Review

A short tutorial on item response theory in rheumatology

L. Siemons[1], E. Krishnan[2]

[1]Arthritis Centre Twente, Department of Psychology, Health and Technology; University of Twente, Enschede, The Netherlands;
[2]School of Medicine, Department of Medicine, Stanford University, USA.

Liseth Siemons, MSc
Eswar Krishnan, MD, MPhil

Please address correspondence and reprint requests to:
Eswar Krishnan, MD, MPhil,
Department of Medicine,
Stanford University,
1000 Welch Road, Palo Alto,
CA 94304, USA.
E-mail: e.krishnan@stanford.edu

Received on October 14, 2013; accepted in revised form on February 4, 2014.

Clin Exp Rheumatol 2014; 32: 581-586.

Key words: item response theory, basic principles, applications, rheumatology, patient-reported outcome measures, clinical measures, computerised adaptive testing

Funding: E. Krishnan was supported by NIAMS (U01AR052158-06). The views expressed herein are those of the authors and do not necessarily reflect the views of the NIH or the PROMIS program.

Competing interests: none declared.

ABSTRACT

Objectives. *The aim is to familiarise physicians and researchers with the most important concepts of item response theory (IRT) and with its usefulness for improving test administration and data collection in health care. Special attention is given to the versatility of its use within the rheumatic field.*

Methods. *This short tutorial describes the most important basic principles of item response theory, including the underlying assumptions, the model parameters, and the different models that can be applied. Practical applications are discussed to demonstrate the potential utility of IRT within clinical practice.*

Results. *IRT has proven to be useful for the development and evaluation of both clinical measures as well as patient reported outcomes used for measuring health status in observational studies and clinical trials. Promising features of IRT for the future of test administration are the assessment of local reliability and differential item functioning, the cross-cultural validation or equation of instruments, the development of large item banks, and the administration of computerised adaptive tests. These modern techniques have the ability to maximise measurement precision while simultaneously minimise response burden.*

Conclusion. *IRT provides a theoretical basis for developing alternatives to the existing tools for assessing health outcome measures in rheumatology.*

Introduction

For both clinical research and routine clinical practice valid and accurate assessment of health status is essential. In the traditional biomedical paradigm, severity and impact of rheumatic diseases were mainly being assessed with laboratory measures (1, 2), such as the erythrocyte sedimentation rate and radiographs in rheumatoid arthritis, serum urate in gout, and urine protein in lupus. In a more contemporary patient centred paradigm, patient experience has gained much interest. Some of these patient experiences can be measured objectively using, for instance, performance tests of mobility, strength, and balance. However, performance tests measure a health trait at a given moment in time, without reflecting function over time, and more importantly without assessing the trait in the context that is of relevance to the patient – his/her daily life. Furthermore, performance measurement can be influenced by biases from the observer, diurnal or temporal variation over time, and most importantly patient effort. As an alternative, one could also use the (more subjective but easier to administer) patient self-report as a metric of health status. Consequently, a wide variety of patient-reported measures have been developed over time and are currently being used, ranging from single item questionnaires like a visual analog scale measuring pain, to multi-item questionnaires as the 36-item short form health survey measuring eight dimensions of health related quality of life (3). Patient reports collect a patient's own description of his/her latent trait (*i.e.* the underlying construct of interest) without any filtering or reinterpretation on the part of the researcher. Over the past 30 years patient reports have obtained a key role in measuring health status in observational studies and clinical trials and they have shown to be useful in monitoring treat-to-target treatment strategies (4).

Traditionally, most health outcome measures have been developed using psychometrics from the classical test theory as described by Lord, Novick, Allen and Yen (5-7). Methods based on classical test theory often strive to maximise the measurement reliability (consistency) and validity (*i.e.* the extent to which the measured quantity

accurately represents the latent trait). Classical test theory has been successfully applied in research for more than 70 years and formed the foundation for measurement theory. It has been useful in the development of most of the well-known questionnaires (instruments) in use in rheumatology such as the Stanford HAQ, Medical Outcomes Short forms 36 and the Arthritis Impact Measurement Scale. These have served their purpose well and are widely used by the academia, Drug and Device development Industry, and regulatory authorities such as the US Food and Drug Administration. Nevertheless, several limitations of these questionnaires have also become apparent over time.

For instance, the scoring of these questionnaires is critically tied to the choice of questions utilised. This means that one cannot use an altered questionnaire without potentially altering the questionnaire performance characteristics. Even after reassuring oneself that the altered questionnaire is valid and reliable (which can be a time-consuming job), the scores are not automatically comparable to the parent questionnaire. Furthermore, once developed and put to use, the questionnaire is considered 'locked' and there are no easy means to reassess, revise or remove individual questions within the questionnaire for any reason (*e.g.* a questionnaire of hand function developed 40 years ago that include items based on the use of a rotary phone and the ability to thread a needle). Another consequence of these static measures is that measures are often either too long which limits their use in clinical practice and places a large burden on respondents, or too short to provide the necessary measurement precision in clinical trials. Additionally, other important limitations of classical test theory based measures are critical but mathematical issues such as the invariance of error across individuals and the rigidity of response scales. Also, some scientists argue that classical test theory mere provides a means to make a rank order of respondents along the trait as opposed to truly measuring the trait; a criticism that can be cited as its major limitation (8).

IRT might offer a solution to some of these problems (9). IRT has already been widely applied for standardised educational testing, from which it originates, but over the past decades it is gaining attention in the medical field as well given its potential to significantly improve the quality of health outcome measures. Modern psychometric techniques as IRT and computerised adaptive testing (CAT) have the ability to maximise measurement precision while simultaneously minimise response burden. Additionally, by administering patient-reported outcome measures in the form of a CAT, the number of patients required for clinical trials can be reduced while remaining an equal statistical power (10).

The goal of this review paper is to familiarise physicians and researchers with the concepts and usefulness of IRT applications. Although Tennant and Conaghan (11) already provided an overview of Rasch analysis in rheumatology this paper goes beyond the Rasch model, providing a broader perspective on the possibilities offered by IRT for improving test administration and data collection in health care. The limited number of IRT-based articles that were published in rheumatic journals over the past decade (12) emphasises the need for this paper even further.

What is item response theory?

An understanding of IRT begins with the recognition of the differences between test theories and test models. Test theories (classical test theory and IRT) provide a theoretical framework for understanding the link between the observed measure and the underlying trait. Models on the other hand operationalise these paradigms for specific situations. Thus within IRT several distinct models has been specified based on, among other things, the number of underlying dimensions or the number of response options to questions.

IRT (also known as latent trait theory) is a paradigm for developing questions and questionnaires where the focus is on the individual questions (or items) as opposed to the total questionnaires. IRT can be described as a collection of probabilistic models. It models the relation between a patient's response to a categorical item and the underlying latent trait being measured (9, 13). As an example, Figure 1 plots the probability (y-axis) of an affirmative response to a question – "can you run 1 mile?" as a function of the underlying physical function, *i.e.* latent trait (x-axis).

An IRT based approach recognises that the response of a person to a single question (item) is a mathematical function of one or more parameters of the "question" on the one hand (difficulty, discrimination, and pseudo-guessing; all explained in a later section) and the latent trait status of the "individual" on the other hand. That IRT explicitly acknowledges the presence of a latent trait, sets IRT based metrics apart from the classical test theory based ones. Additionally, IRT models provide more detailed information on the measurement precision and reliability of an instrument. Hence, they are being used in high-stakes educational testing such as the Graduate record Examination, Graduate Management Admission Test and the Law School Admission test in the United States. More recently the National Institutes of Health has rolled out the Patient Reported Outcomes Measurement Information System (PROMIS) that aims to use IRT based measures for measuring patient outcomes, and the Outcome Measures in Rheumatology (OMERACT) network initiated a special interest group in Rasch to promote the application of Rasch models in rheumatology.

Assumptions of IRT models

Before one can start analysing data with a particular IRT model, several model assumptions should be met first.

At first, most IRT models assume that each latent trait under study is unidimensional (9, 13, 14). This assumption can be assessed using various methods, including a factor analysis, independent *t*-tests comparing person scores on two subsets of items that load in opposite ways (positively *vs.* negatively) on the main component, or a comparison of a unidimensional IRT model with a multidimensional IRT model using a likelihood ratio test as discussed in Siemons *et al.* (12, 15).

A second critical assumption of IRT models is that of local independence of

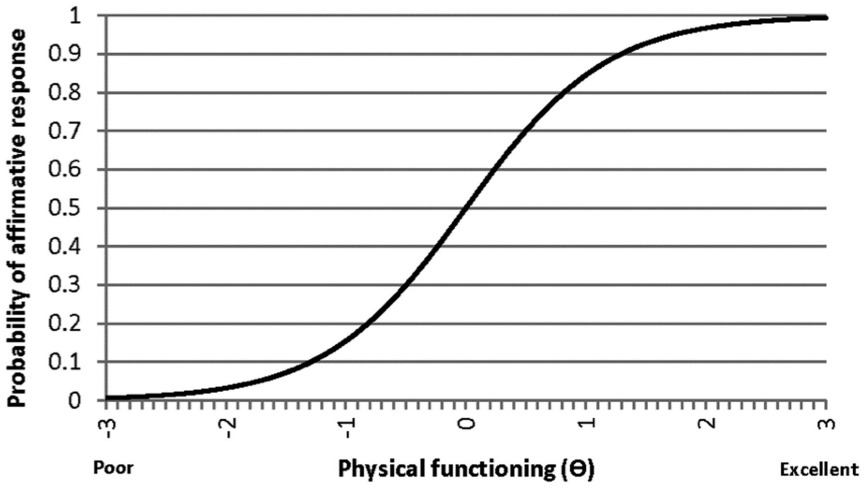

Fig. 1. A simple item characteristic curve. When a question (can you run 1 mile) is presented to an individual, the probability that an individual answers affirmatively increases with the magnitude of physical function (Θ). A person with the highest Θ, 3, is almost certainly going to answer yes whereas one with a very poor physical function will most likely answer no.

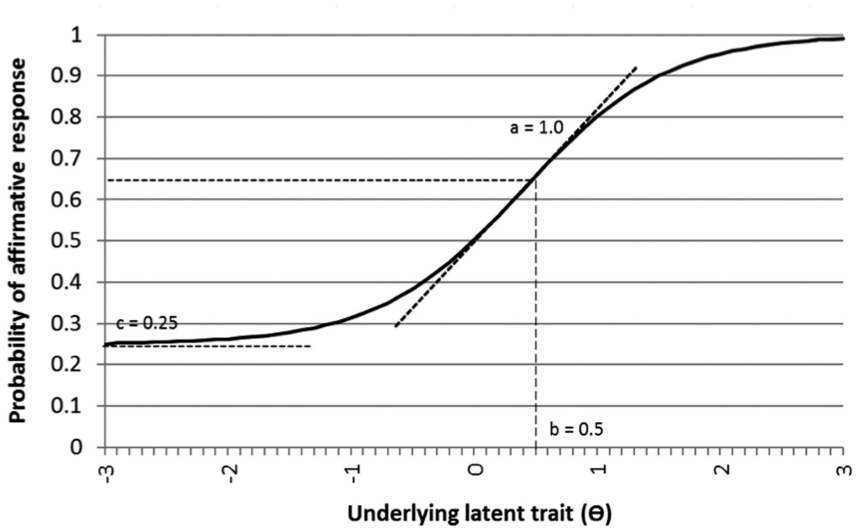

Fig. 2. A 3 parameter model. This curve shows the probability of responding to an item (y axis,) as a function of the magnitude of the underlying latent trait (Θ) on the x-axis. In the above example, the maximum slope is the measure of the discrimination parameter (parameter **a**) and is 1.0. The position on the x-axis where the slope is at its maximum is known as the location or difficulty parameter (parameter **b**). Parameter **c** is the guessing parameter. In the case of an item with 4 response options as the one above, the magnitude of the **c**-parameter is ¼ or 0.25.

items (9, 13, 14). This means that items are not related to each other except for the fact that they are measuring the same latent trait. This assumption is often violated when the items are related to similar content or if they are applied sequentially – *e.g.* an item on ability to shampoo hair followed by an item on the ability to comb hair. One method for verifying local independence is to use a confirmatory factor analysis. Excess covariation among items in the residual matrix of a single factor confirmatory factor analysis model is suggestive of violation of the assumption of local independence. Examining this matrix carefully, or looking at the modification indices associated with the one-factor solution, can reveal potential local dependence.

Finally, model-data fit should be examined to make sure the model reflects the true relationship between the underlying latent trait being measured and the item responses (9, 13, 14). Item fit statistics will show whether there are any relevant deviations between the predicted and observed response-frequencies.

Parameters of IRT models

The term parameter refers to a variable that can be used to describe a model. The best way to understand the parameters in an IRT model is to study the item characteristic curve in Figure 2. It shows the 3 parameters (difficulty, discrimination, and pseudo-guessing) that can be used to describe the item response curve of a single item. The terminology of these items might seem a bit odd in a health setting, but that is because IRT originates from educational settings. The a-parameter shows how well an item can discriminate between patients with various levels of the underlying trait; higher values mean steeper slopes and better discrimination (16). The b-parameter shows where on the underlying scale the item provides most information and it gives insight into the interrelationships of the items. Finally, the c-parameter shows the probability that one would give an affirmative answer to an item purely by random guessing. This parameter has a value of zero if no guessing is involved. By convention the populations mean value of the underlying latent trait is assigned zero and each unit of the x-axis is one standard deviation of the underlying latent trait in the population.

Types of IRT models

Originally, IRT models were developed based on normal probability distribution models also known as normal ogive models. The fundamental statistical assumption here is that the distribution of the error factor is Gaussian. However, since these models were computationally difficult prior to the advent of powerful computers, logistic models has been the preferred modelling methodology for IRT analyses.

Many different IRT models exist. Although the choice of a model is up to the user, a number of basic principles can be used to guide this decision (9, 13, 14, 17) , including whether the data is dichotomous (*e.g.* yes/no answer cat-

egories) or polytomous (Likert scales) and whether the response options are ordered or not. However, the primary distinction between these models is the number of parameters used for describing the items. The most commonly used model is the Rasch model or 1-parameter logistic model (1-PL model). A unique property of this model is that it has the ability to transform an ordinal scale to an interval scale measure when the data meet the model's expectations (15). This is the simplest model, assuming equal discrimination parameters for all items. An extension of this is the 2-parameter logistic model (2-PL model), which assumes that the items have varying abilities in discriminating among patients with different levels of the underlying latent trait. Generalisations of both models are available for polytomous data. Table I provides an overview of the most commonly used IRT models.

A less commonly used model in the medical field but widely used in educational settings is the 3-parameter logistic model, which also takes the guessing factor into account. Furthermore, all models described so far assume that the underlying latent trait is unidimensional, which means that the measured trait is just one dimension. In reality, however, several health models are multidimensional as a trait can be the result of multiple factors. For instance, when an index measure is being analysed like the Disease Activity Scale for 28 joints (18). In this case, a multidimensional IRT model might be more appropriate to use. However, these models have not received much attention in rheumatology so far (12).

Irrespective of which model will eventually be used, the model choice should always be motivated by taking into account the dimensionality, discrimination equality, and response options.

Table I. Overview of commonly applied IRT models.

Model#	Data type	Response options
Rasch / 1-PL* model	Dichotomous	–
2-PL model	Dichotomous	–
Rating scale model	Polytomous (1-PL model)	Ordered
Partial credit model	Polytomous (1-PL model)	Ordered
Generalised partial credit model	Polytomous (2-PL model)	Ordered
Graded response model	Polytomous (2-PL model)	Ordered
Modified graded response model	Polytomous (2-PL model)	Ordered
Nominal response model	Polytomous (2-PL model)	Not ordered

*PL: parameter logistic.
#For the interested reader, some of the key publications belonging to these models are: Rasch (32), 2-PL model (33), Rating scale model (34), Partial credit model (35), Generalised partial credit model (36), Graded response model (37), Modified graded response model (38), Nominal response model (39).

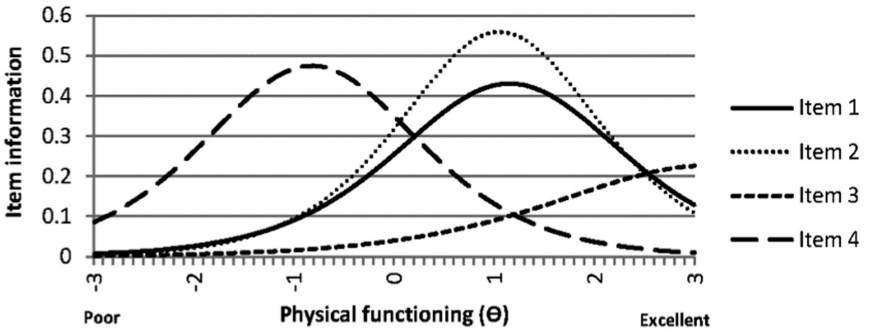

Fig. 3. Item information functions. Each curve describes along which range of Θ an item contributes most to the total instrument (*i.e.* provides the most precise measurements of a patient's physical functioning). It can be observed that item 4 measures most precise in patients with a poor physical functioning, whereas the other items perform better in patients with higher functioning levels. Item 1 and 2 function along the same range of the scale and it could be decided to remove item 1, given its lower information value. Although the information value of item 3 is low, it is still a valuable item to include since it measures best at one of the extremes of the scale.

Applications of IRT in rheumatology

Interest in IRT is increasing, not only in health care research in general (19) but also in the rheumatic field (12). After choosing an appropriate model, new measures can be developed, existing measures can be re-evaluated and improved, and alternate or short form versions can be built (20).

IRT has already proven to be very useful for the development and evaluation of patient-reported outcome measures (PROs) measuring different aspects. For instance, Conaghan *et al*. (21) used a Rasch model to evaluate the psychometric quality of the Oxford Knee Scale, whereas Helliwell *et al*. (22) used it to developed a new questionnaire called the Foot Impact Scale for measuring the foot status in rheumatoid arthritis patients and Davis *et al*. (23) evaluated the measurement properties of the Western Ontario McMaster (WOMAC) Osteoarthritis Index.

However, its use is not limited to PROs. Clinical measures can be assessed as well. Siemons *et al*. (24) used the generalised partial credit model to evaluate whether the assessment of forefoot joints could improve the measurement range and measurement precision of the 28-tender and 28-swollen joint counts. Bode *et al*. (25) used IRT to examine psychometric properties of a disease activity scale for children with juvenile dermatomyositis, and Wolfe *et al*. (26) developed a Short Erosion Scale to measure radiographic severity.

As shown by Siemons *et al*. (12), the main focus within rheumatology is on patient-reported outcome measures (PROs). Cross-sectional Rasch modelling clearly prevails and is mostly being used for developing or evaluating existing PROs measuring the patient's physical functioning or quality of life. Most of this research has been carried out in the US and UK and especially on rheumatoid arthritis or osteoarthritis patients (12). However, IRT is not yet being used at its full capacity within the rheumatic field. A similar conclusion was drawn by Leung *et al*. (15), who focused on the application of the Rasch model over time. Although the

Rasch model has been increasingly used during the past decades, complex features like item banking and CAT applications, but also more basic aspects as differential item functioning and reliability testing, offer promising future research directions.

IRT methods are very suitable for assessing differential item functioning (DIF), which is present when patients from different groups (*e.g.* gender, age, countries/cultures), but with equal scores on the latent trait, do not have the same probability of responding to an item (9, 13). For instance, when a male and female patient with the same level of physical functioning do not have the same probability of answering affirmatively to the question whether they are able to walk 1 mile. Although the assessment of DIF across gender and age has been widely recognised (12), its use for the cross-cultural validation of outcome measures (*i.e.* to determine whether different language versions of the same outcome measure function equally (27)) is still very limited. If DIF across countries is absent scores from different countries will be comparable, but when DIF is present country-specific adjustments should be made to correct for these cultural differences in order to make scores comparable.

Apart from checking whether "the same" outcome measure functions equally over countries, IRT can also be applied to equate "different" outcome measures measuring the same construct. To illustrate this, think of all the measures which are available to assess a patient's physical functioning (28). Scores on these separate instruments are not comparable since they include (a) different (number of) items and response options and they cover different aspects of the physical functioning construct. IRT helps making these scores comparable to each other by equating the measures, which means that it places the scores from these different instruments on a common metric.

Another key feature of IRT which lacked attention so far is that not only the global reliability (Cronbach's alpha) of an instrument can be obtained, but also the local reliability. So called test and item information functions provide information about the range of the underlying latent trait where the instrument and the items provide most precise and reliable measurements and discriminate best among individual patients (9, 13). You may think of it as a weighing scale. Some weighing scales measure very precise in the lighter regions (*e.g.* from 0–2 kg), whereas others are meant for weighing heavier objects (*e.g.* from 20–150 kg). Likewise, some items are especially relevant for people with minor complaints, whereas others are more relevant for people with major complaints. A physician needs to be able to measure well in all patients, not just in part of them. By examining item information functions the best items can be selected and floor and ceiling effects can be reduced to a minimum by making sure to include the items which provide important information at the extremes of the scale. Where classical test theory would often remove items at the extremes of the underlying latent scale because almost all patients answered it either at the lowest or at the highest response category, IRT methods include these items (Fig. 3) (29). When all the item information functions are taken together, a test information function is being obtained, which shows for which patients the instrument as a whole gives the best estimates of their underlying trait level.

Item selection using information functions is also particularly useful for the development of large item banks and computerised adaptive tests (9, 19, 30, 31). When developing a patient-reported outcome measure, it is important to cover the whole spectrum of the concept you are interested in. This could mean that you would have to include many items, which will increase the patient's administration burden significantly. However, when all these items are collected into large item banks, which can be a hard and time-consuming job and requires large sample sizes for evaluating the psychometric properties of all the items (13), computerised adaptive tests (CATs) can be developed to bring relief. A CAT provides every patient a test which is tailored to his or her level on the underlying latent trait being measured (9). As a result, each patient will answer a different number and sequence of items drawn from the item bank, but IRT offers a framework which enables one to compare the resulting latent trait estimates of the individual patients. The items which a patient receives depend on his or her answers on prior items in the test. This process continues until a specified stopping rule has been reached, for instance when a desired level of measurement precision has been obtained. This assessment technique is often referred to as the promise of IRT and the future of test administration.

From a practical perspective, applications of IRT based measures using CATs offer tremendous improvement from the existing methods of measuring and interpreting results of clinical trials. By equating measures, changes in patient reported outcomes such as physical function are being set on the same scale and, consequently, can be compared across trials. This can enable more reliable meta-analyses and subgroup analyses among those with varying measures of the underlying trait. Lastly, the use of IRT measures and CAT can also remove one of the barriers in comparing and integrating data from clinical trials and observational studies.

Conclusion

It is important to realise that applying IRT based methods does not imply abandoning classical test theory; they are two distinct statistical methods and they can be used together, depending on the research questions. As has been shown, however, the growth potential of IRT within clinical medicine is high. IRT has been proven to be very useful for the development and evaluation of patient-reported outcome measures as well as clinical measures, measuring different aspects and domains. Promising features of IRT that were discussed in this review and which may be the future of test administration include the assessment of local reliability and differential item functioning, the cross-cultural validation or equation of instruments, the development of large item banks, and the administration of computerised adaptive tests. Although this

paper showed only part of the versatility of IRT modelling; its concept, usefulness, and possibilities for improving test administration and data collection in health care should be clear by now. The next step is to start using it.

References

1. FRIES JF: The promise of the future, updated: better outcome tools, greater relevance, more efficient study, lower research costs. *Fut Rheumatol* 2006; 1: 415-21.
2. FRIES JF, BRUCE B, CELLA D: The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. *Clin Exp Rheumatol* 2005; 23: S53-S7.
3. WARE JE, SHERBOURNE CD: The MOS 36-item short form health survey (SF-36): I. Conceptual framework and item selection. *Med Care* 1992; 30: 473-83.
4. CASTREJÓN I, PINCUS T: Patient self-report outcomes to guide a treat-to-target strategy in clinical trials and usual clinical care of rheumatoid arthritis. *Clin Exp Rheumatol* 2012; 30 (Suppl. 73): S50-5.
5. LORD FM, NOVICK MR: Statistical theories of mental test scores. Reading, MA, Addison-Welsley Publishing Company, 1968.
6. NOVICK MR: The axioms and principal results of classical test theory. *J Math Psychol* 1966; 3 1-18.
7. ALLEN MJ, YEN WM: Introduction to Measurement Theory. Long Grove, IL, Waveland Press, 2002.
8. LUCE RD, TUKEY JW: Simultaneous conjoint measurement: A new type of fundamental measurement. *J Math Psychol* 1964; 1: 1-27.
9. HAMBLETON RK, SWAMINATHAN H, ROGERS HJ: Fundamentals of item response theory. Newbury Park, CA, Sage Publications, 1991.
10. FRIES JF, KRISHNAN E, ROSE M, LINGALA B, BRUCE B: Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. *Arthritis Res Ther* 2011; 13 (R147).
11. TENNANT A, CONAGHAN PG: The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a rasch paper? *Arthritis Rheum* 2007; 57: 1358-62.
12. SIEMONS L, TEN KLOOSTER PM, TAAL E, GLAS CAW, VAN DE LAAR MAFJ: Modern psychometrics applied in rheumatology - a systematic review. *BMC Musculoskelet Disord* 2012; 13: 216.
13. REEVE BB, FAYERS P: Applying item response theory modeling for evaluating questionnaire item and scale properties. *In* FAYERS P, HAYS RD (Eds.): *Assessing Quality of Life in Clinical Trials: Methods of Practice*. Oxford, NY, Oxford University Press, 2005: 55-73.
14. ORLANDO M: Critical issues to address when applying item response theory (IRT) models. Conference on Improving Health Outcomes Assessment Based on Modern Measurement Theory and Computerized Adaptive Testing. Bethesda, MD, Hyatt, 2004.
15. LEUNG Y-Y, PNG M-E, 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: 159-64.
16. BAKER FB: The basics of item response theory. College Park (MD): ERIC Clearinghouse on Assessment and Evaluation, 2001.
17. EMBRETSON SE, REISE SP: Item response theory for psychologists. Mahwah, NJ, Lawrence Erlbaum Associates, 2000.
18. PREVOO ML, VAN 'T HOF MA, KUPER HH, VAN LEEUWEN MA, VAN DE PUTTE LB, VAN RIEL PL: Modified disease activity scores that include twenty-eight-joint counts: development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. *Arthritis Rheum* 1995; 38: 44-8.
19. HAYS RD, MORALES LS, REISE SP: Item response theory and health outcomes measurement in the 21st century. *Med Care* 2000; 38: II28-II42.
20. FRIES JF, BRUCE B, BJORNER J, ROSE M: More relevant, precise, and efficient items for assessment of physical function and disability: moving beyond the classic instruments. *Ann Rheum Dis* 2006; 65 (Suppl. III): iii16-iii21.
21. CONAGHAN PG, EMERTON M, TENNANT A: Internal construct validity of the Oxford knee scale: Evidence from Rasch measurement. *Arthritis Care Res* 2007; 57: 1363-7.
22. HELLIWELL P, REAY N, GILWORTH G *et al*.: Development of a foot impact scale for rheumatoid arthritis. *Arthritis Rheum* 2005; 53: 418-22.
23. DAVIS AM, BADLEY EM, BEATON DE *et al*.: Rasch analysis of the Western Ontario McMaster (WOMAC) Osteoarthritis Index: results from community and arthroplasty samples. *J Clin Epidemiol* 2003; 56: 1076-83.
24. SIEMONS L, TEN KLOOSTER PM, TAAL E *et al*.: Contribution of assessing forefoot joints in early rheumatoid arthritis patients: Insights from item response theory. *Arthrit Care Res* 2013; 65: 212-9.
25. BODE RK, KLEIN-GITELMAN MS, MILLER ML, LECHMAN TS, PACHMAN LM: Disease activity score for children with juvenile dermatomyositis: Reliability and validity evidence. *Arthritis Rheum* 2003; 49: 7-15.
26. WOLFE F, VAN DER HEIJDE DM, LARSEN A: Assessing radiographic status of rheumatoid arthritis: introduction of a short erosion scale. *J Rheumatol* 2000; 27: 2090-9.
27. TENNANT A, PENTA M, TESIO L *et al*.: Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PROESOR project. *Med Care* 2004; 42: I37-48.
28. MCHORNEY CA, COHEN AS: Equating health status measures with item response theory: illustrations with functional status items. *Med Care* 2000; 38: II-43-II-59.
29. TENNANT A, P. MS, HAGELL P: Application of Rasch analysis in the development and application of quality of life instruments. *Value Health* 2004; 7: S22-S6.
30. MCHORNEY CA: Generic health measurement: Past accomplishments and a measurement paradigm for the 21st century. *Ann Intern Med* 1997; 127: 743-50.
31. REVICKI DA, CELLA DF: Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. *Qual Life Res* 1997; 6: 595-600.
32. RASCH G: Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut; 1960.
33. LORD FM: A theory of test scores. (Psychometric Monograph No. 7). Iowa City, IA, Psychometric Society, 1952.
34. ANDRICH D: A rating formulation for ordered response categories. *Psychometrika* 1978; 43: 561-73.
35. MASTERS GN: A Rasch model for partial credit scoring. *Psychometrika* 1982; 47: 149-74.
36. MURAKI E: A generalized partial credit model: Application of the EM algorithm. *Applied Psychological Measurement* 1992; 16: 159-76.
37. SAMEJIMA F: Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs* 1969; 34: No. 17.
38. MURAKI E: Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement* 1990; 14: 59-71.
39. BOCK RD: Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 1972; 37: 29-51.