*

*Clinica del Lavoro e della RiabilitazioneIRCCS, Nervi (GE), Italy*

*[4]PhD program in Advanced Sciences and Technology in Rehabilitation Medicine and*

*Sports, Tor Vergata University, Rome, Italy*

*Address for correspondence:*
*Prof.dr.sc. Franco Franchignoni,*
*School of Specialization*
*in Physical and Rehabilitation Medicine,*
*Tor Vergata University,*
*Rome, Italy*

## Summary

In order to be useful for their intended purposes, outcome measures (rating scales and questionnaires) must provide information that allows valid inferences and decisions to be made.

Basic classical test theory is still widely used in peer-reviewed, indexed journals for validating these tools. Classical test theory methods mainly focus on an instrument's total score, whichis simply asserted as the relevant statistic. But, this approach neglects a series of criteria that need to be considered when

evaluating the psychometric properties of a measurement tool, and that can be analysed by Rasch analysis. The validation activities performed by RA are numerous; the most important ones are those connected with the analysis of:
**a.** dimensionality;
**b.** functioning of rating scale categories;
**c.** internal construct validity of the measure;
**d.** reliability of the scale, in terms of 'separation' (i.e. the ratio of the true spread of the measures with their measurement error).

Thus, RA is being increasingly used in the development and evaluation of clinical tools for health care.

The purpose of the present paper is to describe the main features of Rasch analysis in assessing outcome measures, and summarize some results of our recent psychometric studieson outcome measures in Physical and Rehabilitation Medicine, in order to provide insights for the appropriate selection and use of outcome measures.Physiatrists have a responsibility to ensure that measures used in clinical settings are psychometrically sound, and that they are administered thoughtfully and analysed correctly. The contents of this article can bring the final users to critically inspect each outcome measure and the related literature before adopting it for clinical practice, decision making, or policy development.

**Key words:** outcome measure, Rasch analysis, psychometrics, Physical and Rehabilitation Medicine

## Introduction

In recent years there has been an increasing use of outcome measures in clinical practice, audit procedures and quality control. Accordingly, physicians need to acquire specific expertise to be able to select the appropriate outcome measure, administer itthoughtfully, and interpret the results correctly (1).

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). 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 outcome measures (rating scales and questionnaires) measuring 'latent traits' must provide information that allows valid inferences and decisions to be made. In Physical and Rehabilitation Medicine, the International Classification of Functioning, Disability and Health (ICF) provides a unified and standard language, and a conceptual framework for the description of health-related states and the general classification of the outcome measures.

Basic classical test theory (CTT) is still widely used in peer-reviewed, indexed journals for validating these tools, in both original and translated versions. These studies are mainly focused on an instrument's total score, and largely based on analysis of internal consistency [using Cronbach's alpha, well known for its limits (4)], reproducibility, and criterion-related validity (usually the demonstration of a moderate to good correlation with some other measure of the trait under study). But, this approach is not sufficient, because it neglects a series of criteria that need to be considered when evaluating the psychometric properties of a measurement tool (5-11), and that can be analysed by Rasch methods, such as the evaluation of how well an item performs in terms of its relevance or usefulness for measuring the underlying construct, the amount of the construct targeted by each question, the possible redundancy of anitem relative to other items in the scale, and the appropriateness of the response categories (12).

Thus, Rasch analysis (RA) represents the top of the pyramid in psychometric analysis, and is being increasingly used in the development and evaluation of clinical tools for health care (13) (Figure 1).

The purpose of the present paper is to describe the main features of RA in assessing outcome measures, and summarize some results of our recent psychometric studies, in order to provide insights for the appropriate selection and use of outcome measures. The following text is structured in three sections:
**1.** What RA is;
**2.** Some recent Rasch studies on outcome measures in Physical and Rehabilitation Medicine;
**3.** How to select an outcome measure in clinical practice and research.

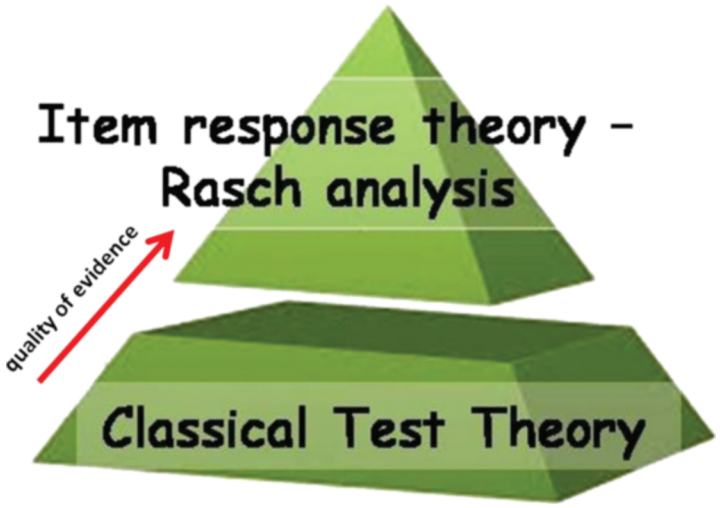


**Figure 1.** The evidence-pyramid graphic related to psychometric analysis of outcome measures. As you move up to studies applying item-response-theory and Rasch analysis, the results provide in-depth psychometric information on basic measurement properties that is not obtainable through Classical Test Theory.

## What is rasch analysis

Traditional psychometric approaches mainly focus on an instrument's total score, whichis simply asserted as the relevant statistic. Conversely, in RA (Rasch analysis, an original item-response theory analysis based on latent-trait modelling, and named after the Danish mathematician Georg Rasch) the total score summarizes completely a person's standing on a variable just if the scale fits the Rasch model, that prescribes how data should be in order to obtain correct measurements from the data. The model postulates that the probability of a particular response to an item is modelled as a function of 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". 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) (3).

This is a criterion for successful measurement: RA estimates, amongst other things, how much the modelled measure is supported by the actual observed scores (the so-called 'data-model fit'). The validation activities performed by RA are numerous; the most important ones are those connected with the analysis of:

**a.** dimensionality;

**b.** functioning of rating scale categories;

**c.** internal construct validity of the measure (by evaluating how well an item performs in terms of its relevanceor usefulness for measuring the underlying construct, and comparing the consistency of item difficulties with the expectations of the construct;

**d.** reliability of the scale, in terms of 'separation' (i.e. the ratio of the true spread of the measures with their measurement error) (14).

## Dimensionality

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), but also 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:

**I.** a cut-off of 50% of the variance explained by the measures;

**II.** an eigenvalue of the first residual factor smaller than 3;

**III.** a percentage variance explained by the first contrast of 5%.

## Functioning of rating scale categories

In order to investigate whether a rating scale is being used in the intended manner (in terms of type and number of the rating scale categories), usually a procedure of 'rating scale diagnostics' based on RA is applied. The performance of the response categories is usually evaluated according to a set of common sense criteria (adequate number of responses per category, even use of the categories, monotonic increase of the difficulty of each category, fair coverage of the possible responses, etc.) that have been formalized statistically in the framework of Rasch models (12), such as: at least 10 observations per category; even distribution of category use;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 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.

Where necessary, categories are collapsed (using different collapsing schemes) to optimize the rating scale. The aim is to select the solution that maximizes statistical performance and clinical meaningfulness (14): the rating scale should be conveyed with categories and labels that elicit unambiguous responses. A typical graphic presentation of the results of 'rating scale diagnostics' – taken from a recent paper of our group (15)–is shown in Figure 2.

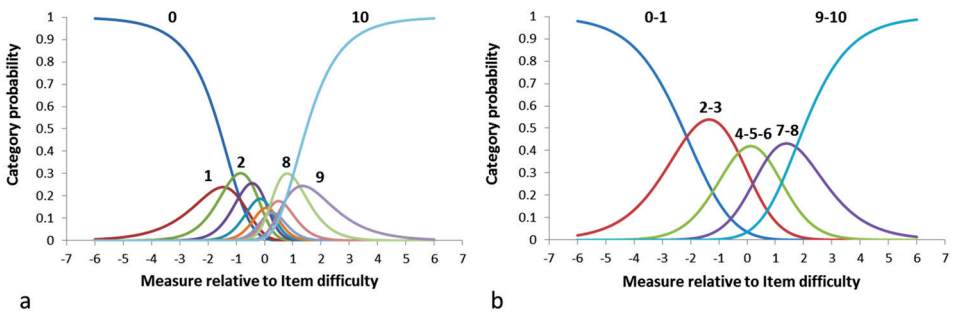


**Figure 2** –Category probability curves: a) Curves of the original 11 categories (0–10) of Bath Ankylosing Spondylitis Functional Index; and b) of the 5 revised categories obtained by combining the original categories as follows: 0-1=0; 2-3=1; 4-6=2; 7-8=3; 9-10=4 (15). The y axis represents the probability (0–1) of responding to one of the rating categories and the x axis the different performance values (patient ability minus item difficulty) in logits. The ideal plot should look like an ordered even succession of hills [as in (b)], with an 'emerging' crest where each category is modal over a certain range. Conversely, in (a) the probability of using some categories is never higher than that of other adjacent ratings.

The intersection of probability curves of rating scale categories shows the point at which there is an equal probability of choosing either of two adjacent response category options (threshold estimates), i.e. where – on the trait continuum – there is a transition from answering with one response option to the next.

**Internal construct validity**

The validity of the test items for their intended application and population is the most important aspect to consider. Depending on the string of ordinal raw scores, RA 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 misfitting: 'underfitting' where MnSq > 1.3 (suggesting the presence of unexpectedly high variability), and 'overfitting' where MnSq < 0.7 (indicating a too predictable pattern) (3).

One needs to be careful about deleting misfitting items from an outcome measure based on statistical results only: data analysis is an aid to thought, not a substitute (14). The items to consider for deletion are those that: i) do not fit the Rasch model; ii) show redundancy, i.e. share the same span of difficulty (as items 2, 3 and 11 and items 1, 8, and 9 in figure 2), thus introducing a risk of inflation of the cumulative raw score when the scores of individual items reflecting the same level of ability are summed (3); iii) present local dependence (i.e. a large positive correlation at principal component analysis of the standardized residuals after Rasch modelling) (7). 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; iv) 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; and, last but not the least; v) are judged by expert review as not very relevant for measuring the construct in question.

**Reliability, sensitivity to change and responsiveness**

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 indexgives an estimate, in standard error units, of the spread or separationof items along the measurement construct, whereas the person separation indexgives an estimate of the spread or separation of respondents. This index reflects the number of strataof 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. The person separation reliability is based on the same concept as Cronbach alpha, and the mistargeting of the range of item difficulties to the range of respondents' abilities (i.e. the extent to which the items are not appropriately difficult for that sample) can negatively affect its value.

As for measuring the change in outcome measures, there are two main types of approach: distribution-based methods and anchor-based methods. The distribution-based methods are based on the statistical characteristics of the obtained sample and analyze the ability to detect change in a state, regardless of whether this change is relevant or meaningful to the decision-maker (parameters such as the Standard Error of Measurement and the Minimum Detectable Change are calculated). Conversely, the anchor-based methods require an external criterion to determine if changes in outcome scores are clinically meaningful: the external assessmentcan be obtained through a Global Rating of Change (i.e. an ordinal rating scale designed to quantify patients' improvement or deterioration over time); then, the mean change approach

and the Receiver Operating Characteristic (ROC) curve approach are usually applied, in order to obtain the Minimal Clinically Important Difference.

## Some recent rasch studiesrelated to outcome measures for physical and rehabilitation medicine

Recently, a review of the application and quality of reporting of RA in musculoskeletal disorders over the past two decades showed that the Rasch measurement model has been increasingly used in these fields (16). Our group has published several papers examining with RA the psychometric properties of a series of outcome measures. We briefly summarizethe main results of these papers, in order to show how much this kind of analyses can help in the assessment and refinement of the measurement tools. The ideal scale is the one where the items have 'expert-certificated' validity (after evaluation of both the construct being measured and the conceptual model underlying the measurement of that construct), fit the model, make an independent contribution to the construct, and be uniformly spaced in terms of difficulty over the measurement range.

LEQUESNE ALGOFUNCTIONAL INDEX FOR THE SEVERITY OF OSTEOARTHRITIS OF THE HIP (LAI-hip), and KNEE (LAI-knee) – A recent study performed a comprehensive psychometric analysis of these outcome measures in the two samples respectively representing a wide spectrum of hip and knee osteoarthritis severity, in order to better understand the strengths and weaknesses of both instruments (17). Using both CTT and RA, the main problem was the presence of a mix of items assessing 'function' and 'mobility' with others evaluating 'pain and discomfort'. Moreover, the rating categories of the item 'Maximum distance walked' did not comply with the criteria for category functioning, in both scales: this because respondents had difficulty in consistently discriminating between the seven response options (being too many, or with a confusing labeling).

In conclusion, the LAI-hip and LAI-knee showed a series of drawbacks, which render both questionnaires inadequate in relation to their metric properties and severely limit their ability to perform, as a composite measure, in line with the main aims of their developers.

KNEE INJURY AND OSTEOARTHRITIS OUTCOME SCORE - PHYSICAL FUNCTION SHORT FORM (KOOS-PS7) – In 2008, Perruccio et al.(18) used RA to analyze the responses of the 22 KOOS items comprising the 2 subscales of ADL function

and sport/recreation function (KOOS-PF22). Their objective was to develop a short measure of physical function across the ostheoarthritis spectrum (from early to late disease) for individuals with ostheoarthritis of the knee. The result was a 7-item scale, dubbed KOOS-Physical Function Short Form (KOOS-PS7), that received further validation in studies performed by both multinational and local groups (19-21), using CTT only. However, the item selection process was mainly data-driven and sometimes based on uncertain item bias, leading to a very short version with the risk of reliability values borderline for clinical application in individuals (22). Therefore, an independent replication of the procedure seemed called for.

The replication study was performed in a sample of Italian patients with ostheoarthritis (23). Its main results were as follows: i) the KOOS-PF22 showed an underlying response structure sufficiently unidimensional to allow further analysis using Rasch methods; ii) the replication of RA on the KOOS-PF22 was not able to confirm the selection of the items included in KOOS-PS7, but selected a unidimensional pool of 12 items (KOOS-PS12) with promising psychometric characteristics; iii) the reliability levels of the KOOS-PS7 indicated that this instrument is more useful for group decisions than for everyday clinical application in single patients.

Overall, if the aim is to achieve a short measure of physical function for a wide spectrum of individuals with ostheoarthritis of the knee (able to optimize coverage and technical quality with an appropriate number of items) starting from the KOOS (20), further large independent studies -based on modern psychometric approaches and re-analysing in depth the full scale with the support of expert opinion- are recommended. The KOOS-PS12 represents a useful starting point for creating a scale with good psychometric properties and seems to contain a number of items able to optimize coverage of the construct (internal validity) and technical quality.

DISABILITIES OF THE ARM, SHOULDER AND HAND (DASH) – A comprehensive psychometric analysis of DASH was performed in 2010, to examine its properties and provide insights for an improved version of the questionnaire (24). This study did neither confirm unidimensionality of the scale, nor the key domains identified by the original developers as the theoretic framework of DASH. The dimensional complexity of DASH is justified by some considerations: the DASH Outcome Measure User's Manual states that the questionnaire mixes up symptoms and disability, there are items that do not exclusively rely on

the function of the upper limb, and the tool contains some items that measure different constructs, such as impairment, activity limitations, and participation restriction (24). Another critical point of DASH is the number (and/or wording) of its response categories: rating scale diagnostics provided evidence that respondents were unable to discern appreciably the five response levels. In addition, two DASH items (#21 'Sexual Activities', and #26 'Tingling') showed a clear misfit at RA.

Overall, further detailed investigations of DASH are warranted, both to confirm these results in different health conditions and cultures, and to reanalyse in depth content validity issues regarding the questionnaire.

DISABILITIES OF THE ARM, SHOULDER AND HAND QUESTIONNAIRE, SHORT VERSION (QUICKDASH) – The metric properties of QuickDASH have recently been examined in detail (25) and – as a didactic example – the thresholds-persons map for the Quick DASH is reported in Figure 3.

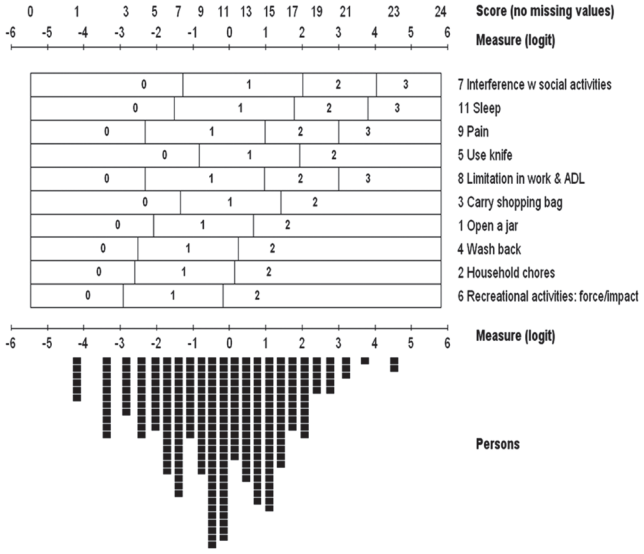


**Figure 3**. Thresholds-persons map for the QuickDASH (without item 10 'tingling', that was misfitting), and with the items in location order. The first two lines contains the Rasch nomogram, which allows the conversion of total raw score (no missing data) into a logit measure (centered at the mean item difficulty). The latent trait (disability of upper limb) increases toward the right with the severity of dysfunction. In the middle, there is the threshold map for QuickDASH items. The rating scales have been collapsed from 5 to 3-4 categories. At the bottom, the distribution of the subjects (Persons) in the study sample according to their ability: each marker is a single patient.

In summary, the main results of the study were as follows: i) the unidimensionality of QuickDASH has not been confirmed. Both factor analysis and RA showed that at least one item (#10 'Tingling') does not belong to the dominant trait. This is not surprising because tingling is a symptom specific for a limited range of upper limb pathologies (e.g. nerve entrapments); ii) the number (and/or wording) of the QuickDASH response categories should undergo further investigation; iii) the reliability indexes of the questionnaire are good but not excellent: it seems more useful for group decisions than for everyday clinical application in monitoring outcome in single patients. On the basis of these results future studies were recommended, in order to revise the QuickDASH, restarting from the original full-length DASH.

In the meantime, another study of our group determined the Minimal Clinically Important Difference for DASH and QuickDASH in patients with upper-limb musculoskeletal disorders, using a triangulation of distribution-based (Minimum DetectableChange) and anchor-based (mean change method; Receiver Operating Characteristic curve analysis) approaches (Figure 4).

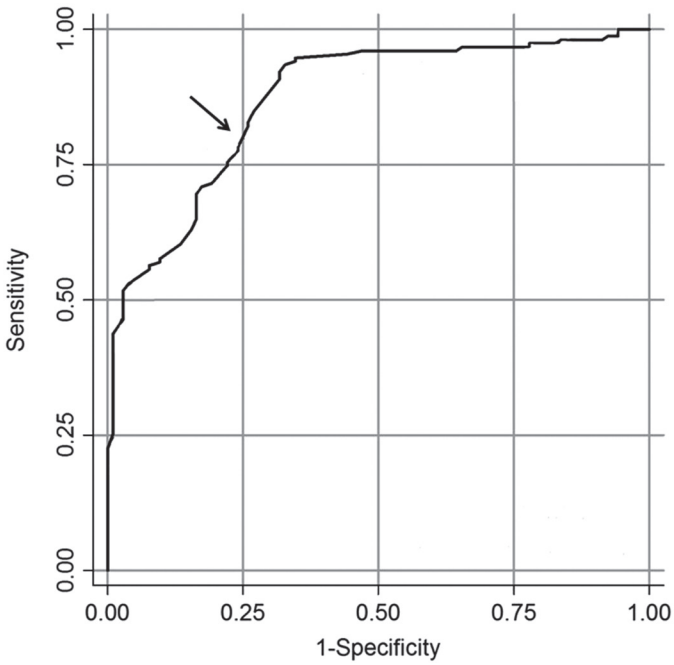


**Figure 4** – Receiver Operating Characteristic curves of DASH, showing its overall accuracy in identifying an improvement according to theGlobal Rating of Change Scale < +2 vs. ≥ +2. Arrows show the Minimal Clinically Important Differences point (see text).

The proposed ranges of Minimal Clinically Important Differencesvalues -reasonably useful for different populations and contextual characteristics- were as follows: DASH from 10.83 to 15 points; QuickDASH from 15.91 to 20 points (26).

IOWA LEVEL OF ASSISTANCE SCALE(ILAS) – A psychometric analysis, using both CTT and RA methods of the Iowa Level of Assistance Scale (ILAS) administered in patients with recent total hip arthroplasty or total knee arthroplasty, has been recently performed, to examine its metric properties and provide insights for a refined version. The scale includes two domains: the ILAS for assistance needed during functional activities (ILAS-funct) and need for assistive devices (ILAS-dev). The two domains showed a good correlation. According to rating scale diagnostics, ILAS-funct showed two disordered response category thresholds (of the seven different response levels of 'assistance', only five were appreciably discernible). All five ILAS-funct items fitted the model and did not show either local dependence or differential item functioning across age groups or sex. Conversely, ILAS-dev presented two unused response categories, which precluded Rasch calibration and subsequent analyses. Overall, ILAS-funct showed sound psychometric properties, but the rating system of ILAS-funct could be simplified. In ILAS-dev, there is need for a reconsideration of its scaling options and methods (27).

## How to select an outcome measure in clinical practice and research

The use of an outcome measure is an important aspect of the Rehabilitation practice, audit and research. The user has to choose an outcome measure on the basis of its internal structure, the psychometric properties required for the intended purpose and the previous use of that measure in similar clinical situations and contexts. If the appropriate outcome measure has been selected, the information gained can be used to measure the variable of interest developing a comprehensive list of clinical problems that is useful for establishing short- and long-term goals optimizing the patient's plan of treatment (1, 2).

Considerable care needs to be taken to ensure that the selected outcome measure is the most appropriate, because instruments may have varying strengths and weaknesses, depending on the population and the reasons for their use, so the user's final decision must be context-specific. Thus, clinicians had to analyse the match of each instrument to the specific purposes,

circumstances and questions of a trial, carefully considering: i) aims and end points of the trial; ii) nature of the study intervention; iii) features of the patient group; iv) internal construct validity of each measure, in comparison with alternative candidate instruments; v) previous use in the literature of the measure, in similar contexts.

Moreover, there is a series of additional practical and technical attributes that investigators should take into account in selecting the most appropriate measure, such as: i) interpretability (measures should give results which are easily understood by others); ii) acceptability (how acceptable it is for respondents to complete: response rate, time to complete, cultural applicability, and so on); iii) feasibility (ease of administration and processing, i.e. extent of effort, burden and disruption to staff and clinical care arising from the use, including for example the professional expertise required to apply or interpret the instrument, and the presence of a clear instruction manual); andiv) presence of a correct cross-cultural adaptation of the instrument to the new country, culture and/or language (ensuring the attainment of equivalence between the source and target measures) (28).

Last but not the least, the selected measure should conform to modern quality standards for measurement, in terms of reliability, validity and responsiveness to change.

In the present paper, we have discussed how RA can help in examining essential psychometric properties of the outcome measures, that cannot be analysed by traditional techniques (12,14, 29).

In conclusion, the above considerations and suggestions can bring the final users to critically inspect the content of each outcome measure and the related literature, before adopting it for clinical practice, decision making or policy development. Physiatrists have a responsibility to ensure that measures used in our clinical settings are psychometrically sound, and that they are administered thoughtfully and analysed correctly.

Future research in Rehabilitation Medicine should address a better use of modern psychometric methods for measurement validation, better calibration and responsiveness of the instruments, studies on comparability across different populations.

## Izjava o sukobu interesa

Autori izjavljuju da nemaju sukob interesa.

## References:


1. Franchignoni F, Michail X. Selecting an outcome measure in Rehabilitation Medicine. Eura Medicophys 2003;39:67-68.
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;35:105-115.
4. Hattie J. Assessing unidimensionality of tests and items. Appl Psychol Meas 1985;9:139-164.
5. 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-S105.
6. Snyder CF, Watson ME, Jackson JD, Cella D, Halyard MY; Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. Patient-reported outcome instrument selection: designing a measurement strategy. Value Health 2007;10 Suppl 2:S76-S85.
7. Turner RR, Quittner AL, Parasuraman BM, Kallich JD, Cleeland CS; Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. Patient-reported outcomes: instrument development and selection issues. Value Health 2007;10 Suppl 2:S86-S93.
8. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45(5 Suppl 1):S22-S31.
9. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60(1):34-42.
10. Rothman M, Burke L, Erickson P, Leidy NK, Patrick DL, Petrie CD. Use of existing Patient-Reported Outcome (PRO) instruments and their modification: The ISPOR good research practices for evaluating and documenting content validity for the use of existing instruments and their modification - PRO Task Force Report. Value Health 2009; 12:1075-1083.
11. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539-549.
12. Wolfe EW, Smith EV Jr. Instrument development tools and activities for measure validation using Rasch models: part II – validation activities. J Appl Meas 2007;8:204-234.

13. Conrad KJ, Smith EV Jr. International conference on objective measurement: applications of Rasch analysis in health care. Med Care 2004;42(1 Suppl):I1-I6.

14. Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. 2nd ed. Mahwah: Lawrence Erlbaum Associates; 2007.

15. Franchignoni F, Salaffi F, Ciapetti A, Giordano A. Searching for optimal rating scales in the Bath Ankylosing Spondylitis Functional Index (BASFI) and Bath Ankylosing Spondylitis Disease Activity Index (BASDAI). Rheumatol Int 2014;34:171-173.

16. Leung YY, Png ME, Conaghan P, Tennant A. A Systematic Literature Review on the Application of Rasch Analysis in Musculoskeletal Disease -- A Special Interest Group Report of OMERACT 11. J Rheumatol 2014;41:159-164.

17. Franchignoni F, Salaffi F, Giordano A, Ciapetti A, Carotti M, Ottonello M. Psychometric properties of self-administered Lequesne Algofunctional Indexes in patients with hip and knee osteoarthritis: an evaluation using classical test theory and Rasch analysis. Clin Rheumatol 2012;31:113–121.

18. Perruccio AV, Lohmander LS, Canizares M, et al. The development of a short measure of physical function for knee OA KOOS-Physical Function Shortform (KOOS-PS) - an OARSI/OMERACT initiative. Osteoarthritis Cartilage 2008;16:542-550.

19. Ornetti P, Parratte S, Gossec L, et al. Cross-cultural adaptation and validation of the French version of the Knee Injury and Osteoarthritis Outcome Score (KOOS) in knee osteoarthritis patients. Osteoarthritis Cartilage 2008;16:423-428.

20. Davis AM, Perruccio AV, Canizares M, et al. Comparative, validity and responsiveness of the HOOS-PS and KOOS-PS to the WOMAC physical function subscale in total joint replacement for osteoarthritis. Osteoarthritis Cartilage 2009;17:843-847.

21. Goncalves RS, Cabri J, Pinheiro JP, Ferreira PL, Gil J. Reliability, validity and responsiveness of the Portuguese version of the Knee Injury and Osteoarthritis Outcome Score Physical Function Short-Form (KOOS-PS). Osteoarthritis Cartilage 2010;18:372-376.

22. Streiner DL, Norman GR. Health measurement scales: Apractical guide to their development and use. 2nd ed. Oxford: Oxford University; 1995.

23. Franchignoni F, Salaffi F, Giordano A, Carotti M, Ciapetti A, Ottonello M. Rasch analysis of the 22 Knee injury and Osteoarthritis Outcome Score-physical function items in Italian patients with knee osteoarthritis. Arch Phys Med Rehabil 2013;94:480-487.

24. Franchignoni F, Giordano A, Sartorio F, Vercelli S, Pascariello B, Ferriero G. Suggestions for refinement of the Disabilities of the Arm, Shoulder and Hand Outcome Measure (DASH): a factor analysis and Rasch validation study. Arch Phys Med Rehabil 2010;91:1370-1377.

25. Franchignoni F, Ferriero G, Giordano A, Sartorio F, Vercelli S, Brigatti E. Psychometric properties of QuickDASH - a classical test theory and Rasch analysis study. Man Ther 2011;16:177-182.

26. Franchignoni F, Vercelli S, Giordano A, Sartorio F, Bravini E, Ferriero G. Minimal Clinically Important Difference of the Disabilities of the Arm, Shoulder and Hand Outcome Measure (DASH) and Its Shortened Version (QuickDASH). J Orthop Sports Phys Ther 2014;44:30-39.

27. Benedetti MG, Franchignoni F, Morri M, Franchini N, Natali E, Giordano A. Rasch analysis of the Iowa Level of Assistance Scale in patients with total hip and knee arthroplasty. Int J Rehabil Res 2014;37:118-24.

28. Küçükdeveci AA, Tennant A, Grimby G, Franchignoni F. Strategies for assessment and outcome measurement in physical and rehabilitation medicine: an educational review. J Rehabil Med 2011;43:661-72.

29. Franchignoni F, Giordano A, Michail X, Christodoulou N. Practical lessons learned from use of Rasch analysis in the assessment of outcome measures. Portuguese Journal of Physical and Rehabilitation Medicine 2010;19:5-12.