Editorial

Practical lessons learned from use of rasch analysis in the assessment of outcome measures

Franco Franchignoni[1] I Andrea Giordano[2] I Xanthi Michail[3] I Nicolas Christodoulou[4]

In a time of increasing use of outcome measures in clinical practice, quality control and audit procedures, physiatrists need to acquire the necessary expertise to be able to select the appropriate tools, administer them thoughtfully, and interpret correctly the results[1].

An outcome measure is essentially an evaluative tool for assessing the magnitude of some longitudinal change (in impairment, functioning, activities, participation, etc.) in an individual or group[2]; in Physical and Rehabilitation Medicine 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 assessed by a series of 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. Basic classical test theory is still widely used in peer-reviewed, indexed journals for validating these tools, in both original and translated versions. These papers are based mainly 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). This is a superficial approach that neglects standard criteria and practical attributes that need to be considered when evaluating the psychometric properties of a measurement tool [5-11], and it does not provide information about many essential psychometric characteristics, 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 the item relative to other items in the scale, and the appropriateness of the response categories [12].

Furthermore, the caveats emerging from the use of modern (e.g. Rasch) measurement methods are often neglected or disregarded, probably due to a lack of familiarity with these methods and their results. As an example, some years ago the measurement properties of the Lequesne index of severity for osteoarthritis of the hip in elderly people were examined and major limitations were found with the convergent validity and the unidimensional structure of the measure [13]. This paper was cited also by a review on assessment of disability associated with osteoarthritis, commenting that this finding adds to the literature indicating problems with this measure [14], but the Lequesne index continues to be a favourite amongst clinicians, and several recent papers in PubMed include it as an outcome measure. The same applies for the Berg Balance Scale, a measure with a rating scale structure needing refinements [15], but still used in its original version by dozens of recent studies.

In addition, papers reporting the validation of a scale in different languages give little insight to readers if detailed methods of cross-cultural adaptation and validation are not applied [16-18].

In the last few years our group has published a number of papers reporting psychometric analyses – using both classical test theory and Rasch analysis – of outcome measures to investigate a wide range of metric characteristics [19-29].

The purpose of this paper is to summarize some basic results of these studies, in order to provide insights for selecting and/or revising outcome measures in Physical and Rehabilitation Medicine (PRM). We focus our

(1) Unit of Occupational Rehabilitation and Ergonomics - Salvatore Maugeri Foundation, Clinica del Lavoro e della Riabilitazione, IRCCS, Veruno (NO), Italy; Past-President of the UEMS PRM Board.
(2) Unit of Bioengineering - Salvatore Maugeri Foundation, Clinica del Lavoro e della Riabilitazione, IRCCS, Veruno (NO), Italy.
(3) Professor of Rehabilitation Medicine in Physiotherapy Department, Technological University, Athens, Greece; Past-President of the UEMS PRM Board & Incoming President of the European Society of PRM.
(4) School of Sciences, European University, Cyprus; President of the UEMS PRM Section & Past President of the Mediterranean Forum of PRM
Corresponding Author: Franco Franchignoni, MD, Fondazione Salvatore Maugeri, Clinica del Lavoro e della Riabilitazione, IRCCS,
Via Revislate 13, I-28010 Veruno (NO), Italy. Tel + 39. 0322-884.624. Mobile +39. 3395608857. Fax + 39. 0322-830.294. E-mail: franco.franchignoni@fsm.it.

comments on the following practical issues related to the use of outcome measures in clinical practice: 1) content validity; 2) rating scale structure; 3) cross-cultural adaptation.

1. Content validity

There are many procedures for analysing content validity of an outcome measure [12], but the most exhaustive are based on Rasch analysis methods. Rasch analysis is an original item-response theory analysis based on latent-trait modelling. Briefly, the model postulates that the probability of a person's response to each category of a rating scale item is assumed to be 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 represented by a given item, referred to as 'item difficulty' [3, 30]. 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 model conceptualizes the scale resulting from the Rasch analysis like a ruler. The same ruler is used to measure item difficulty and subject ability and it has the properties of an interval scale (i.e. is linear and quantitative, which is particularly important when measuring change and responsiveness to treatment). Conversely, 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'). The typical non-linear relationship between raw scores and Rasch-transformed measures is shown in figure 1, using as an example a scale measuring the degree of manual functioning after a unilateral upper limb amputation [23]. Moreover, item weighting is the same regardless of the difficulties or complexity inherent in the items (e.g. certain items are more difficult than others). Thus, it must be remembered that raw scores can be misleading and there is a potential for misinference when ordinal scales are used.

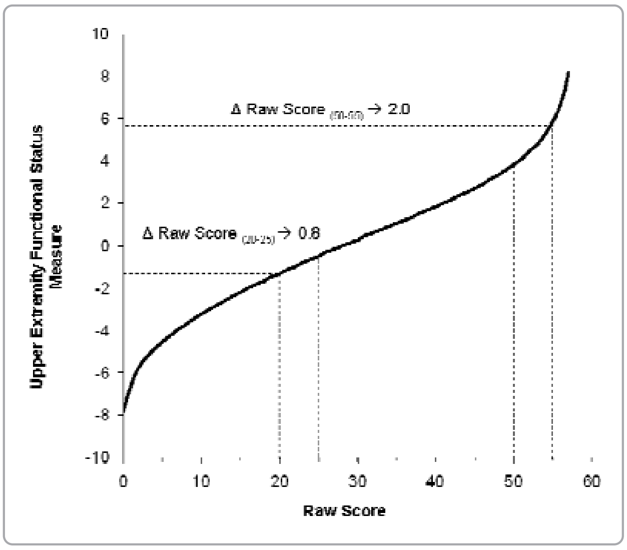


Figure 1 - Upper Extremity Functional Status module of the Orthotics and Prosthetics User Survey (OPUS): S-shaped function between raw scores (x axis) and equal-interval Rasch measures in logits (y axis). In the diagram, you can see that a change of 5 points in the raw score between 20 to 25 corresponds to a modification of 0.8 logits in the measure of manual functioning, whereas a change of 5 points between 50 to 55 corresponds to a modification of 2 logits.

Before applying Rasch analysis it is necessary to evaluate the core assumptions of the model, in particular unidimensionality (i.e. whether items are measuring one underlying dimension or several separate dimensions). Factor analysis for categorical data is usually performed to search for additional dimensions or to evaluate the fit of the scale to a unidimensional model [31]. The extent to which the model can be used to reproduce the sample data is determined by examining a series of indexes (26, 29).

In the Rasch ruler, 'subject ability' and 'item difficulty' are expressed in logit units (figure 2), a logit being the natural logarithm of the ratio (odds) of mutually exclusive alternatives (e.g. higher response vs. lower response) [30]. Moreover, Rasch analysis assesses the extent to which the observed responses to the items accord with the responses predicted by the mathematical model.

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 [32].

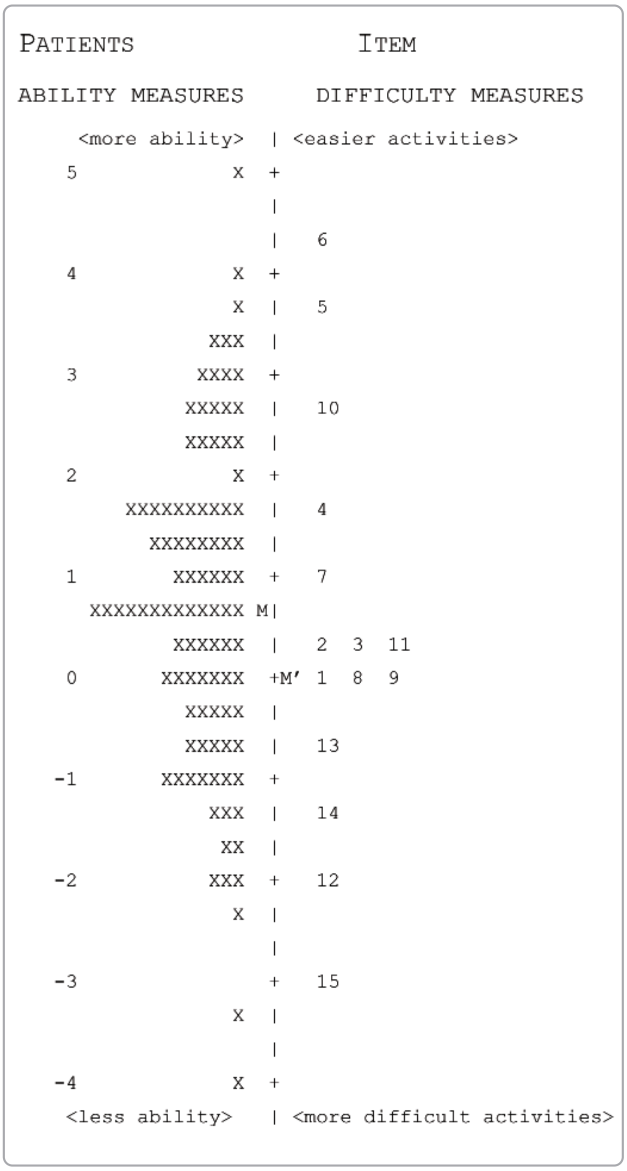
The items to consider for deletion are those that:

1) do not fit the Rasch model;
2) show redundancy, i.e. share the same span of difficulty (as items 2, 3 and 11 and items 1, 8, and 9 in figure 1), 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];
3) 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;
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; and, last but not the least,

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.



```
PATIENTS                ITEM
ABILITY MEASURES     DIFFICULTY MEASURES
     <more ability>  | <easier activities>
  5              X  +
                    |
                    |   6
  4              X  +
                 X  |   5
               XXX  |
  3           XXXX  +
             XXXXX  |   10
             XXXXX  |
  2              X  +
        XXXXXXXXXX  |   4
          XXXXXXXX  |
  1         XXXXXX  +   7
     XXXXXXXXXXXXX M|
            XXXXXX  |   2  3  11
  0        XXXXXXX  +M' 1  8  9
             XXXXX  |
             XXXXX  |   13
 -1        XXXXXXX  +
               XXX  |   14
                XX  |
 -2            XXX  +   12
                 X  |
                    |
 -3                 +   15
                 X  |
                    |
 -4              X  +
    <less ability>  | <more difficult activities>
```

Figure 2 - Patient-ability and item-difficulty maps of a hypothetical outcome measure. The vertical line represents the measure of the variable, in linear logit units. The left-hand column locates each patient's ability: each 'x' is one person. The right-hand column locates the relative difficulty of each item (indicated by its number in the scale). From bottom to top, measures indicate more ability for patients and less difficulty of items. By convention, the average difficulty of items in the test is set at 0 logits (and indicated with M', while patients with average ability are located at M).

TAKE HOME MESSAGES

c. **Raw scores can be misleading and there is a potential for misinference when ordinal scales are used**

a. **A series of validation methods must be applied to analyze validity evidence** [12]

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

c. **If data fit the model, Rasch-transformed scores are at interval-scale level**

d. **Many parameters should be considered to select the set of items with best coverage and technical quality**

e. **Data analysis is an aid to thought, not a substitute for clinical reasoning.**

2. Rating scale structure

In order to investigate whether a rating scale is being used in the intended manner, usually a procedure of ‘rating scale diagnostics’ based on Rasch analysis is applied. The performance of the response categories can be 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 by Linacre [33].

Where necessary, categories are collapsed to optimize the rating scale. 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 [30].

A typical graphic presentation of the results of ‘rating scale diagnostics’ is shown in figure 2. 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. As an example, in

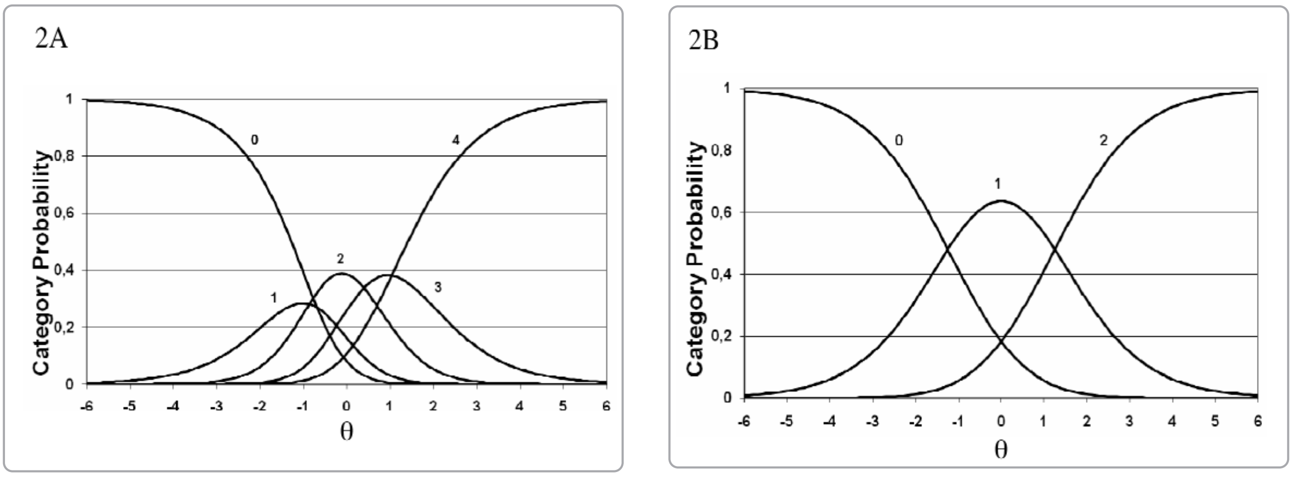


Figure 2 - Category probability curves: (A) original scale with 5 categories (0-4), (B) revised scale after collapsing categories 1 and 2, and 3 and 4 and renumbering (01122). The y-axis represents the probability (0 to 1) of responding to one of the rating categories and the x-axis represents the different performance values in logits (theta,0).

figure 2A one can see that the probability of using category 1 was never higher than that of adjacent ratings, while category 3 shows a narrow threshold difference (< 0.80 logits), whereas the plot should look (as in figure 2B, after category collapsing) like a range of hills, with each peak clearly visible.

a) Number of categories

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]. Streiner and Norman write "*It is reasonable to presume that the upper practical limit of useful levels on a scale can be set at seven. Certainly these findings clearly suggest that the 'one in a hundred' precision of the VAS is illusory; people are probably mentally dividing it into about seven segments*" [34]. Then they add "*when a large number of individual items are designed to be summed to create a scale score, it is likely that reducing the number of levels to five or three will not result in significant loss of information*". This is our experience. If it is true that healthy adults can rarely distinguish more than seven levels of a rating scale, we agree that some people with special needs would be unable to appreciably discern between more than five categories as indicating different levels of a variable. Accordingly, Smith and Wolfe state that the better fit to the Rasch rating scale model is usually obtained with three to five categories [35]. This is in agreement with other authors who state that six or more categories in a rating scale obscures the distinction between categories [36, 37]. 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.

Table 1 shows the main results of our studies on the category functioning in outcome measures with from 5 to 11 rating categories. Inevitably, Rasch analysis suggested to reduce them to 3-5 categories. In particular, when the questionnaire contains a retrospective frequency-related response scale, this is not surprising because often these categories are the so-called 'vague quantifiers' (38). This to emphasize that there is neither a formal nor an informal definition given for the meaning of these terms ('sometimes', 'rarely', 'occasionally', etc.) and as such they have fuzzy boundaries [34].

b) Central category and question wording

From the point of view of classical theory, for a bipolar scale (like a typical five-level Likert item from 'strongly agree' to 'strongly disagree') the provision of an odd number of categories should allow raters the choice of expressing a position of neutrality. Conversely, an even number of boxes forces the raters to commit themselves to one or other side (35, 39). But, when the middle category is labelled with a so-called 'non-response category' (i.e. with an indifferent, neutral or undecided phrase, such as 'Neither agree nor disagree', 'Neither dissatisfied nor satisfied', 'Does not apply' "Don't know", etc.), there is often the trend to use it as a 'dumping ground', i.e. respondents make their responses for construct-irrelevant reasons and their rating contributes construct-irrelevant variance to the measure. Thus, it seems an easy way out rather than taking time to think through to the correct answer.

For this reason, although the inclusion of negatively phrased items may theoretically control or offset acquiescence tendencies, their actual effect may be to reduce response validity. As a result, it has been argued that - from a measurement perspective - it would be better that the 'middle' alternative – if really needed - is presented aside, as the last response category, and

Table 1 - Main results regarding category collapsing in different outcome measures. Italics and brackets indicate the changes suggested by Rasch analysis

QUESTIONNAIRE	QUESTION	ORDINAL LEVELS
Locomotor Capability Index [19]	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 [22]	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 [20]	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
Geriatric Oral Health Assessment Index [24]	FREQUENCY How often have you experienced difficulties?	0 = never *1 = seldom* } *2 = sometimes* *3 = often* } *4 = always*
Parkinson's Disease Questionnaire -8 [21]	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.
Amputee Body Image Scale [16]	FEELINGS How do you see and feel about your body image? (20 questions)	1 = none of the time *2 = rarely* } *3= some of the time* *4 = most of the time* } *5 = all of the time*
Prosthesis Evaluation Questionnaire [17]	ABILITY 0-10 numeric rating scale	0-4 numeric rating scale
Fatigue Severity	SEVERITY OF SYMPTOMS	Three- to five-point response format
Scale [39]	Seven-point response format	
Disabilities of the Arm, Shoulder and Hand [26]	DIFFICULTY TO PERFORM Rate your ability to do the following activities...	0= no difficulty *1= mild difficulty* } *2= moderate difficulty* *3= severe difficulty* 4= unable

not considered for global raw score calculation.

In summary, the art of asking questions is a crucial point for an outcome measure. Both Wolfe & Smith [35] and McColl et al. (40) suggest detailed guidelines for writing rating scale items, in order to maximize the measurement validity.

c) Negatively phrased items

Even if traditional wisdom would suggest that designing a scale with an equal number of positively and negatively worded statements could obviate the problem of acquiescence bias (i.e. always agree with statements as presented), evidence shows that caution should be exercised in the use of negatively phrased attitudinal items and their inclusion may impair rather than increase the validity of survey results [41].

Moreover, when a scale contains items written in the opposite direction this could contribute to show a separate factor in factor analysis [42, 43]. As an example, in the revised version of Trinity Amputation and Prosthesis Experience Scales [25] a negatively worded item ("*I have difficulty in talking about my limb loss in conversation*") inserted in the middle of a rating scale based on a series of positively worded items was reversed ("*I find it easy to talk about my limb loss in conversation*") in order to avoid confusion for some readers.

Thus, we can conclude that further research on the impact of mixing positive and negative statements is recommended. But, at present we suggest that a mix of 'positive' and 'negative' statements in a questionnaire should be avoided whenever possible.

TAKE HOME MESSAGES
a. Rating scale structure should be as simple as possible. In most cases three to five well-selected categories are enough
b. Conversely, too many categories introduce background noise
c. Question wording has a major impact on validity and reliability of an instrument

3. Cross-cultural adaptation

The great majority of instruments have been developed in English-speaking countries and, when measures have to be used in other than the source context, there is need for a cross-cultural adaptation to the new country, culture and/or language, in order to maximise the attainment of semantic, idiomatic, experiential and conceptual equivalence between the source and target measures [44]. This means analysing many times in different ways if the instrument functions as required with 'real' target people. A correct translation process is just the first step. Full adaptation requires that scaling and psychometric properties of the new language version are assessed and compared with those of the source version, applying item response theory methods [45]. In particular, 'Differential Item Functioning' analysis is performed to test whether the difficulty hierarchy of the items is similar across versions (as well as other testing situations).

The main steps of the adaptation process usually include forward and backward translations, consensus by an expert committee, field testing of a pre-final version (with face-to-face interview with respondents: 'cognitive debriefing') and development of the final version, but many different approaches exist [46].

However, sometimes even these procedures are not enough. For example, comparing the different cross-cultural adaptations of the Disabilities of Arm, Shoulder and Hand Questionnaire (DASH), Alotaibi [47] found that some problems were identified (and solved) only in some languages. For example, item 20 'manage transportation needs' was found difficult to understand and unclear, and items 18 and 19 (regarding recreational activities) were judged to include activities unknown or infrequently used for some cultures (e.g. playing frisbee or badminton).

This example underlines that not only is there a need for high quality cross-cultural adaptation, but also for measurement tools well-designed right from the first stage of their development.

TAKE HOME MESSAGES
a. There is a need for a series of technical procedures and psychometric controls to verify equivalence of source version and its adaptation to other languages / cultures
b. Detailed processes of cross-cultural adaptation may provide useful insights for scale revision aiming at a universal applicability

Conclusion

The use of an outcome measure is an important aspect of clinical practice, audit and research. Considerable care needs to be taken to ensure that the best possible selection for the task in hand is always performed, and that, wherever possible, the selected measure conforms to modern quality standards for measurement.

In this editorial we have discussed some practical issues related to outcome measures, underlining the complexity of this field. At present, Rasch analysis 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 [12, 30].

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 using it in clinical practice, decision making or policy development.

Unfortunately, there is a generally little attention given to the theoretical framework of health outcome measures and a large variation in the methodological development and validation of commonly used tools[48].

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.

From a European perspective, we believe that in order to promote the use of outcome measures in clinical practice, decision-making and policy development there is need for strong international multidisciplinary cooperation, under the umbrella of the main European PRM bodies (European Society of Physical and Rehabilitation Medicine, UEMS PRM Section & Board, and European Academy of Rehabilitation Medicine).

Key words: outcome assessment, Rasch analysis, Physical & Rehabilitation Medicine

Referências / *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:S76-85.
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:S86-93.
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): S22-31.
9. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007 Jan;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, et al. 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;19:539-49.
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. Dawson J, Linsell L, Doll H, Zondervan K, Rose P, Carr A, et al. Assessment of the Lequesne index of severity for osteoarthritis of the hip in an elderly population. Osteoarthritis Cartilage 2005;13:854-60.
14. Pollard B, Johnston M. The assessment of disability associated with osteoarthritis. Curr Opin Rheumatol 2006;18:531-6
15. Kornetti DL, Fritz SL, Chiu YP, Light KE, Velozo CA. Rating scale analysis of the Berg Balance Scale. Arch Phys Med Rehabil 2004;85:1128-35.
16. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol 1993;46:1417-32.
17. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 2000;25:3186-91.
18. Hagell P, McKenna SP. International use of health status questionnaires in Parkinson's disease: translation is not enough. Parkinsonism Relat Disord 2003; 10: 89-92.
19. Gallagher P, Horgan O, Franchignoni F, Giordano A, MacLachlan M. Body image in people with lower-limb amputation: a Rasch analysis of the Amputee Body Image Scale. Am J Phys Med Rehabil 2007;86:205-15.
20. Franchignoni F, Giordano A, Ferriero G, Orlandini D, Amoresano A, Perucca L. Measuring mobility in people with lower limb amputation: Rasch analysis of the mobility section of the Prosthesis Evaluation Questionnaire. J Rehabil Med 2007;39:138-44.
21. Franchignoni F, Ferriero G, Giordano A, Guglielmi V, Picco D. Rasch psychometric validation of the Impact on Participation and Autonomy questionnaire in people with Parkinson's disease. Eura Medicophys 2007; 43: 451-461.
22. 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.
23. 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;40:393-399.
24. Franchignoni F, Giordano A, Ferriero G. Rasch analysis of the short form 8-item Parkinson's Disease Questionnaire (PDQ-8). Qual Life Res 2008; 17:541-548.
25. 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: 2023-2030.
26. 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; 42:323-331.
27. Franchignoni M, Giordano A, Levrini L, Ferriero G, Franchignoni F. Rasch analysis of the Geriatric Oral Health Assessment Index (GOHAI). Eur J Oral Sci 2010; 118: 278-283.
28. Gallagher P, Franchignoni F, Giordano A, MacLachlan M. Trinity Amputation and Prosthesis Experience Scales: a psychometric assessment using classical test theory and Rasch analysis. Am J Phys Med Rehabil 2010;89:487-96.
29. 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;91:1370-7.
30. Bond TG, Fox CM. Applying the Rasch model: fundamental

measurement in the human sciences. 2nd ed. Mahwah: Lawrence Erlbaum Associates; 2007.

31. Gessaroli ME, De Champlain AF. Test dimensionality: assessment of. In: Everitt BS, Howell DC, eds. Encyclopedia of Statistics in Behavioral Science. Chichester: John Wiley & Sons 2005;2014-21.

32. McHorney CA, Monahan PO: Applications of Rasch analysis in health care. Med Care 2004;42 (1 Suppl): I73-8

33. Linacre JM. Optimizing rating scale category effectiveness. J Appl Meas 2002;3:85-106.

34. Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. 2nd ed. Oxford: Oxford University Press, 1995; p. 28-53.

35. Wolfe, E. W., Smith, E. V. Jr. Instrument development tools and activities for measure validation using Rasch models: part I - instrument development tools. J Appl Meas 2007; 8: 97-123

36. Penta M, Thonnard JL, Tesio L. ABILHAND: a Rasch-built measure of manual ability. Arch Phys Med Rehabil 1998; 79:1038-1042.

37. Chen CC, Granger CV, Peimer CA, Moy OJ,Wald S. Manual ability measure (MAM-16): a preliminary report on a new patient-centered and task-oriented outcome measure of hand function. J Hand Surg Br 2005; 30:207-216.

38. Bradburn NM, Miles C. Vague quantifiers. Public Opin Q 1979; 43: 92-101.

39. Burger H, Franchignoni F, Puzic N, Giordano A. Psychometric properties of the Fatigue Severity Scale in polio survivors. Int J Rehabil Res 2010, Sep 6 [Epub ahead of print].

40. 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.

41. Schriesheim CA, Hill KD. Controlling acquiescence response bias by item reversals: the effect on questionnaire validity. Educ Psychol Meas 1981;41:1101-14..

42. Waugh RF, Chapman ES. An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: what is the difference? Which method is better? J Appl Meas 2005; 6, 80-99.

43. Spector PE, van Katwyk PT, Brannick MT, Chen PY. When two factors don't reflect two constructs: How item characteristics can produce artifactual factors. J Manage 1997;23:659-677.

44. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 2000; 25: 3186-91

45. Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, et al. ISPOR Task Force for translation and cultural adaptation. Principles of good practice for the translation and cultural adaptation process for Patient-Reported Outcomes (PRO) measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health 2005; 8: 94-104.

46. Acquadro C, Conway K, Hareendran A, Aaronson N; European Regulatory Issues and Quality of Life Assessment (ERIQA) Group. Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. Value Health 2008;11:509-21.

47. Alotaibi NM. The cross-cultural adaptation of the Disability of Arm, Shoulder and Hand (DASH): a systematic review. Occup Ther Int 2008;15:178-90.

48. 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.