Musculoskeletal Science and Practice 83 (2026) 103544

Contents lists available at ScienceDirect

Musculoskeletal Science and Practice

journal homepage: www.elsevier.com/locate/msksp

Rasch analysis in the development of self-reported outcome measures to assess physical function

Andrés Pierobon [a,*], Chris Krägeloh [b]

[a] Department of Primary Health Care, University of Otago, Wellington, New Zealand
[b] Auckland University of Technology, Auckland, New Zealand

A B S T R A C T

Introduction: The assessment of physical function is central to clinical decision-making in rehabilitation and musculoskeletal care. Patient-reported outcome measures (PROMs) are widely used because they are simple, cost-effective, and patient-centred. However, many PROMs were developed using Classical Test Theory, which assumes equal distances between ordinal response options and overlooks differences in item difficulty and person ability. These limitations can reduce measurement precision and cause ceiling effects, particularly among individuals with high physical function. Rasch analysis, a modern psychometric approach based on Item Response Theory, addresses these issues and enhances the measurement properties of PROMs.

Purpose: This article introduces Rasch analysis as a methodological framework for developing and refining PROMs to assess physical function. It explains the principles of the Rasch model, its application to dichotomous and polytomous data, and how it transforms ordinal scores into interval-level measurements. Example figures illustrate key outputs such as category probability curves, person–item maps, and threshold ordering. Advantages, limitations, and practical considerations for integrating Rasch analysis into outcome measure development are discussed.

Implications: Rasch analysis enables clinicians and researchers to better understand item difficulty and estimate patients' functional ability with greater precision. Incorporating Rasch-developed PROMs enhances the validity, interpretability, and responsiveness of functional assessments. Clinicians can use these measures with increased confidence when monitoring progress and evaluating treatment outcomes, supporting more accurate goal setting and improved rehabilitation practice.

1. Introduction

Physical function, defined as the ability to perform basic activities of daily living that are essential for maintaining independence and carrying out more complex activities, is a core domain included in most musculoskeletal outcome sets (Sabet et al., 2025), and its assessment is recommended by most clinical guidelines (Lin et al., 2020). Physical function can be assessed using performance-based outcome measures or subjectively using patient-reported outcome measures (PROMs) (Pierobon et al., 2025). The assessment of physical function using PROMs is common practice among physiotherapists and other healthcare professionals as these are relatively easy to administer, quick, cost-effective, and, in most cases, provide valid and reliable results to estimate respondents' physical function (Chiarotto, 2019; Cook et al., 2021; Lam et al., 2020; Pierobon et al., 2025). PROMs are crucial for understanding the impact of musculoskeletal conditions and the effectiveness of various treatment programs, and are commonly used in clinical trials as primary outcome measures (Black, 2013; Blasco et al., 2020; Froud et al., 2016; Page et al., 2015; Reiter et al., 2024).

PROMs have known disadvantages related to their self-report nature, including vulnerability to response bias and the influence of external factors (such as culture and language) (Cook et al., 2021). There are also other common limitations related to the development of these types of measures, such as a lack of challenging activities that can lead to a ceiling effect and an incomplete coverage of the physical function range (Eckhard et al., 2021; Hysing-Dahl et al., 2025). Furthermore, traditional sum-scoring methods assume equal item difficulty and treat ordinal responses as interval-level data, assumptions that are rarely tested and may not hold in practice, potentially compromising measurement precision (Tennant and Conaghan, 2007).

The majority of the PROMs commonly used to assess physical function, such as the Knee injury and Osteoarthritis Outcome Score (KOOS) for knee conditions (Roos et al., 1998), the Roland Morris questionnaire for back pain (Roland and Morris, 1983), and the Upper Extremity Functional Index (UEFI) for upper limb conditions (Gabel et al., 2006), were developed based on Classical Test Theory (CTT), which has been dominating test construction and validation since the early 20th century (Siegert et al., 2025; Thomas, 2019). CTT focuses on test-level properties such as reliability, validity, and responsiveness, typically employing methods like exploratory and confirmatory factor analysis to establish

* Corresponding author. University of Otago, 23A Mein Street, Newtown, Wellington, 6242, New Zealand.
E-mail address: andi.pierobon@postgrad.otago.ac.nz (A. Pierobon).

https://doi.org/10.1016/j.msksp.2026.103544
Received 17 October 2025; Received in revised form 9 March 2026; Accepted 12 March 2026
Available online 14 March 2026

the psychometric properties of measurement instruments (de Vet et al., 2011). Publications using Rasch analysis for the development and analysis of PROMs have increased steadily year on year over the past 25 years, and Rasch analysis is now a well-established measurement approach (Belvedere and de Morton, 2010; Mallinson et al., 2022). However, while CTT is more widely understood and applied by clinicians and researchers, familiarity with Rasch analysis remains limited (De Champlain, 2010). This article aims to introduce Rasch analysis briefly and discuss the advantages and potential disadvantages of using this approach to develop PROMs for assessing physical function.

2. What is Rasch analysis?

Rasch analysis is a measurement approach from the family of the item response theory (IRT), developed to improve the precision of PROMs, assess instrument quality, and compute respondents' performance (Boone, 2016). IRT encompasses a family of probabilistic models that describe the relationship between respondents and their probability of responding in a particular way to individual items (Engelhard and Wang, 2025). Unlike CTT, which evaluates item primarily through aggregate statistics such as item-total correlations and can accommodate multidimensional structures through composite or domain subscale scoring, IRT models explicitly specify an item-level mathematical relationship between individual items and a latent trait. In contrast to other IRT models that estimate item-specific discrimination parameters, Rasch analysis is a single-item-parameter model in which item discrimination is constrained to be equal and item difficulty is the only estimated item parameter. Rasch analysis, like other IRT models, is based on the assumption of unidimensionality, whereby all items of a PROM are intended to measure a single underlying latent construct. For measures of physical function, this implies that variation in item responses should primarily reflect differences along a single continuum of physical function rather than multiple distinct traits (Brentari and Golia, 2007).

Rasch analysis can be applied with dichotomous measures (e.g., yes/no, correct/incorrect) such as the Roland-Morris questionnaire (Davidson, 2009) or with polytomous measures (e.g., Likert scales) such as the KOOS (Franchignoni et al., 2013; Tennant and Küçükdeveci, 2023). For dichotomous measures, the Rasch analysis is a probabilistic model that states that the probability of passing or failing a test depends on the difference between a person's ability and the difficulty of the items. For example, if the test is standing up from a chair, a person with much higher ability than the task difficulty will have a high probability of success, a person with much lower ability will have a low probability of success, and a person whose ability is about the same as the task difficulty will have around a 50% chance of success (Fig. 1).

In polytomous measures, the Rasch model estimates the probability of the transition from one category (such as a response option in a 5-point Likert scale) to the next, with this transition point being known as the threshold. The item threshold refers to the level of a latent trait when the probability of choosing one of two subsequent response options is the same (Fig. 2). When item thresholds are arranged in an orderly way, progressing from the lowest to the highest value in both dichotomous and polytomous scales, they are said to fit the Rasch model (top panel of Fig. 2). An example of disordered thresholds is shown at the bottom panel of Fig. 2.

Many of the most common PROMs to assess physical function developed using CTT, such as the KOOS and Roland Morris questionnaire, have been re-analysed using Rasch analysis (Comins et al., 2008; Kent et al., 2015). It is common for such studies to have found subsequent misfit to the Rasch model, leading to the proposal of new versions, with some PROMs having more than two different versions (Comins et al., 2008; Davidson, 2009; Franchignoni et al., 2013; Perruccio et al., 2008; Soh et al., 2021). Using Rasch analysis in the development of the PROMs, rather than for re-assessment, could be helpful for building psychometrically robust measures from the very beginning. This is because Rasch analysis enables researchers to scrutinise scales in detail,

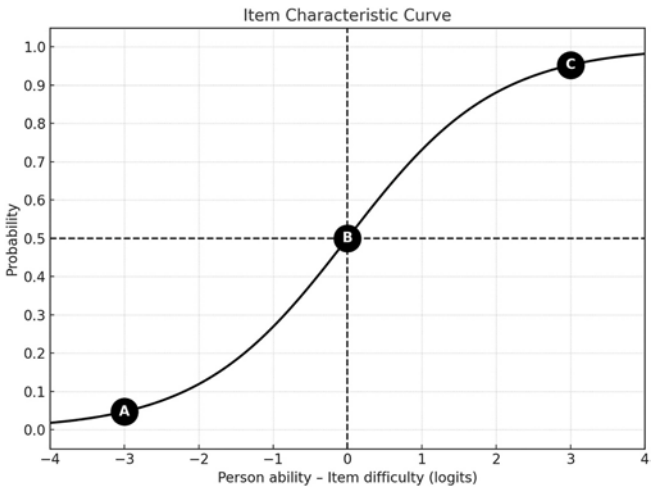

Fig. 1. Item characteristic curve in a dichotomous Rasch model. The curve represents the probability of a person successfully completing an item (e.g., standing up from a chair) (Y axis) as a function of their ability relative to the item's difficulty (X axis), both measured on the same logit scale. When a person's ability equals the item difficulty (e.g., at 0 logits), the probability of success is 50% (B). As the person's ability increases beyond the item difficulty, the probability of success increases (C). Conversely, when the person's ability is lower than the item difficulty, the probability of success decreases (A). The curve reflects the logistic function that underlies the Rasch model, demonstrating how the probability of success changes continuously across the latent trait.

looking at a range of indicators, as outlined below.

3. What are the advantages of Rasch analysis?

Rasch analysis offers several advantages when applied to both dichotomous and polytomous outcome measures: 1) It places estimates of item difficulty and person ability on a common metric (the logit scale), enabling direct comparison between item demands and respondent capabilities; 2) it provides detailed fit statistics that identify misfitting items and aberrant response patterns; 3) it provides information on the way that items may be interpreted differently by different groups (differential item functioning); and 4) when the data fit the model, it enables the transformation of ordinal raw scores into interval-level measures, which permits direct comparisons of scores from different groups (Boone, 2016; Tennant and Küçükdeveci, 2023). The first three advantages are not unique to the Rasch model and are also features of other IRT approaches, such as two-parameter models.

3.1. Item difficulty and person ability on a common scale

PROMs assessing physical function should be comprehensive and cover the entire spectrum of physical abilities (de Vet et al., 2011). Including items that range from very easy to very challenging activities ensures that the measure can accurately capture functional limitations across varying levels of ability within the target population. This can potentially minimise floor and ceiling effects of the total or sub-scale score, which can occur when the items are too easy or too difficult for certain individuals, limiting the instrument's ability to detect change or distinguish between patients with different levels of ability. PROMs with broad item content enhance the generalisability of findings across different populations and clinical contexts, making them more suitable for use in both clinical practice and research. Many of the most commonly used PROMs to assess physical function (both for lower and upper limbs) have reported ceiling effects (Eckhard et al., 2021; Hsu et al., 2010; Jo et al., 2021; Ra et al., 2014; Saithna and Cote, 2024), likely due to the absence of a sufficient number of items inquiring about

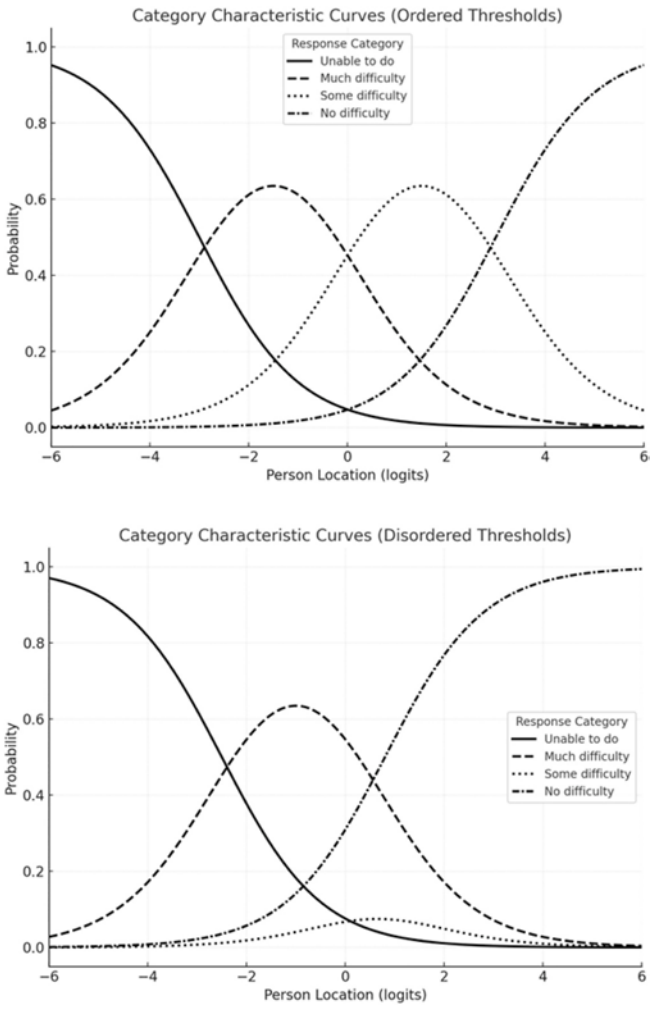

Fig. 2. Likert scale responses transition. The figure displays the category characteristic curves for each response option (four options), showing the probability of endorsing a given category across different levels of the latent trait. The top panel represents ordered thresholds, while the bottom panel represents disordered thresholds, with the response "some difficulty" never being the most likely option along the latent trait.

challenging activities (Pierobon et al., 2025).

Rasch analysis estimates thresholds between response options of an item while simultaneously estimating both item difficulty and person abilities, with both parameters expressed on the same logit scale. This common metric enables direct comparison of where persons and items are located relative to one another. A person-item threshold distribution plot provides a visual representation of how well the range of individual abilities on a latent trait is covered by the range of items (Fig. 3). When item thresholds representing the latent construct adequately cover sample abilities, this indicates good targeting of the instrument and with minimal floor and ceiling effects (Medvedev and Krägeloh, 2025). Clinically, this information allows practitioners to determine whether a PROM is appropriate for a specific patient population. For example, if most patients in a sports rehabilitation setting cluster above the highest item thresholds, the measure may lack sufficient challenging activities and fail to detect improvement. Conversely, poor targeting in the lowest spectrum of ability may result in floor effects and reduced sensitivity to functional decline. Gaps in the person-item threshold distribution reveal areas where additional items of specific difficulty levels may be needed to improve the instrument's precision and coverage of the construct. Using Rasch analysis in the initial stages of the development process allows for a detailed assessment of the difficulty of the items in the context of all candidate items, which can eventually lead to the removal of redundant items or the inclusion of items if gaps in difficulty are present.

3.2. *Identification of misfitting items and persons*

Rasch analysis provides detailed fit statistics (e.g., infit and outfit mean square residuals) that identify both items and respondents whose response patterns deviate significantly from the expectations of the model (Tennant and Küçükdeveci, 2023). While factor analytic approaches can also identify problematic items through low factor loadings or high residual correlations (Tennant and Conaghan, 2007), Rasch fit statistics offer complementary information by evaluating whether individual response patterns conform to the probabilistic expectations of the measurement model. Importantly, item misfit in Rasch analysis can serve as an indicator of potential multidimensionality, signalling items that may tap into constructs other than the intended latent variable and thereby threatening the unidimensional structure required for valid summation of item scores (Tennant and Conaghan, 2007).

The above-mentioned features are particularly valuable for refining the PROM by highlighting items that may be ambiguous, poorly worded, or not aligned with the underlying construct of physical function. Additionally, misfitting person responses can indicate individuals who may not have engaged meaningfully with the questionnaire, for example, those who answered randomly, did not pay attention, misunderstood the questions, or intentionally provided misleading responses (Felt et al., 2017). Unlike factor analysis, which operates at the aggregate level, Rasch analysis enables identification of aberrant response patterns at the individual respondent level (Tennant and Conaghan, 2007). Identifying these misfitting responses allows researchers to assess data quality and consider excluding such cases from analysis, thereby

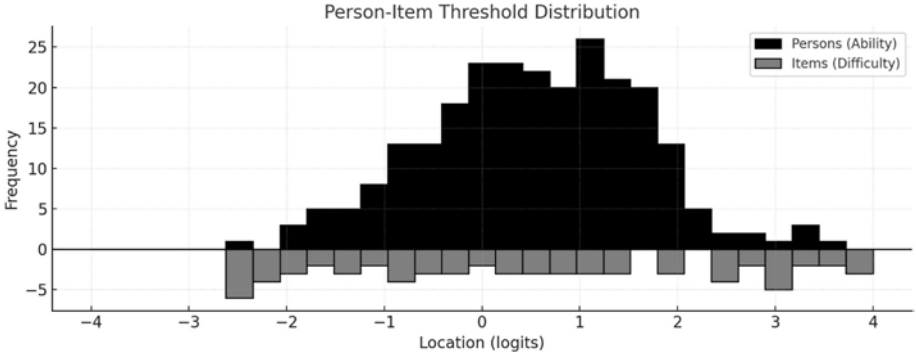

Fig. 3. Person-item threshold distribution plot. This figure displays the alignment between the distribution of participants' abilities (black columns) and the distribution of item difficulties (grey columns) along the same logit scale. Locations on the right part of the scale indicate greater ability (for persons) or higher difficulty (for items). Locations on the left part of the scale indicate lower ability (for persons) or lower difficulty (for items).

improving the reliability and validity of the final instrument. Moreover, repeated detection of misfitting items across samples may suggest content that requires revision or removal to enhance the measurement properties of the measure.

3.3. Evaluation of differential item functioning

One of the key strengths of Rasch analysis, also common to other IRT models, is the ability to evaluate differential item functioning (DIF), which occurs when individuals from different subgroups (e.g., age, gender, height) with the same underlying level of physical function respond differently to a particular item (Hagquist and Andrich, 2017; Tennant and Küçükdeveci, 2023; Tesio et al., 2024b). In the context of PROMs assessing physical function, DIF analysis is essential to ensure that items are not biased or unfairly favour one group over another. For example, an activity such as "hopping forward" might appear more difficult for short people in comparison with tall people, not because of their physical function per se, but due to differences in limb length that affect hop distance. Detecting and addressing DIF helps ensure that the PROM yields comparable and valid scores across diverse patient populations, which is critical for equitable clinical decision-making, accurate assessment of functional status, and valid comparisons in research settings. By flagging items that exhibit significant DIF, Rasch analysis allows researchers to revise or remove problematic items, thereby enhancing the fairness, generalisability, and interpretability of the measure.

3.4. Transformation of ordinal scores into interval-level data

Most PROMs used by clinicians and researchers to assess physical function comprise multiple items with different levels of responses (e.g., "no difficulty", "some difficulty", "much difficulty", "unable to do") that generally assess the difficulty of engaging in several activities. Each response has a score, and all the scores are summed to generate a total score for a particular measure. The assumption that the difference in value between responses is exactly the same is fundamentally incorrect. Although it is indisputable that the option "much difficulty" represents a higher level of difficulty than "some difficulty", the true quantitative difference between these two options is unknown and is instead expressed by the ordinal score that the developer of the measure has assigned to the response options (Fig. 4). Furthermore, assuming that each item contributes equally to the total score (i.e., same difficulty) is also not warranted. This means that PROMs developed using CTT are essentially ordinal scales, where the order of the responses is known but the magnitude of the differences between them is unknown.

Using ordinal scales as if they were interval scales can lead to significant measurement and interpretation errors (Bishop and Herron, 2015; Liddell and Kruschke, 2018; Sönning, 2024). To avoid this issue, Rasch analysis can transform ordinal data into interval-level measures (Boone, 2016). The mathematical basis for this property lies in the Rasch model's unique requirement that raw scores serve as sufficient statistics for person ability estimates—meaning no information is lost in summing item responses—combined with specific objectivity, whereby person and item parameters are estimated independently on a common logit scale (Andrich, 1988; Perline et al., 1979). These properties enable the construction of invariant interval-level scales. In contrast, other IRT model, such as the two-parameter and three-parameter models, which include item discrimination and guessing parameters, sacrifice specific objectivity: Item difficulties become dependent on the particular sample used for estimation, precluding the same claims of invariant interval-level measurement (Engelhard and Wang, 2025). The production of ordinal-to-interval algorithms constitutes the final step of the psychometric analyses, as recommended by commonly accepted Rasch guidelines (Leung et al., 2014). Users of the scale can then be guided to conversion rules through instructions or software syntax files that illustrate how each raw ordinal score is to be changed to an interval score (Medvedev et al., 2018).

By converting these ordinal responses into interval-level data, where the difference between each point on the scale is equal, Rasch analysis enables more accurate and valid statistical analyses. This is particularly valuable when comparing results over time—such as before and after an intervention—because it allows for valid calculations of change, effect sizes, and responsiveness that would not be appropriate with raw ordinal scores. Additionally, differences in the interpretation of items detected through DIF can be adjusted so that different groups can be compared directly. For example, if one group of participants with certain shared characteristics (such as gender or height) generally finds a particular item more difficult than another group, separate ordinal-to-interval tables can be calculated for the two groups. This adjusts for differences in interpretation and ensures the new converted scores are on the same metric.

4. What are the potential disadvantages of Rasch analysis?

4.1. Complexity and technical expertise requirements

One of the main challenges of using Rasch analysis is that it requires advanced statistical knowledge and the use of specialised software (such as RUMM 2030, Winsteps, or R packages like eRm or TAM) (Tennant and Küçükdeveci, 2023). This can be a barrier for researchers or

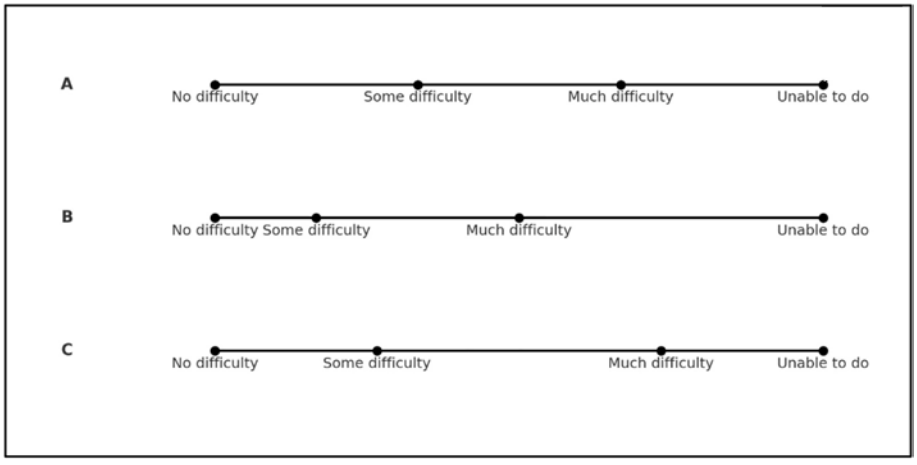

Fig. 4. Ordinal versus interval scaling. The scaling in Item A represents an unlikely perfect jump between categories used in PROMs developed using Classical Test Theory. The real responses are more likely to behave like Item B or C, where the distance between categories is not the same.

clinicians who are not trained in psychometrics or who do not have access to statistical support. In addition, the results produced by Rasch analysis—such as item difficulty and person ability estimates (called "logits"), item-person maps, and fit statistics—are not always easy to understand for people unfamiliar with the method (Linacre, 2006). These outputs often require interpretation by experts and translation into practical terms to be clinically useful. For example, clinicians may find it difficult to make decisions based on a person's logit score unless this score is linked to meaningful functional categories or cut-offs. Without clear guidance on how to interpret and apply the results, there is a risk that the benefits of Rasch analysis—such as its precision and fairness—may not be fully realised in everyday clinical practice. Therefore, collaboration between psychometricians, researchers, and clinicians is essential to ensure that complex Rasch outputs are translated into simple information that supports patient care and decision-making. Recently, research papers have been published and guidelines developed for helping readers to understand Rasch analysis specifically in the field of rehabilitation (Mallinson et al., 2022; Tennant and Küçükdeveci, 2023; Tesio et al., 2024a, 2024b).

4.2. Statistical fit overweighting clinical relevance

Before psychometric evaluation using Rasch analysis or other methods, it is essential that PROMs undergo rigorous assessment of content validity. Content validity refers to the extent to which a PROM's items are relevant, comprehensive, and comprehensible from the perspective of patients with the target condition. This qualitative evaluation typically involves directly consulting patients to determine whether items adequately capture the construct of interest and whether important aspects are missing or unclear. Terwee et al. (2018) developed consensus-based standards for systematically evaluating PROM content validity, encompassing criteria for item relevance, appropriateness of response options and recall periods, comprehensiveness, and comprehensibility.

A potential limitation of using Rasch analysis in the development of PROMs for assessing physical function is the risk that useful or clinically important items may be excluded if they do not meet strict statistical fit criteria (Tennant and Küçükdeveci, 2023). Automatically discarding items on statistical grounds alone could undermine the content validity of the instrument by narrowing its scope or omitting aspects of physical function that are highly relevant to patients and clinicians. Therefore, it is crucial to strike a balance between psychometric rigor and clinical judgment. This involves critically evaluating misfitting items in the context of their functional importance, consulting expert panels, and potentially retaining certain items with known clinical value despite misfitting values, provided they do not compromise the overall measurement properties of the scale. A thoughtful integration of Rasch statistics with expert clinical input can help ensure that the developed PROM remains both psychometrically valid and clinically meaningful.

4.3. Assumption of equal item discrimination

The Rasch measurement model specifies that all items share a common discrimination, such that the probability of endorsing an item is solely a function of the difference between a person's location on the latent trait and the item's difficulty (Tennant and Küçükdeveci, 2023). This contrasts with CTT, where item-total correlations are used descriptively to evaluate item discrimination, and with general IRT models that estimate item-specific discrimination parameters (Siegert et al., 2025). Although the Rasch model's equal-discrimination requirement is more restrictive, it is empirically testable: Violations of this requirement manifest as item misfit, prompting investigation of potential sources such as multidimensionality, local dependence, or DIF. When model assumptions are adequately met, the Rasch framework supports specific objectivity and justifies the transformation of ordinal raw scores into interval-level measures.

5. Conclusion

Rasch analysis can help researchers to develop robust, comprehensive and more precise PROMs to assess physical function. By providing information about the difficulty of the items, detailed insights into item functioning, and interval-level measurement, Rasch analysis addresses many of the limitations inherent in PROMs developed using CTT. However, its strict model assumptions and requirement for specialised expertise must be carefully considered, and alternative IRT models, such as two-parameter and three-parameter models, which allow item discrimination parameters to vary, may be more appropriate when Rasch assumptions are not met. Close collaboration among psychometricians, researchers, and clinicians is key to making Rasch analysis more interpretable and easier to apply within rehabilitation settings. Content validity should be established during initial item development and reassessed after Rasch analysis, as item modification or removal may alter the comprehensiveness and relevance of the construct representation.

CRediT authorship contribution statement

Andrés Pierobon: Writing – review & editing, Writing – original draft, Project administration, Investigation. **Chris Krägeloh:** Writing – review & editing, Visualization, Supervision, Conceptualization.

Funding sources

This research has been funded by the Health Research Council (HRC 21/826).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

No specific acknowledgements to declare.

Glossary

Classical Test Theory (CTT) A traditional measurement framework in which observed scores are assumed to consist of a true score plus measurement error

Content validity The extent to which a PROM adequately covers and represents all relevant aspects of the concept it is intended to assess

Structural validity The degree to which the scores of a PROM are an adequate reflection of the dimensionality of the construct to be measured

Construct validity The degree to which the scores of a PROM are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments or differences between relevant groups) based on the assumption that the PROM validly measures the construct to be measured

Reliability The degree to which the measurement is free from measurement error

Internal consistency The degree of interrelationship or homogeneity among the items on a test, such that they are consistent with one another and measuring the same construct

Responsiveness The ability of a PROM to detect meaningful change over time when it has occurred

Exploratory Factor Analysis (EFA) A data-driven technique used to identify the underlying factor structure of a set of items

without imposing a predefined model. Items may load on one or more factors, and the number of dimensions is determined empirically

Confirmatory Factor Analysis (CFA) A hypothesis-driven technique used to test whether data fit a prespecified factor structure. Model fit indices indicate how well the proposed structure represents the observed data

Item Response Theory (IRT) A family of probabilistic measurement models that describe the relationship between a person's level on a latent trait and their probability of responding to an item in a particular way

Rasch analysis A specific IRT model that requires data to fit strict measurement criteria. It assumes that the probability of a given response is determined solely by the difference between person ability and item difficulty, enabling invariant measurement and transformation of ordinal raw scores into interval-level measures

Item difficulty (or item location) In Rasch analysis, item difficulty represents the location of an item on the latent trait continuum. More difficult items require higher levels of the trait to be successfully endorsed or completed

Item discrimination In general IRT models, discrimination reflects how well an item differentiates between individuals at different levels of the trait. In the Rasch model, discrimination is fixed to be equal across items and is therefore not estimated separately

Person ability (or person location) An estimate of an individual's level on the latent trait (e.g. physical function), derived from their pattern of responses across items

Logit The unit of measurement used in Rasch analysis. Logits represent the natural logarithm of the odds of a successful response and form an interval-level scale on which both person ability and item difficulty are expressed.

Fit statistics (item and person fit) Indices that indicate how well observed responses align with model expectations. Misfit suggests that an item or respondent may not conform to the assumptions of the Rasch model

Item thresholds Points along the latent trait continuum at which the probability of endorsing one response category equals that of endorsing the adjacent category, relevant for polytomous items

Differential Item Functioning (DIF) DIF occurs when individuals from different groups (e.g. sex, age, condition) with the same underlying ability have different probabilities of endorsing an item, indicating potential bias

References

Andrich, D., 1988. Rasch Models for Measurement. Sage Publications.

Belvedere, S.L., de Morton, N.A., 2010. Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. J. Clin. Epidemiol. 63 (12), 1287–1297. https://doi.org/10.1016/j.jclinepi.2010.02.012.

Bishop, P.A., Herron, R.L., 2015. Use and misuse of the likert item responses and other ordinal measures. Int. J. Exerc. Sci. 8 (3), 297–302. https://doi.org/10.70252/lanz1453.

Black, N., 2013. Patient reported outcome measures could help transform healthcare. Bmj 346, f167. https://doi.org/10.1136/bmj.f167.

Blasco, J.M., Acosta-Ballester, Y., Igual-Camacho, C., Hernández-Guillén, D., Gómez, M. C., Roig-Casasús, S., Puigcerver-Aranda, P., 2020. Preferred outcome measures used in randomized clinical trials of total knee replacement rehabilitation: a systematic review. PM&R 12 (7), 706–713. https://doi.org/10.1002/pmrj.12312.

Boone, W.J., 2016. Rasch analysis for instrument development: why, when, and how? CBE-Life Sci. Educ. 15 (4). https://doi.org/10.1187/cbe.16-04-0148.

Brentari, E., Golia, S., 2007. Unidimensionality in the Rasch model: how to detect and interpret. Statistica 67, 253–261. https://doi.org/10.6092/issn.1973-2201/3508.

Chiarotto, A., 2019. Patient-reported outcome measures: best is the enemy of good (but what if good is not good enough?). J. Orthop. Sports Phys. Ther. 49 (2), 39–42. https://doi.org/10.2519/jospt.2019.0602.

Comins, J., Brodersen, J., Krogsgaard, M., Beyer, N., 2008. Rasch analysis of the Knee injury and Osteoarthritis Outcome Score (KOOS): a statistical re-evaluation. Scand. J. Med. Sci. Sports 18 (3), 336–345. https://doi.org/10.1111/j.1600-0838.2007.00724.x.

Cook, C.E., Wright, A., Wittstein, J., Barbero, M., Tousignant-Laflamme, Y., 2021. Five recommendations to address the limitations of patient-reported outcome measures. J. Orthop. Sports Phys. Ther. 51 (12), 562–565. https://doi.org/10.2519/jospt.2021.10836.

Davidson, M., 2009. Rasch analysis of 24-, 18- and 11-item versions of the Roland-Morris disability questionnaire. Qual. Life Res. 18 (4), 473–481. https://doi.org/10.1007/s11136-009-9456-4.

De Champlain, A.F., 2010. A primer on classical test theory and item response theory for assessments in medical education. Med. Educ. 44 (1), 109–117. https://doi.org/10.1111/j.1365-2923.2009.03425.x.

de Vet, H.C.W., Terwee, C.B., Mokkink, L.B., Knol, D.L., 2011. Measurement in Medicine: a Practical Guide. Cambridge University Press. https://doi.org/10.1017/CBO9780511996214.

Eckhard, L., Munir, S., Wood, D., Talbot, S., Brighton, R., Walter, B., Baré, J., 2021. The ceiling effects of patient reported outcome measures for total knee arthroplasty. Orthop. Traumatol. Surg. Res. 107 (3), 102758. https://doi.org/10.1016/j.otsr.2020.102758.

Engelhard Jr., G., Wang, J., 2025. Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences, second ed. Routledge.

Felt, J.M., Castaneda, R., Tiemensma, J., Depaoli, S., 2017. Using person fit statistics to detect outliers in survey research. Front. Psychol. 8, 863. https://doi.org/10.3389/fpsyg.2017.00863.

Franchignoni, F., Salaffi, F., Giordano, A., Carotti, M., Ciapetti, A., Ottonello, M., 2013. Rasch analysis of the 22 knee injury and osteoarthritis outcome score-physical function items in Italian patients with knee osteoarthritis. Arch. Phys. Med. Rehabil. 94 (3), 480–487. https://doi.org/10.1016/j.apmr.2012.09.028.

Froud, R., Patel, S., Rajendran, D., Bright, P., Bjørkli, T., Buchbinder, R., Eldridge, S., Underwood, M., 2016. A Systematic Review of Outcome Measures Use, Analytical Approaches, Reporting Methods, and Publication Volume by Year in Low Back Pain Trials Published between 1980 and 2012: respice, adspice, et prospice. PLoS One 11, e0164573. https://doi.org/10.1371/journal.pone.0164573.

Gabel, C.P., Michener, L.A., Burkett, B., Neller, A., 2006. The upper limb functional index: development and determination of reliability, validity, and responsiveness. J. Hand Ther. 19 (3), 328–348. https://doi.org/10.1197/j.jht.2006.04.001.

Hagquist, C., Andrich, D., 2017. Recent advances in analysis of differential item functioning in health research using the Rasch model. Health Qual. Life Outcome 15 (1), 181. https://doi.org/10.1186/s12955-017-0755-0.

Hsu, J.E., Nacke, E., Park, M.J., Sennett, B.J., Huffman, G.R., 2010. The disabilities of the arm, shoulder, and hand questionnaire in intercollegiate athletes: validity limited by ceiling effect. J. Shoulder Elb. Surg. 19 (3), 349–354. https://doi.org/10.1016/j.jse.2009.11.006.

Hysing-Dahl, T., Faleide, A.G.H., Waaler, P.A.S., Inderhaug, E., 2025. The ability of the knee osteoarthritis outcome score to detect changes over time is limited in patients with patellar instability due to substantial ceiling effect. J. Exp. Orthop. 12 (2), e70146. https://doi.org/10.1002/jeo2.70146.

Jo, Y.H., Lee, K.H., Jeong, S.Y., Kim, S.J., Lee, B.G., 2021. Shoulder outcome scoring systems have substantial ceiling effects 2 years after arthroscopic rotator cuff repair. Knee Surg. Sports Traumatol. Arthrosc. 29 (7), 2070–2076. https://doi.org/10.1007/s00167-020-06036-y.

Kent, P., Grotle, M., Dunn, K.M., Albert, H.B., Lauridsen, H.H., 2015. Rasch analysis of the 23-item version of the Roland Morris disability questionnaire. J. Rehabil. Med. 47 (4), 356–364. https://doi.org/10.2340/16501977-1935.

Lam, K.C., Marshall, A.N., Snyder Valier, A.R., 2020. Patient-reported outcome measures in sports medicine: a concise resource for Clinicians and researchers. J. Athl. Train. 55 (4), 390–408. https://doi.org/10.4085/1062-6050-171-19.

Leung, Y.-Y., Png, M.-E., Conaghan, P., Tennant, A., 2014. A systematic literature review on the application of Rasch analysis in musculoskeletal disease — a special interest group report of OMERACT 11. J. Rheumatol. 41 (1), 159. https://doi.org/10.3899/jrheum.130814.

Liddell, T.M., Kruschke, J.K., 2018. Analyzing ordinal data with metric models: what could possibly go wrong? J. Exp. Soc. Psychol. 79, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009.

Lin, I., Wiles, L., Waller, R., Goucke, R., Nagree, Y., Gibberd, M., Straker, L., Maher, C.G., O'Sullivan, P.P.B., 2020. What does best practice care for musculoskeletal pain look like? Eleven consistent recommendations from high-quality clinical practice guidelines: systematic review. Br. J. Sports Med. 54 (2), 79. https://doi.org/10.1136/bjsports-2018-099878.

Linacre, J., 2006. Rasch: too complicated or too simple? https://www.rasch.org/rmt/rmt203b.htm.

Mallinson, T., Kozlowski, A.J., Johnston, M.V., Weaver, J., Terhorst, L., Grampurohit, N., Juengst, S., Ehrlich-Jones, L., Heinemann, A.W., Melvin, J., Sood, P., Van de Winckel, A., 2022. Rasch Reporting Guideline for Rehabilitation Research (RULER): the RULER statement. Arch. Phys. Med. Rehabil. 103 (7), 1477–1486. https://doi.org/10.1016/j.apmr.2022.03.013.

Medvedev, O.N., Krägeloh, C.U., 2025. Rasch measurement model. In: Medvedev, O.N., Krägeloh, C.U., Siegert, R.J., Singh, N.N. (Eds.), Handbook of Assessment in Mindfulness Research. Springer, pp. 131–147. https://doi.org/10.1007/978-3-030-77644-2_4-1.

Medvedev, O.N., Turner-Stokes, L., Ashford, S., Siegert, R.J., 2018. Rasch analysis of the UK functional assessment measure in patients with complex disability after stroke. J. Rehabil. Med. 50 (5). https://doi.org/10.2340/16501977-2324, 428-428.

Page, M.J., McKenzie, J.E., Green, S.E., Beaton, D.E., Jain, N.B., Lenza, M., Verhagen, A. P., Surace, S., Deitch, J., Buchbinder, R., 2015. Core domain and outcome measurement sets for shoulder pain trials are needed: systematic review of physical

therapy trials. J. Clin. Epidemiol. 68 (11), 1270–1281. https://doi.org/10.1016/j.jclinepi.2015.06.006.
Perruccio, A.V., Stefan Lohmander, L., Canizares, M., Tennant, A., Hawker, G.A., Conaghan, P.G., Roos, E.M., Jordan, J.M., Maillefert, J.F., Dougados, M., Davis, A. M., 2008. The development of a short measure of physical function for knee OA KOOS-Physical Function Shortform (KOOS-PS) - an OARSI/OMERACT initiative. Osteoarthr. Cartil. 16 (5), 542–550. https://doi.org/10.1016/j.joca.2007.12.014.
Pierobon, A., Taylor, W., Caya, R., Villalba, F., Soliño, S., Policastro, P.O., Siegert, R., Darlow, B., 2025. Physical functions assessed by lower limb performance-based and self-reported outcome measures for knee musculoskeletal conditions: a scoping review. Braz. J. Phys. Ther. 29 (1), 101166. https://doi.org/10.1016/j.bjpt.2024.101166.
Ra, H.J., Kim, H.S., Choi, J.Y., Ha, J.K., Kim, J.Y., Kim, J.G., 2014. Comparison of the ceiling effect in the Lysholm score and the IKDC subjective score for assessing functional outcome after ACL reconstruction. Knee 21 (5), 906–910. https://doi.org/10.1016/j.knee.2014.06.004.
Reiter, C.R., Abraham, V.M., Riddle, D.L., Patel, N.K., Goldman, A.H., 2024. Patient reported outcome measures (PROMs) as primary and secondary outcomes in total hip and knee arthroplasty randomized controlled trials: a systematic review. Arch. Orthop. Trauma Surg. 144 (5), 2257–2266. https://doi.org/10.1007/s00402-024-05242-4.
Roland, M., Morris, R., 1983. A study of the natural history of back pain: part I: development of a reliable and sensitive measure of disability in low-back pain. Spine 8 (2), 141–144. https://doi.org/10.1097/00007632-198303000-00004.
Roos, E.M., Roos, H.P., Lohmander, L.S., Ekdahl, C., Beynnon, B.D., 1998. Knee Injury and Osteoarthritis Outcome Score (KOOS)–development of a self-administered outcome measure. J. Orthop. Sports Phys. Ther. 28 (2), 88–96. https://doi.org/10.2519/jospt.1998.28.2.88.
Sabet, T.S., Anderson, D.B., Stubbs, P.W., Buchbinder, R., Terwee, C.B., Chiarotto, A., Gagnier, J., Verhagen, A.P., 2025. Pain and physical function are common core domains across 40 core outcome sets of musculoskeletal conditions: a systematic review. J. Clin. Epidemiol. 180, 111687. https://doi.org/10.1016/j.jclinepi.2025.111687.
Saithna, A., Cote, M.P., 2024. Editorial commentary: ceiling effects are a limitation of frequently used patient-reported outcomes measures used to assess shoulder function: appropriate selection of shoulder patient-reported outcomes measures is required-especially in athletes. Arthroscopy 40 (3), 711–713. https://doi.org/10.1016/j.arthro.2023.11.001.
Siegert, R.J., Krägeloh, C.U., Medvedev, O.N., 2025. Classical test theory and the measurement of mindfulness. In: Medvedev, O.N., Krägeloh, C.U., Siegert, R.J., Singh, N.N. (Eds.), Handbook of Assessment in Mindfulness Research. Springer, Nature Switzerland, pp. 51–64. https://doi.org/10.1007/978-3-031-47219-0_3.
Soh, S.E., Harris, I.A., Cashman, K., Heath, E., Lorimer, M., Graves, S.E., Ackerman, I.N., 2021. Implications for research and clinical use from a Rasch analysis of the HOOS-12 and KOOS-12 instruments. Osteoarthr. Cartil. 29 (6), 824–833. https://doi.org/10.1016/j.joca.2021.02.568.
Sönning, L., 2024. Ordinal response scales: psychometric grounding for design and analysis. Res. Method Appl. Linguist. 3 (3), 100156. https://doi.org/10.1016/j.rmal.2024.100156.
Tennant, A., Conaghan, P.G., 2007. 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 Care Res. 57 (8), 1358–1362.
Tennant, A., Küçükdeveci, A.A., 2023. Application of the Rasch measurement model in rehabilitation research and practice: early developments, current practice, and future challenges. Front. Rehabil. Sci. 4, 1208670. https://doi.org/10.3389/fresc.2023.1208670.
Terwee, C.B., Prinsen, C.A.C., Chiarotto, A., Westerman, M.J., Patrick, D.L., Alonso, J., Bouter, L.M., de Vet, H.C.W., Mokkink, L.B., 2018. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual. Life Res. 27 (5), 1159–1170. https://doi.org/10.1007/s11136-018-1829-0.
Tesio, L., Caronni, A., Kumbhare, D., Scarano, S., 2024a. Interpreting results from Rasch analysis 1. The “most likely” measures coming from the model. Disabil. Rehabil. 46 (3), 591–603. https://doi.org/10.1080/09638288.2023.2169771.
Tesio, L., Caronni, A., Simone, A., Kumbhare, D., Scarano, S., 2024b. Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. Disabil. Rehabil. 46 (3), 604–617. https://doi.org/10.1080/09638288.2023.2169772.
Thomas, M.L., 2019. Advances in applications of item response theory to clinical assessment. Psychol. Assess. 31 (12), 1442–1455. https://doi.org/10.1037/pas0000597.