



This is a repository copy of *Patient reported outcome measures require scale metrification and quantified precision: evidence from the assessment of breathlessness in people with ALS/MND*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/229814/>

Version: Published Version

Article:

Young, C.A. orcid.org/0000-0003-1745-7720, Chaouch, A., Mcdermott, C.J. orcid.org/0000-0002-1269-9053 et al. (6 more authors) (2025) Patient reported outcome measures require scale metrification and quantified precision: evidence from the assessment of breathlessness in people with ALS/MND. Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration. ISSN 2167-8421

<https://doi.org/10.1080/21678421.2025.2533870>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

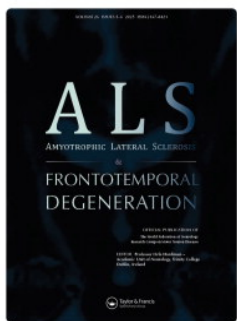
<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>



Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration

ISSN: 2167-8421 (Print) 2167-9223 (Online) Journal homepage: www.tandfonline.com/journals/iafd20

Patient reported outcome measures require scale metrification and quantified precision: evidence from the assessment of breathlessness in people with ALS/MND

Carolyn A. Young, Amina Chaouch, Christopher J. Mcdermott, Ammar Al-Chalabi, Suresh Kumar Chhetri, Nicola Waters, Richard Buccleuch, Roger J. Mills, Alan Tennant & On behalf of the TONiC-ALS study group

To cite this article: Carolyn A. Young, Amina Chaouch, Christopher J. Mcdermott, Ammar Al-Chalabi, Suresh Kumar Chhetri, Nicola Waters, Richard Buccleuch, Roger J. Mills, Alan Tennant & On behalf of the TONiC-ALS study group (20 Jul 2025): Patient reported outcome measures require scale metrification and quantified precision: evidence from the assessment of breathlessness in people with ALS/MND, Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, DOI: [10.1080/21678421.2025.2533870](https://doi.org/10.1080/21678421.2025.2533870)

To link to this article: <https://doi.org/10.1080/21678421.2025.2533870>



© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 20 Jul 2025.



Submit your article to this journal [↗](#)



Article views: 119






View related articles [↗](#)



View Crossmark data [↗](#)

RESEARCH ARTICLE

Patient reported outcome measures require scale metrification and quantified precision: evidence from the assessment of breathlessness in people with ALS/MND

CAROLYN A. YOUNG^{1,2} , AMINA CHAOUCH³, CHRISTOPHER J. MCDERMOTT⁴ , AMMAR AL-CHALABI⁵, SURESH KUMAR CHHETRI⁶ , NICOLA WATERS⁷, RICHARD BUCCLEUCH⁸, ROGER J. MILLS^{1,2}, ALAN TENNANT⁹ & On behalf of the TONiC-ALS study group

¹Department of Neurology, Walton Centre NHS Foundation Trust, Liverpool, UK, ²Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, UK, ³Department of Neurology, Greater Manchester Centre for Clinical Neurosciences, Salford, UK, ⁴Sheffield Institute for Translational Neuroscience, Sheffield, UK, ⁵Department of Basic and Clinical Neuroscience, Maurice Wohl Clinical Neuroscience Institute, King's College London, London, UK, ⁶Department of Neurology, Lancashire Teaching Hospital, Preston, UK, ⁷Lay author, UK MND Research Institute, London, UK, ⁸Lay author, UK, and ⁹Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, Leeds, UK

Abstract

Introduction: Precision (how closely repeated measures match) and responsiveness (ability to detect change over time) are critical properties of patient reported outcome measures (PROMs). Smallest Detectable Difference (SDD) is a useful statistic regarding precision; Minimal Detectable Change (MDC) and Minimal Important Change (MIC) assess responsiveness. **Methods:** We examined measurement properties of Numeric Rating Scale for Breathlessness, ALSFRS-R respiratory subscale and Dyspnea-12, contributed by participants in the Trajectories of Outcome in Neurological Conditions-ALS study. Rasch analysis converts ordinal scale data to interval equivalents. **Results:** Data from 1120 people with ALS showed ALSFRS-R Respiratory is only valid as ordinal data. The NRS Breathlessness requires computation from a wider NRS set for Rasch analysis; its SDD is 3.2, MDC 2.59, MIC 2.39, with score range of 0–10. The Dyspnea-12 has SDD 7.0, MDC 6.14, MIC 4.5, with score range of 0–36. The %MDC, indicating smallest change detectable above measurement error as % of scale range, is superior for the Dyspnea-12 (17.1%) compared to the NRS Breathlessness (25.9%). Another advantage of Dyspnea-12 is transformation of raw ordinal to interval equivalent data using published conversion tables. Both NRS and Dyspnea-12 have disadvantages of MIC < MDC. **Conclusions:** Accurate measurement underpins optimal clinical decision making and high-quality research. Informed choice of PROMs reduces risk of misinterpreting clinical and research data. Patients want PROMs which they feel give an accurate account of their progression when participating in research and communicating with their clinical team. The Dyspnea-12 is preferable for clinical and research use based on its psychometric properties.

Keywords: Rasch analysis, patient reported outcome measure, TONiC-ALS study, breathlessness, dyspnea

Introduction

Patient reported outcome measures (PROMs) are increasingly used in clinical practice and research, necessitating an understanding of terms such as validity (measures what is intended), reliability (does so consistently) and responsiveness (can

measure change). These are well documented in the “Consensus-based Standards for the selection of health Measurement INstruments” (COSMIN) literature (1). There are many guidelines for choosing PROMs, including for electronic implementation (2–4). However, there remains a critical yet neglected issue, which relates to the level of

Correspondence: Carolyn A. Young, Institute of Systems Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7BE, UK. E-mail: Cayoung@liverpool.ac.uk

(Received 9 January 2025; revised 3 July 2025; accepted 8 July 2025)

ISSN 2167-8421 print/ISSN 2167-9223 online © 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

DOI: 10.1080/21678421.2025.2533870

In considering any analysis it is crucial to understand the type of measurement available and, as a result, the type of statistical analysis that can be used.

Nominal: A nominal (or categorical) scale simply names categories to assign variables to, such as marital status or blood type. This is the lowest level of measurement because one can only group the observations, and it is not possible to order the groups.

Ordinal: Ordinal scales also name groups in the data, but these groups can be placed in a natural order; examples include Likert scales (strongly disagree to strongly agree). While it is possible to rank the values, the differences between values might not be consistent.

Interval: For interval scales, the order of values and the interval between any two points is meaningful. An example would be the person's temperature, the change in temperature from 32°C to 34°C is the same as the rise from 37°C to 39°C. But zero Celsius represents a temperature rather than a lack of temperature, so there is no true zero and it is not valid to employ ratios and say that 30°C is three times as hot as 10°C.

Ratio: With ratio scales, intervals are still meaningful and zero measurements represent a lack of the attribute, e.g. zero kilograms indicates absence of weight. Consequently, measurement ratios are valid for these scales: 120 kg is twice as heavy as 60 kg.

Figure 1. Classification of measurement.

measurement achieved by the PROM, outlined by Stevens (5) and described in Figure 1. This is important because it creates a significant risk of introducing bias into findings.

The key issue is that the total score, or subscale scores, from PROMs are ordinal scales, where results can be ranked in order, but where the distances between those ranks are not known (6). This rules out any calculations such as standard deviations and change scores including Minimal (Clinically) Important Change (MIC) (7,8). The interval, or metric, level of measurement is required for these calculations, i.e. where the interval between any two scale points is constant.

Increasingly, conversion tables are being produced which allow for a transformation of the ordinal raw score to the interval level metric (9), a process termed metrification. This can be achieved through the application of the Rasch measurement model (10), a process widely known as Rasch analysis. Rasch analysis has recently been extensively reviewed, including guidelines on how to go about the task (11). Data contributed when participants complete PROMs constitute raw data. The items from such data are typically added together to produce a score, which is at the ordinal level. This item set must be tested to assess if it is working together properly, i.e. its internal construct validity. One way of doing this is by using Rasch analysis to check if the PROM is measuring a single concept (unidimensionality) and whether a PROM repeatedly measuring the same stable concept yields similar results (reliability). In addition, Rasch analysis can identify items which are answered differently by different groups of people, even though they actually are the same for the

concept being measured (Differential Item Functioning (DIF) (12).

If the PROM was likened to weighing scales, it should measure how heavy you are (internal construct validity), weight (unidimensionality), produce the same result every time you step on them unless you have changed (reliability), and two people the same weight should see the same result irrespective of their age/gender (absence of DIF). The added advantage of Rasch analysis is, that if all these aspects are satisfactory, then the ordinal score can be converted to an interval level metric. Extending our weighing analogy, a result on the scales of "light/medium/heavy" can be converted to kilograms.

The measurement of dyspnea (breathlessness) is a useful exemplar of the need for PROMs; dyspnea is a "perception of an abnormal or distressing *internal* state", so it can only be measured by patient self-report (13). Thus, PROMs are invaluable as they convert patients' symptoms, such as "I feel breathless", to measurement, thereby enabling comparison with other subjects and monitoring in the same person across time.

Dyspnea has been shown to influence quality of life and risk of depression for people with amyotrophic lateral sclerosis (ALS), also known as motor neuron disease (MND) (14). Despite the impact of this symptom in ALS/MND, there is no consensus about how to measure it. Systematic reviews were unable to highlight any single scale of the many options for dyspnea in adults and considered the Visual Analogue Scale (VAS) or Numeric Rating Scale (NRS) suitable for further evaluation (15,16). The VAS typically asks the subject to mark a point denoting their symptom level on a 10-cm horizontal line labeled on left as "None" and right at "Maximum" of the symptom, typically scored as 0–100. The NRS asks the subject to rate their symptom 0–10 anchored to 0: "None" and 10: 'Maximum'. A systematic review found that NRS had better compliance and ease of use than the VAS¹⁷.

The ALS Functional Rating Scale-Revised (ALSFRS-R) incorporates questions about respiratory function (18). Earlier work has shown that the ALSFRS-R should be considered as a profile of mean scores from three different domains (bulbar, motor and respiratory) (19), raising the possibility of using the respiratory domain to assess dyspnea. Since the reviews cited above, a novel generic scale, the Dyspnea-12, has been recently validated for ALS/MND as a measure which quantifies breathlessness (20,21).

In this analysis, we examine some of the measurement properties of these three outcome measures: an NRS for breathlessness, ALSFRS-R respiratory subscale, and the Dyspnea-12, using data from the UK Trajectories of Outcomes in

Neurological Conditions-ALS (TONiC-ALS) study. Our objectives are to: (1) evaluate the published assertion that an NRS is suitable, since this would present a simple option for clinicians, attractive as it quickly provides a single number for clinical convenience; (2) assess if the respiratory subscale of the ALSFRS-R is suitable, again convenient due to its brevity and (3) answer whether there is a rationale to adopt the Dyspnea-12, despite it requiring patient self-report of 12 items.

In doing so, we present the analysis in a format which educates readers about desirable properties of measurement scales such as precision, and the pitfalls of not appreciating the limitations of raw scale scores. We present measurement concepts simply, as a refresher for some and as a primer for others, since we know that these topics are important to people living with ALS/MND who read the scientific literature.

Methods

Data collection

Patient recruitment. The Trajectories of Outcomes in Neurological Conditions-ALS (TONiC-ALS) collaboration recruited people with ALS/MND (pwALS) across many centers and asked them to complete a questionnaire pack containing a variety of PROMs at intervals of at least four months. Clinical data, including whether onset type was limb, bulbar, respiratory or unknown, were taken from the medical record. All participants received written information and written informed consent was obtained prior to enrollment into the study. Ethical approval was granted from research committees (#11/NW/0743).

Patient reported outcome measures. In the current paper three PROMs are considered:

1. *NRS Breathlessness* – a scale from 0 to 10, with 0 representing no breathlessness, and 10 representing extreme breathlessness. A single item scale, such as an NRS, cannot be applied to the Rasch model, so to obtain interval *NRS Breathlessness* data, breathlessness, fatigue, pain and spasticity NRS's were applied to the Rasch model as an impairment set, where each NRS spanned 0 for no symptom and 10 for extreme levels of that symptom. Each NRS was considered as an item in this set, and once fit to the model was achieved, the linearization of each NRS was translated back to a metric range of 0–10. Note that for single items such as the NRS, reliability is determined through a correlation based upon a test-retest, usually using repeated measures about 14 days apart (although could be longer), where respondents are unchanged.

2. *Amyotrophic Lateral Sclerosis Functional Rating Scale-Revised (ALSFRS-R) Respiratory* (18) – a validated patient-reported ALSFRS-R (22), from which this analysis used the respiratory subscale, consisting of 3 items from the original 12, covering breathlessness during activity or related to sleep, or use of ventilator support. Maximum score is 12, derived from 3 items each scored 0–4, where higher scores indicate fewer breathing problems.
3. *Dyspnea-12* – 12 items each scored 0–3 with total range 0–36, where a high score represents extreme breathlessness (20). The questionnaire was prefaced by a “skip” instruction such that if the patient perceived no breathing problems, they could omit the questionnaire and this would be scored as zero.

Precision. When dealing with cross-sectional PROM data, such as the differences between two patients receiving different care or treatment versus control groups in a clinical trial, it is important to understand the precision of the PROM. The following are two useful statistics:

- a. *Standard Error of Measurement (SEM)* is an indicator of how much a measured test score is spread around its “true” score and can be calculated as $SEM = SD \cdot \sqrt{(1 - \text{reliability})}$ (23–25). Consequently, the SEM is based on the Standard Deviation (SD) of the scale, a statistic measuring the dispersion of a dataset relative to its mean; since a mean score is required, data to calculate the SD must be interval. Similarly, the SEM calculation requires interval data which can be derived from Rasch analysis.
- b. *Smallest Detectable Difference (SDD)*: This is defined as the smallest difference that can be detected beyond measurement error. It represents the minimum difference that must be observed to be sure that the observed difference is real. The SDD is calculated as $SDD = \pm 1.96 \cdot \sqrt{2} \cdot SEM$, again calculated using interval data (26). A difference greater than the SDD represents a true difference.

Responsiveness. Responsiveness refers to the extent to which a scale can measure change when change has occurred. There are many ways to measure responsiveness (27). For longitudinal data, the *Standard Error of Measurement of the change (SEM_c)* can be calculated on the difference between baseline and first follow-up as $[SD \text{ of difference} \cdot \sqrt{(1 - \text{reliability})}]$ (28). The *Minimal Detectable Change (MDC)* is defined as the change below which there is more than a 95% chance that no real change has occurred and can be calculated as $\pm 1.96 \cdot \sqrt{2} \cdot SEM_c$ (25). Note that once again the calculation of change requires interval data.

The *Minimal (Clinically) Important Change (MIC)* reflects the smallest measured change in score that respondents perceive as important (29,30). The MIC is calculated in different ways in the literature. In one example the median change score of those worsening established the MIC⁸. We included a question at each follow-up as to whether dyspnea had worsened, stayed the same, or improved, allowing us to calculate the MIC, using the “anchor-based” method (31).

Results

Patients

This analysis uses data from 1120 pwALS. Mean age was 65.1 years (SD 10.6), 60.4% were male, and median duration since diagnosis was 9 months (IQR 3.5–24.1). Onset type was limb (71.1%), bulbar (26.9%), and respiratory (2.0%).

Rasch analysis

Statistical fit to the Rasch model is shown in Table 1 with ideal fit values at the foot of the table. The *Dyspnea-12* fit the model with a bi-factor equivalent solution which retained 90% of the explained common variance (ECV) (21). Raw ordinal *Dyspnea-12* data can be easily converted to interval level scores ranging from 0 to 36 using a conversion table (21). The combined NRS impairment set was also found to fit the model, enabling interval conversion of the *NRS Breathlessness*. Fit of the *ALSFRS-R Respiratory* was unsatisfactory (Table 1: Analyses 1–4). Thus, the *ALSFRS-R Respiratory* can only be analyzed as ordinal data.

Comparative analysis

Almost half (47.6%) of the cohort of 1120 pwALS had some level of breathlessness on the *Dyspnea-12*. On the *Dyspnea-12* interval level metric, where a skip gave a value of zero, the mean was 5.1 (SD 7.3). There was a significant difference by onset type, with respiratory onset highest at 16.3, compared to bulbar 5.2 and limb 4.5 (ANOVA, $F_{30.4}$ (df2, 1069); $p \leq 0.001$). The Spearman

correlation of the *Dyspnea-12* metric with the *ALSFRS-R Respiratory* was 0.77.

A strong gradient for the *Dyspnea-12* metric was shown across the *NRS Breathlessness* ordinal categories, ranging from a mean of 0.182 in category zero (range 0–9), to 20.7 in category 10 (range 11–36) ($F_{153.1}$ df (10, 1080); $p \leq 0.001$). Having converted *NRS Breathlessness* to interval level, there was also a significant difference in the *NRS Breathlessness* across onset type where the mean level was 5.41 (SD 3.33) for respiratory onset compared to 1.61 (SD 1.77) for limb and 2.02 (SD 1.88) in bulbar ($F_{48.0}$ (df 2,1058); $p < 0.001$).

The SEM and SDD for each scale are shown in Table 2, with the SEM and SDD for *ALSFRS-R Respiratory* calculated (inappropriately) on the ordinal scores solely for comparison purposes. A striking finding was the magnitude of SDD for the *NRS Breathlessness*, requiring movement of almost a third, 32.0% of the scale, to overcome the error.

The measurement of precision applies in the same way to the measurement of change. Again the *ALSFRS-R Respiratory* scale is included for reference only (Table 3). What is of interest is the gradient of the %MDC, i.e. the percentage of the scale that must be traversed to overcome the error, ranging from 17.1% for the *Dyspnea-12* to 34.1% for the *ALSFRS-R Respiratory*. In each instance the MDC values are higher than their respective MIC, suggesting that meaningful change could occur within the range of measurement error; in this circumstance the MDC provides the level of change important for patient.

Table 2. Standard error of measurement (SEM) and smallest detectable difference (SDD) for scales, showing SDD as percentage of scale range.

Scale	Range	SEM	SDD	%SDD
Dyspnea-12	0–36	2.5	7.0	19.4
NRS-breathlessness	0–10	1.1	3.2	32.0
ALSFRS-R respiratory	0–12	1.7	4.7	39.2

NRS: Numeric Rating Scale; ALSFRS-R: Amyotrophic Lateral Sclerosis Functional Rating Scale-Revised.

Table 1. Fit of scales to Rasch model.

PROMs	Fit residuals		Chi-Square Interaction Fit			Reliability		Unidimensionality % t values (LCI)	DIF	ECV
	Items	Persons	Total	df	p	PSI	α			
Dyspnea-12										
Item Set LD clusters	1.763	0.858	46.9	27	0.010	0.72	0.87	4.1	None	0.90
ALSFRS-R respiratory	6.949	0.525	19.5	4	<0.001	0.36	0.58	1.4	None	–
NRS										
Impairments	0.891	1.020	41.0	24	0.017	0.66	0.67	0.9	None	–
Ideal values	1	1			>0.05*	>0.7	>0.7	<5.0 (LCI <5.0)	–	–

PROMs: patient reported outcome measures; df: degrees of freedom; PSI: Person Separation Index; α : Cronbach’s alpha; LCI: Latent Correlation Index; DIF: Differential Item Functioning; ECV: Expected Common Variance; LD: Local Dependency; ALSFRS-R: Amyotrophic Lateral Sclerosis Functional Rating Scale-Revised; NRS: Numeric Rating Scale; *: Bonferroni adjusted value 0.007.

Table 3. Standard error of measurement of change (SEMc), minimal detectable change (MDC) and minimal important change (MIC) for scales.

Scale	Range	SEMc	MDC	% MDC	MIC	% MIC
Dyspnea-12	0–36	2.21	6.14	17.1	4.5	12.5
NRS-breathlessness	0–10	0.93	2.59	25.9	2.39	23.9
ALSFRS-R respiratory	0–12	1.48	4.09	34.1	2.39	19.9

NRS: Numeric Rating Scale; ALSFRS-R: Amyotrophic Lateral Sclerosis Functional Rating Scale-Revised – calculations only for comparison but not valid as only ordinal.

Discussion

This study demonstrates some of the factors and pitfalls to consider when selecting PROMs for clinical or research use. The literature suggests that an NRS might be suitable (15,16), and certainly during the SARS-COV2 pandemic, in an effort to adhere to the regular respiratory monitoring mandated for ALS care (32), an NRS might have seemed an attractive option for remote screening of respiratory symptoms by patient self-report; asking on a telephone or video review “How bad is your breathlessness on a scale of 0–10 where 0 is no breathlessness and 10 is severe breathlessness?”. However, it is important to remember that the ordered numbers from 0 to 10 do not indicate that NRS data is interval-scaled. Use of numbers to represent the different levels may suggest that it is reasonable to calculate change scores, such as assuming that a person who goes from NRS 5 to 7 at one review and then from 7 to 9 at the next review is deteriorating at a steady rate but this is not true. Analyses of VAS, which also yield results spanning 0–100 frequently inappropriately treated as interval level, show that the differences between values vary across these scales (33). The nature of ordinal scales, which are bounded by their minimum and maximum scores, translates into a sigmoidal curve when contrasted against the metric, so that respondents move more readily in the mid-range but it requires a more marked change in their state to make them change their responses at the extremes of the scale. As such, any use of change scores or means with raw NRS or VAS data is invalid.

The *NRS Breathlessness* can generate interval level data following fit of a set of similar symptom NRS’s to the Rasch model; however, in a real-world setting clinicians may not have data from other NRS’s and access to Rasch analysis. The *Dyspnea-12* fit the Rasch model and can be easily converted to interval level metrics for calculation of means and change scores (21). We suggest the evidence indicates *Dyspnea-12* is preferable for clinical and research use out of the three options examined; the MDC exceeding the MIC so that

not all meaningful change is detectable is a weakness of all three options. The *ALSFRS-R Respiratory* did not fit the Rasch model, providing further evidence that it requires revision (34). For those using the *ALSFRS-R* total score, or limb or bulbar subscales, a conversion table to convert to interval level data is available (9). The methods of our study do not permit any assessment of the correlation between dyspnea measured by PROMs and measurements of ventilatory strength, such as vital capacity.

The study also highlights the importance of the SDD for a PROM. The measures studied varied markedly in their SDD, e.g. *NRS Breathlessness* requires movement of 32.0% of its operational range before error is overcome. In contrast the *Dyspnea-12* requires a movement of 19.4% of its range to overcome error. In clinical practice, many changes in the *NRS Breathlessness* should be disregarded due to falling within measurement error. For example, it might be thought that two patients with *NRS Breathlessness* levels of 4 and 7 have different degrees of breathlessness because of the different scores but in fact, their breathlessness may be similar. In research, reading that the *NRS Breathlessness* levels of the treated group changed by 15% and the controls by 45% might suggest the treatment made a difference but with the high SDD of the *NRS Breathlessness*, these groups may also be similar.

Clinicians and researchers must be aware of the risks of incorrectly employing inappropriate statistics on PROMs, such as means for raw PROMs scores. In addition, pwALS need to be fully informed: a longer PROM may be necessary to achieve accurate measurement. Furthermore, large data sets are essential for Rasch analysis, demonstrating the importance of extensive data collection studies like TONiC-ALS <https://www.finders-study.org/mnd-study> and Precision ALS <https://www.precisionals.ie/>. Without these fundamental analyses to enable meticulous measurement, accurate high-quality research to investigate new treatments cannot be done.

The measurement properties of these three PROMs for dyspnea reveal they have very different utility for research or clinical care. If a PROM is not capable of conversion to interval level data, e.g. *ALSFRS-R-Respiratory*, it can only be used as ordinal data employing non-parametric analyses. It would be possible to detect if a patient had deteriorated or improved, but not by how much. Any research studies using inappropriate parametric analyses for ordinal data require re-analysis. In clinical care, simple change scores cannot be used without prior transformation. For pwALS, using a PROM which they feel gives an accurate account of their progression is an important part of

participating in research and communicating with their clinical team.

In clinical or research practice, presentation of data regarding precision and responsiveness of PROMs should ideally include:

- Analysis of any ordinal PROM data using appropriate non-parametric statistical techniques;
- The basis for conversion if ordinal PROM data is converted to interval;
- If interval conversion is possible, SEM and SDD, and for longitudinal research, the SEMc, MDC and MIC, all calculated using interval data.

In conclusion, attention to the measurement properties of PROMs is vital to avoid misinterpretation of clinical and research data. Earlier work suggested use of the *NRS Breathlessness* (15,16), but in ALS/MND, differences between two groups, such as treatment and control, would need to be more than 32.0% to be able to overcome measurement error, whereas for the *Dyspnea-12* it would be 19.4%. Those who use PROMs, or who read treatment studies using PROMs, benefit from understanding their measurement properties. Valuable large-scale studies to improve the utility of these measures reflect the contribution of pwALS, the clinical and research teams, and funders and sponsors.

Acknowledgments

We sincerely thank all the people with ALS/MND and their families who contributed to this study; staff from MND Care Centres in Basildon, Brighton, Cambridge, Cumbria, Dartford & Gravesham NHS Trust, Edinburgh, Exeter, Kings' College London, Leicester, Liverpool, London North West, London, North East, Maidstone, Manchester, Newcastle, North Devon, Norwich, Oxford, Peterborough, Portsmouth, Plymouth, Preston, Sheffield, Shropshire, Southampton, South Wales MND Care Network (Cardiff, Swansea, Cwm Taf and Hywel Dda Clinics), Stoke on Trent, Swansea, West Suffolk, Worcester for identifying and caring for study patients; the research and clinical staff for recruitment and data collection; and the TONiC team. We are particularly grateful for the insightful feedback on this manuscript from Dr Jane Haley of MND Scotland and Dr Brian Dickie of Motor Neurone Disease Association. We thank NIHR Clinical Research Network and the NIHR UK MND Clinical Studies Group for support.

Disclosures of interest

The authors report no conflicts of interest. No funding source were involved in the study design,

analysis, interpretation of data, or preparation of the manuscript.

Funding

This work was supported by the Motor Neurone Disease Association (UK) under grant Young/Jan15/929-794, and also received research support from the NIHR Clinical Research Network, and the Neurological Disability Fund 4530. AAC is an NIHR Senior Investigator and is supported through the following funding organizations under the egis of JPND – www.jpnd.eu (United Kingdom, Medical Research Council (MR/L501529/1; MR/R024804/1) and Economic and Social Research Council (ES/L008238/1)) and through the Motor Neurone Disease Association. This study represents independent research part-funded by the National Institute for Health Research (NIHR) Biomedical Research Center at South London and Maudsley NHS Foundation Trust and King's College London. CJM and this research are supported by the NIHR Sheffield Biomedical Research Center, an NIHR Research Professor Award and the NIHR Sheffield Clinical Research Facility. CAY and this research are supported by NIHR CRN NWC.

ORCID

Carolyn A. Young  <http://orcid.org/0000-0003-1745-7720>

Christopher J. Mcdermott  <http://orcid.org/0000-0002-1269-9053>

Suresh Kumar Chhetri  <http://orcid.org/0000-0002-3325-5582>

Data availability statement

Data supporting this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request. Data are located in controlled access data storage at Walton Center NHS Trust.

References

1. Mokkink LB, Prinsen CA, Bouter LM, Vet HC, Terwee CB. The CONsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. *Braz J Phys Ther.* 2016;20:105–13.
2. Al Sayah F, Jin X, Johnson JA. Selection of patient-reported outcome measures (PROMs) for use in health systems. *J Patient Rep Outcomes.* 2021;5:99.
3. Fontaine G, Poitras ME, Sasseville M, Pomey MP, Ouellet J, Brahim LO, et al. Barriers and enablers to the implementation of patient-reported outcome and experience measures (PROMs/PREMs): protocol for an umbrella review. *Syst Rev.* 2024;13:96.
4. Mowlem FD, Elash CA, Dumais KM, Haenel E, O'Donohoe P, Olt J, et al. Best practices for the electronic

- implementation and migration of patient-reported outcome measures. *Value Health*. 2024;27:79–94.
5. Stevens SS. On the theory of scales of measurement. *Science*. 1946;103:677–80.
 6. Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. *Arch Phys Med Rehabil*. 1989;70:308–12.
 7. Forrest M, Andersen B. Ordinal scale and statistics in medical research. *Br Med J (Clin Res Ed)*. 1986;292:537–8.
 8. Ørnbjerg LM, Christensen KB, Tennant A, Hetland ML. Validation and assessment of minimally clinically important difference of the unadjusted Health Assessment Questionnaire in a Danish cohort: uncovering ordinal bias. *Scand J Rheumatol*. 2020;49:1–7.
 9. Young CA, Chaouch A, McDermott CJ, Al-Chalabi A, Chhetri SK, Talbot K, et al. Improving the measurement properties of the Amyotrophic Lateral Sclerosis Functional Rating Scale-Revised (ALSFRS-R): deriving a valid measurement total for the calculation of change. *Amyotroph Lateral Scler Frontotemporal Degener*. 2024;25:400–9.
 10. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1960.
 11. Tennant A, Küçükdeveci AA. Application of the Rasch measurement model in rehabilitation research and practice: early developments, current practice, and future challenges. *Front Rehabil Sci*. 2023;4:1208670.
 12. Holland PW, Wainer H. Differential item functioning. Hillsdale, NJ: Psychology Press; 1993.
 13. Parshall MB, Schwartzstein RM, Adams L, Banzett RB, Manning HL, Bourbeau J, et al. An official American Thoracic Society statement: update on the mechanisms, assessment, and management of dyspnea. *Am J Respir Crit Care Med*. 2012;185:435–52.
 14. Young C, Ealing J, McDermott C, Williams T, Al-Chalabi A, Majeed T, et al. Fatigue and anxiety mediate the effect of dyspnea on quality of life in amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Frontotemporal Degener*. 2022;23:390–8.
 15. Dorman S, Byrne A, Edwards A. Which measurement scales should we use to measure breathlessness in palliative care? A systematic review. *Palliat Med*. 2007;21:177–91.
 16. Bausewein C, Farquhar M, Booth S, Gysels M, Higginson I. Measurement of breathlessness in advanced disease: a systematic review. *Respir Med*. 2007;101:399–410.
 17. Hjermstad M, Fayers P, Haugen D, Caraceni A, Hanks G, Loge J, et al. Studies comparing Numerical Rating Scales, Verbal Rating Scales, and Visual Analogue Scales for assessment of pain intensity in adults: a systematic literature review. *J Pain Symptom Manage*. 2011;41:1073–93.
 18. Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *BDNF ALS Study Group (Phase III)*. *J Neurol Sci*. 1999;169:13–21.
 19. Franchignoni F, Mora G, Giordano A, Volanti P, Chiò A. Evidence of multidimensionality in the ALSFRS-R Scale: a critical appraisal on its measurement properties using Rasch analysis. *J Neurol Neurosurg Psychiatry*. 2013;84:1340–5.
 20. Yorke J, Moosavi SH, Shuldham C, Jones PW. Quantification of dyspnoea using descriptors: development and initial testing of the Dyspnoea-12. *Thorax* 2010;65:21–6.
 21. Young CA, Chaouch A, McDermott CJ, Al-Chalabi A, Chhetri SK, Talbot K, et al. Dyspnea (breathlessness) in amyotrophic lateral sclerosis/motor neuron disease: prevalence, progression, severity, and correlates. *Amyotroph Lateral Scler Frontotemporal Degener*. 2024;25:475–85.
 22. Montes J, Levy G, Albert S, Kaufmann P, Buchsbaum R, Gordon PH, et al. Development and evaluation of a self-administered version of the ALSFRS-R. *Neurology* 2006;67:1294–6.
 23. Harvill LM. Standard error of measurement: an NCME instructional module on. *Educ Measur*. 1991;10:33–41.
 24. Altman DG, Bland JM. Standard deviations and standard errors. *BMJ*. 2005;331:903.
 25. Ries JD, Echternach JL, Nof L, Gagnon Blodgett M. Test-retest reliability and minimal detectable change scores for the timed “up & go” test, the six-minute walk test, and gait speed in people with Alzheimer disease. *Phys Ther*. 2009;89:569–79.
 26. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge: Cambridge University Press; 2011.
 27. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol*. 2000;53:459–68.
 28. Young CA, Rog DJ, Tanasescu R, Kalra S, Langdon D, Tennant A, et al. Multiple Sclerosis Vision Questionnaire (MSVQ-7): reliability, validity, precision and discrimination. *Mult Scler Relat Disord*. 2023;80:105115.
 29. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989;10:407–15.
 30. Kovacs FM, Abaira V, Royuela A, Corcoll J, Alegre L, Tomás M, et al. Minimum detectable and minimal clinically important changes for pain in patients with nonspecific neck pain. *BMC Musculoskelet Disord*. 2008;9:43.
 31. Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N, et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. *BMJ*. 2020;369:m1714.
 32. NICE. Motor neurone disease: assessment and management. In: Centre NCG, ed. London: National Institute for Health and Care Excellence; 2016.
 33. Kersten P, White PJ, Tennant A. Is the pain visual analogue scale linear and responsive to change? An exploration using Rasch analysis. *PLoS One*. 2014;9:e99485.
 34. Pinto S, de Carvalho M. The R of ALSFRS-R: does it really mirror functional respiratory involvement in amyotrophic lateral sclerosis? *Amyotroph Lateral Scler Frontotemporal Degener*. 2015;16:120–3.