Journal of Clinical Epidemiology 163 (2023) 21−28

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Psychometric sensitivity analyses can identify bias related to measurement properties in trials that use patient-reported outcome measures: a secondary analysis of a clinical trial using the disabilities of the arm, shoulder, and hand questionnaire

Conrad J. Harrison[a,*], Anower Hossain[b], Julie Bruce[b,c], Jeremy N. Rodrigues[b]

[a]*Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Surgical Intervention Trials Unit, University of Oxford, Oxford, UK*
[b]*Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry, UK*
[c]*University Hospital Coventry and Warwickshire NHS Trust, Clifford Bridge Road, Coventry CV2 2DX, UK*

Accepted 21 September 2023; Published online 27 September 2023

Abstract

Objectives: Demonstrate psychometric sensitivity analyses for testing the stability of study findings to assumptions made about patient-reported outcome measures.

Study Design and Setting: We performed secondary analyses of Disability of Arm, Shoulder, and Hand (DASH) data collected within the Prevention of Shoulder Problems clinical trial, which compared upper limb function scores in women who had undergone breast cancer surgery, randomized to either an exercise program or usual care. We repeated the principal trial analyses after grouping DASH items into subscales suggested by factorial analyses in this dataset and applied item response theory to account for unequal item weighting. We checked for measurement invariance by participant age and response shift bias using established techniques.

Results: Our analyses suggested that the DASH measured two constructs: motor function and sensory symptoms. The majority of the six-month difference in DASH score was driven by motor function. With item response theory scoring, we found differences in both constructs at 12 months ($P = 0.019$ and $P = 0.007$), but in neither construct at 6 months, contrary to the original trial results. We found no differential item function by age or between baseline and 12-month measurements.

Conclusions: Psychometric sensitivity analyses aid in the interpretation of the Prevention of Shoulder Problems trial's results.

Keywords: Psychometric sensitivity analysis; DASH; Item response theory; Randomized controlled trial; Psychometrics; Factor analysis

1. Introduction

1.1. Measurement error in patient-reported outcome measures

Increasingly, patient-reported outcome measures (PROMs) are being used in randomized controlled trials (RCTs) that have potential to directly influence healthcare policy [1−3]. The scores of a PROM are intended to reflect underlying health constructs that are important to patients but cannot be observed directly, such as pain, depression, and fatigue. Unlike a ruler, which directly measures length, PROMs *indirectly* measure these health constructs. For example, pain causes a person to respond to a PROM in a certain way, but the construct of interest (pain level in the mind of the patient) is not necessarily defined by the items (questions) in the PROM. Scores from a PROM might correlate closely with the underlying, unobservable, construct of interest, but they might also be affected by other human characteristics and so should be interpreted

Funding: Conrad J. Harrison is funded by a National Institute for Health Research (NIHR) Doctoral Research Fellowship (NIHR300684). Jeremy N. Rodrigues is funded by an NIHR postdoctoral fellowship (PDF-2017-10-075). This document presents independent research funded by the NIHR (HTA program project 13/84/10). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

* Corresponding author. The Botnar Research Centre, University of Oxford, Old Road, Oxford OX3 7LD, UK. Tel.: +01865 227374; fax: +01865 750 750.

E-mail address: Conrad.harrison@medsci.ox.ac.uk (C.J. Harrison).

https://doi.org/10.1016/j.jclinepi.2023.09.008

What is new?

Key findings

- Psychometric sensitivity analyses demonstrated how measurement assumptions impacted the results of the Prevention of Shoulder Problems trial.

What this adds to what was known?

- Psychometric sensitivity analyses test the stability of patient-reported outcome measure scores to key measurement assumptions.
- This may aid in the interpretation of findings from randomized controlled trials.

What is the implication and what should change now?

- Further work is warranted to assess the potential impact of psychometric sensitivity analyses, and define their role in clinical trials.

judiciously. The developers of PROMs endeavor to create instruments that reflect meaningful health constructs as accurately, reliably, and precisely as possible, but because these measurements are indirect, they will always incur some level of error. By error, we mean that there is discordance between the true level of the health construct and the level implied by the PROM score, for example, if a patient's pain score implies a higher or lower level of pain than the patient is actually experiencing. In some cases, this may be negligible, but due to the great variability in the construct validity of available PROMs, there may be cases where this error has the potential to influence the results and interpretation of findings from clinical trials.

For practical reasons, researchers make assumptions about PROM measurements. This paper firstly focuses on three key assumptions trialists may make about PROM scores: unidimensionality, correct (and often equal) weighting of items, and measurement invariance. We describe each of these before presenting a case study using PROM data from a recently published National Institute for Health Research−funded clinical trial.

1.2. Scale unidimensionality

A unidimensional PROM is one which either measures a single health construct or health constructs that are so closely related that they can be assumed to have an equal impact on the responses to items within the questionnaire [4]. It is undesirable to reflect experientially unrelated health constructs in a single score. For example, if we were to combine pain and stiffness into a single disability score, we may find that in some cases, a patient's pain improves, although their stiffness worsens—two important changes that might not be captured by a single PROM score if these changes were to offset each other. This example is further complicated if we give equal weight to pain and stiffness, when these constructs might not be equally important to the patient.

1.3. Item weighting

In a series of items measuring upper limb function, some items may more closely reflect upper limb function than others and some may reflect a greater level of upper limb function than others. For example, the ability to turn a key may reflect upper limb function more closely than one's ability to manage their transport needs, and doing heavy household chores might reflect a higher level of function than playing cards. Where these differences between items exist, their impact on the overall function score should differ. The degree to which the score of an individual item should affect the total score of the scale is multifactorial and complex [4], but for simplicity, we will refer to this general concept as item weighting.

1.4. Measurement invariance

Third, measurement invariance describes a stable relationship between the level of the latent construct and the responses to items in a PROM [4]. There are two types of bias which might be caused by a lack of measurement invariance and could go on to affect the findings of clinical trials and observational studies. The first is response shift bias and the second is differential item functioning by subgroups such as genders, ages, or ethnicities, where relevant.

Response shift bias occurs when the meaning of a PROM response changes over time. This can take three forms: recalibration, reprioritization, and reconceptualization [5]. Recalibration describes a change in the patient's internal frame of reference, for example, what is considered the 'worst pain imaginable' might change following an extremely painful experience. Reprioritization is a change in what the patient finds important, for example, a patient's ability to work may have a lesser impact on their perceived quality of life after retirement. Reconceptualization is a redefinition of the health construct in the mind of the patient, for example, what a patient considers to be independent functioning might change to include the use of compression garments, following lymph node surgery for breast cancer [6].

Differential item functioning by subgroup occurs when population subgroups respond to an item differently, for any given latent construct level. In an assessment of upper limb function, we may ask whether a patient's upper limb interferes with their work. For any given level of upper limb function, patients' answers might vary depending on the nature of their work. In this case, the item will have different measurement properties depending on the patient's job

description and is said to exhibit differential item functioning [7].

1.5. Psychometric sensitivity analyses

Individually, the potential for each of these theoretical assumptions to impact the results of clinical trials has been demonstrated, with both simulated and real data [8−11]. In other fields, such as health economics, assumptions which may go on to impact a trial's results are tested in secondary, post-hoc, sensitivity analyses. Psychometric sensitivity analyses are intended to explore the impact of measurement assumptions in trials that use PROMs [12]. They ask *what measurement properties did the PROM show in the trial dataset* and *how much would the trial results change if the PROM was not functioning as expected*? These are particularly important questions to ask when the PROM has not been developed with contemporary psychometric techniques.

This paper demonstrates the use of psychometric sensitivity analyses in a secondary analysis of patient-reported disability data from a recently completed National Institute for Health Research−funded multicenter rehabilitation RCT. The aim of these analyses is to test the stability of the trial's principal findings against these theoretical psychometric assumptions and to illustrate the potential impact of these analyses more generally.

2. Methods

2.1. The UK PROSPER trial (case study)

The Prevention of Shoulder Problems (PROSPER) trial was a pragmatic, multicentered RCT which compared a physiotherapy-led exercise program to usual National Health Service care in 392 women who had undergone breast cancer surgery between January 2016 and July 2017 [1]. The primary outcome measure was the Disabilities of the Arm, Shoulder, and Hand (DASH) questionnaire score, 12 months after randomization [13]. The trial was powered to detect a seven-point difference, assuming a minimum important difference (MID) of 5−10 points. Upper limb disability (DASH) data were collected by post, preoperatively, at six and 12 months. At 12 months, the adjusted mean difference in DASH sum-score was 7.81 favoring the exercise intervention (95% confidence interval [CI] 3.17 to 12.44, $P = 0.001$) after adjusting for age, baseline DASH score, type of breast surgery, type of axillary surgery, and the use of radiotherapy and chemotherapy, using ordinary least squares regression. The authors concluded that the exercise program was more clinically and cost-effective than standard care and reduced upper limb disability one year after breast cancer surgery.

2.2. The disabilities of the arm, shoulder, and hand questionnaire

The DASH questionnaire is one of the most frequently used PROMs for measuring upper limb disability. It contains 30 items, each with five response categories scored 1, 2, 3, 4, and 5, with a higher score indicating more severe disability. Overall scores are typically calculated by subtracting 1 from the mean item score and multiplying by 25 to generate a 0−100 index. Scores from response sets that are missing more than three item responses are not usually included in primary analyses [14].

The DASH was developed in 1996 by a North American collaborative group, although not using modern psychometric techniques, such as Rasch measurement theory [15] or IRT [4] that ensure unidimensionality and interval-scale scoring. Since its release, several psychometric studies have challenged the structural validity of the DASH and its short-form (11-item) version, the QuickDASH, suggesting that the instruments may be better divided into at least two subscales [16−19]. The PROSPER study did perform a secondary analysis involving the DASH subscales (activity limitations; impairment; participation restriction) proposed by Dixon et al., [20] but the grouping of items into these particular subscales was based on academics' interpretation of the International Classification of Functioning Disability and Health and does not necessarily reflect how the items are interpreted by women undergoing breast cancer surgery nor covariance patterns in their item responses.

2.3. Ethical approval

We obtained ethical approval from the University of Warwick Biomedical and Scientific Research Ethics Committee (BSREC16/22-23, 26/10/2022) to perform these secondary psychometric analyses using anonymized data from the PROSPER trial.

2.4. Stage 1: factorial structure

In our first psychometric sensitivity analysis, we grouped DASH items into two unidimensional subscales, scored these subscales by summing the individual item responses, and repeated the primary PROSPER analysis with these subscale scores. To understand how many subscales items should be grouped into, and which items belong to which subscale, we performed a series of psychometric tests, including a scree plot and Kaiser criterion analysis, exploratory and confirmatory factor analyses, and bifactor modeling (Supplementary material). These techniques aim to reduce the number of dimensions of the item response covariance matrix into subscales that account for the majority of covariance between items [21]. This prevents items that do not correlate well with each other from being combined to produce a single score.

2.5. Stage 2: item response theory modeling

Item response theory (IRT) describes a framework for using probabilistic models to describe the relationship between a set of item responses and the level of the latent construct which they aim to measure [4]. When an IRT model can be fitted to a set of PROM responses, it can provide interval scale measurement (as opposed to ordinal measurement), missing data can be handled directly by the model, and it is potentially possible to derive more granular measurements, as these models account for the *pattern* of item responses (rather than merely the sum of item responses). IRT is believed to handle the naturally unequal weighting of items in the minds of patients by allowing certain items to impact the scale score more than others as the level of the latent construct varies.

In a second psychometric sensitivity analysis, we fitted IRT models (specifically, graded response models [22]) to the unidimensional subscales and used these to produce subscale IRT scores (specifically, expected a posteriori [EAP] scores [23] with a standard normal prior; Supplementary material). These scores are presented on a continuous logit scale, with the majority of scores falling between −2.00 and +2.00, and a higher score indicating a poorer clinical state. After calculating EAP scores for each subscale, we repeated the primary PROSPER analysis once more, using these scores as dependent variables.

2.6. Stage 3: response shift bias and measurement invariance

We then tested for differential item functioning by age (<70 years vs. ≥70 years) using the logistic regression technique described by Choi et al. [24] This involves fitting logistic regression models that aim to predict item responses from overall scale scores. The addition of age as a covariate in these models should not improve model fit by a Nagelkerke pseudo-R^2 value of >2%, unless age influences the relationship between latent construct level and item response (and differential item functioning exists). To check for potential response shift bias, we repeated this process using time point (baseline vs. 12 months) as the covariate.

3. Results

3.1. Sample characteristics

The demographics and clinical characteristics of PROSPER study participants are presented in Table 1.

Table 1. Demographics and clinical characteristics of total PROSPER trial participants (n = 382 at baseline)

Median age (IQR) years	58 (49 to 68)
Type of breast surgery	
Mastectomy	157
Breast conserving surgery	222
Type of Axillary surgery	
Nodal clearance	327
Sentinel lymph node biopsy	52
Radiotherapy status	
Underwent radiotherapy	317
Did not undergo radiotherapy	40
Chemotherapy status	
Underwent chemotherapy	226
No chemotherapy	132
Ethnic group	
White	321
Indian	11
Pakistani	5
Mixed	2
African	1
Bangladeshi	1
Black or Black British	1
Caribbean	1
Chinese	1
Other	4

Abbreviations: IQR, interquartile range; PROSPER, Prevention of Shoulder Problems.

3.2. Stage 1: factorial structure

The analyses of factorial structure suggested that it was reasonable to consider the DASH as two independently functioning subscales in this cohort (the methodology and results of these analyses are reported in the Supplementary Material). Items from 1 through to 21 together measured a common health construct, which we named 'motor function' and items 22 through to 30 measured a second construct, which we named 'sensory symptoms'. The covariance of motor function and sensory symptoms was low (0.59), implying that it may be more appropriate to consider motor function and sensory symptoms separately rather than as a single construct.

This data-driven approach, based on the covariance structure of item responses, is supported by examining the face validity of the items. Items in the motor function subscale primarily ask about limitations in activities and procedural tasks, and items in the sensory symptoms subscale focus on pain and paresthesia. The division of items in this way is also consistent with previous studies of the factorial structure of the DASH and QuickDASH, when applied to hand conditions [18,19].

The primary analysis of the PROSPER trial suggested statistically significant differences in upper limb disability between groups at both six months (adjusted mean difference [MD] −4.60, 95% CI −8.90, −0.30; P = 0.036) and 12 months (MD −7.81, 95% CI −12.44, −3.17; P = 0.001). After accounting for the factorial structure of the DASH in this cohort, we found a statistically significant difference in motor function score at six months (MD

Table 2. Adjusted mean (95% CI) DASH scores by intervention arm for each analyses

Trial arm	Baseline	6 months	12 months
Primary PROSPER analysis using recommended DASH scoring			
Exercise	20.3 [15.4, 25.2]	18.8 [14.2, 23.4]	15.2 [10.4, 19.9]
Usual care	18.9 [13.7, 24.2]	23.4 [18.5, 28.3]	23.0 [18.1, 27.9]
P value	0.572	0.036	0.001
Secondary psychometric analyses			
Motor function sum-score			
Exercise	33.8 [28.9, 38.6]	36.4 [31.7, 41.1]	33.8 [27.9, 39.8]
Usual care	35.3 [30.1, 40.6]	41.2 [35.9, 46.6]	39.5 [33.5, 45.4]
P value	0.509	0.029	0.026
Sensory symptoms sum-score			
Exercise	16.4 [14.5, 18.3]	15.5 [13.9, 17.2]	14.0 [12.1, 15.8]
Usual care	16.3 [14.5, 18.3]	16.1 [14.4, 17.9]	16.9 [14.9, 18.9]
P value	0.907	0.475	0.002
Motor function EAP score			
Exercise	−0.07 [−0.30, 0.17]	−0.002 [−0.20, 0.195]	−0.15 [−0.35, 0.05]
Usual care	−0.08 [−0.33, 0.17]	0.111 [−0.105, 0.33]	0.092 [−0.12, 0.31]
P value	0.912	0.256	0.019
Sensory symptoms EAP score			
Exercise	−0.067 [−0.299, 0.165]	−0.051 [−0.250, 0.148]	−0.171 [−0.377, 0.034]
Usual care	−0.073 [−0.321, 0.175]	0.060 [−0.157, 0.278]	0.113 [−0.106, 0.333]
P value	0.958	0.265	0.007

Baseline scores are adjusted for age, type of breast surgery, type of axillary surgery, and the use of radiotherapy, and chemotherapy. Mean scores at six and 12 months are also adjusted for baseline score (either DASH scores, subscale sum-scores, or subscale expected a posteriori [EAP] scores in the respective analyses). *P* values relate to adjusted between-group differences at each time point.

Abbreviations: CI, confidence interval; DASH, Disability of Arm, Shoulder, and Hand; PROSPER, Prevention of Shoulder Problems.

−4.88, 95% CI −9.23, −0.52; $P = 0.029$) and 12 months (MD −5.61, 95% CI −10.5, −0.67; $P = 0.026$), and a difference in sensory symptom score at 12 months (MD −2.92, 95% CI −4.73, −1.10; $P = 0.002$), but not at six months (MD −0.57, 95% CI −2.14, 1.00; $P = 0.475$). Although the original PROSPER analysis might imply a progressive improvement in upper limb disability in the exercise intervention group, the factorial analysis suggests that motor scores for this group were similar between baseline and 12 months, whereas those in the control group deteriorated over time (Table 2, Fig. 1). Overall, this psychometric sensitivity analysis of the DASH supports the main trial conclusion that at 12 months, participants who received the exercise intervention had better motor function and sensory symptom outcomes than the control group.

3.3. Stage 2: item response theory theory modeling

The item responses to each subscale demonstrated good fit to the graded response model (Supplementary Material). This meant that we were able to generate EAP scores on a continuous scale that account for the weighting of items at different construct levels. With EAP scoring, we found no statistically significant difference between groups in motor function score (MD −0.11, 95% CI −0.308, 0.08; $P = 0.256$) or sensory symptom score (MD −0.11, 95% CI −0.31, 0.09; $P = 0.265$) at six months. At 12 months, there were differences in both motor function score (MD −0.241, 95% CI −0.442, −0.04; $P = 0.019$) and sensory symptom score (MD −0.29, 95% CI −0.49, −0.08; $P = 0.007$) between groups, both favoring the exercise intervention.

3.4. Stage 3: measurement invariance

For both motor and sensory subscales, we found no evidence of differential item functioning by age and no differential item functioning between-item responses at baseline and 12 months. This suggests that the probabilistic relationship between item responses and motor function or sensory symptom levels was constant between age groups ($<$ 70 years vs. $\geq$70 years) in this trial cohort and over the course of the study. Measurement invariance by age or time was unlikely to have biased the findings of the PROSPER study.

4. Discussion

This paper describes some key measurement assumptions (or decisions) that are often made when analyzing PROM data in clinical trials and demonstrates the impact they can have on a trial's results. The underlying

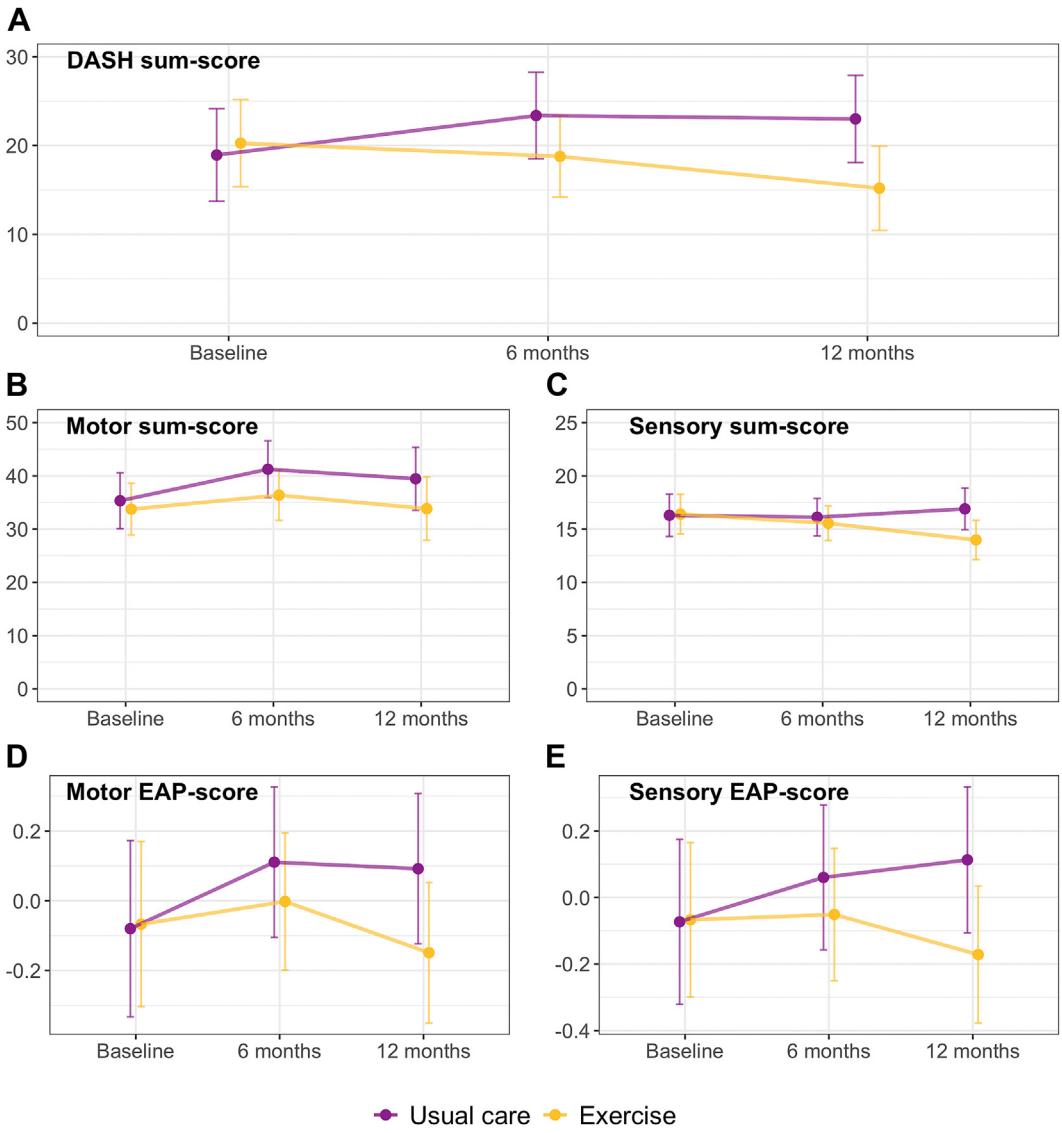

Fig. 1. Adjusted mean (95% CI) DASH sum-scores for each analysis. Panel (A) represents the original PROSPER analysis. Panels (B) and (C) demonstrate the repeat analysis, after accounting for subscales using sum-scoring. Panels (D) and (E) demonstrate the second repeat analysis, which accounts for subscales using expected a posteriori (EAP) scoring. The usual care group is illustrated in purple, whereas the exercise intervention group is illustrated in amber. Baseline scores are adjusted for age, type of breast surgery, type of axillary surgery, and the use of radiotherapy, and chemotherapy. Mean scores at six and 12 months are also adjusted for baseline score (either DASH scores, subscale sum-scores, or subscale EAP scores in the respective analyses). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

assumptions we focus on here are unidimensionality, correct item weighting, and measurement invariance, including response shift bias.

Our factorial analysis suggested an alternative subscale structure of the DASH to that used in PROSPER and other clinical trials following the official recommendations for scale scoring. The subscale structure we propose is consistent with other psychometric studies examining DASH [19,25], and these analyses provide additional, meaningful, clinical insights. By dividing the PROM into motor function and sensory symptom subscales, clinicians can see that apparent early (six-month postoperatively) differences in score were due to motor function item responses, whereas sensory symptoms (pain and paresthesia) were similar between groups at this time point.

Although these subscales are different to those used in the original PROSPER secondary subscale analysis, our conclusions are similar. The main trial found between-group differences in ‘activity limitation’ and ‘participation restriction’ subscale scores, but not in ‘impairment’ subscale scores, at six months. At 12 months, between-group differences were found in all three of the subscales proposed by Dixon et al. [1].

After accounting for the weighting of items with IRT, we found that the differences between groups (both in sensory symptoms and motor function) only became apparent at

12 months. Although the primary analysis of PROSPER might imply that the exercise program improves upper limb disability at 12 months vs. baseline, a reasonable alternative interpretation is that it prevents the deterioration seen in the control group. This is a helpful insight for healthcare practitioners who may need to manage patients' expectations when administering the exercise intervention.

The methods we used to assess factorial structure, item weighting, and measurement invariance are neither exhaustive nor prescriptive. There are many other techniques that could have been reasonably applied, and these may have led to different results. Similarly, we made judgments when interpreting our results that other researchers may reject (e.g., whether or not to exclude certain items from the analysis that were influenced by both sensory symptoms and motor function, such as item 23, which relates to limitations in work and daily activities; Supplementary material). A second challengeable decision we made was to assess differential item functioning by age after categorizing patients into those who were aged <70 years or ≥ 70 years. Common software packages for assessing differential item functioning require dichotomization like this, and so an age threshold separating group that may respond differently must be drawn. But we have no strong empirical reason to draw the threshold at 70 years, different researchers may have chosen a different boundary, and this might have led to different results. For these reasons, psychometric sensitivity analyses could easily become a target for 'p-hacking' [26]. One approach to deal with this may be to prespecify a range of psychometric sensitivity analyses in a trial's protocol, with explicit thresholds for decision-making. This could be incorporated into estimand frameworks for clinical trials using PROMs [27].

Another limitation of this work is that we have not questioned whether the trial's target difference represents a truly meaningful difference. There are several ways to estimate a PROM's MID, with varying rigor, and for the purposes of sample size calculations and result interpretation, trialists must often select an MID based on estimates from different contexts which may not generalize perfectly to the sample population [28]. In the PROSPER trial, the authors recognized this uncertainty by providing a range of plausible MID values for the DASH. These ranges help readers to understand whether a statistically significant difference is also clinically meaningful. Because we altered the scoring of the PROM in our secondary psychometric analyses (through subscale sum-scoring and IRT), it is difficult to compare our observed differences to original MID estimates. This issue should be a target for development in the area of psychometric sensitivity analysis and might be partially addressed through the use of anchors (e.g., global rating of change scales [29]) or standardized effect size heuristics [30].

While psychometric sensitivity analyses may address some of the assumptions surrounding the construct validity of an outcome measure, they do not address assumptions related to the measure's content validity (the relevance and comprehensiveness of the items posed). The DASH was not developed with patients undergoing breast cancer surgery and its content validity in this population could be explored further through cognitive debriefing studies. Finally, psychometric sensitivity analyses do not address broader trial limitations, for example, those relating to intervention adherence, attrition, or the placebo effect. A detailed discussion of the broader strengths and limitation of the PROSPER trial can be found in its primary publication [1]. These considerations must be taken into account together when drawing conclusions based on trial results.

5. Conclusion

Not all PROMs provide unidimensional, interval-scaled, and invariant measurement. The measurement assumptions and decisions that clinical trialists must make when using PROMs as primary outcomes can influence the results of RCTs. Psychometric sensitivity analysis is a framework for testing the stability of a trial's results to these assumptions. This study provides further evidence that the DASH functions as two independent subscales, but this violation of unidimensionality does not alter the principal conclusion of the PROSPER study.

Further work is needed to understand what role psychometric sensitivity analyses could play in RCTs and which techniques might be most appropriate for different circumstances. Additionally, work is needed to improve the accessibility and practicality of these analyses. For now, we recommend that psychometricians and measurement scientists are involved in the planning and execution of RCTs that use PROMs and that trialists aim to select PROMs that demonstrate high standards of content and construct validity, as defined in the consensus-based standards of the selection of health measurement instruments [31].

CRediT authorship contribution statement

Conrad J. Harrison: Conceptualization, Methodology, Formal analysis, Investigation, Project administration, Software, Visualization, Writing − original draft, Writing − review & editing. **Anower Hossain:** Data curation, Formal analysis, Investigation, Project administration, Resources, Writing − review & editing. **Julie Bruce:** Investigation, Project administration, Resources, Writing − review & editing. **Jeremy N. Rodrigues:** Data curation, Formal analysis, Project administration, Writing − review & editing, Supervision.

Declaration of competing interest

C.J.H. reports financial support was provided by National Institute for Health and Care Research. J.N.R. finanacial support was provided by the National Institute for

Health and Care Research. J.B. reports financial support was provided by National Institute for Health and Care Research. A.H. reports financial support was provided by National Institute for Health and Care Research.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2023.09.008.

References

[1] Bruce J, Mazuquin B, Canaway A, Hossain A, Williamson E, Mistry P, et al. Exercise versus usual care after non-reconstructive breast cancer surgery (UK PROSPER): multicentre randomised controlled trial and economic evaluation. BMJ 2021;375:e066542.

[2] Beard DJ, Davies LJ, Cook JA, MacLennan G, Price A, Kent S, et al. The clinical and cost-effectiveness of total versus partial knee replacement in patients with medial compartment osteoarthritis (TOPKAT): 5-year outcomes of a randomised controlled trial. Lancet 2019;394(10200):746−56.

[3] Beard DJ, Rees JL, Cook JA, Rombach I, Cooper C, Merritt N, et al. Arthroscopic subacromial decompression for subacromial shoulder pain (CSAW): a multicentre, pragmatic, parallel group, placebo-controlled, three-group, randomised surgical trial. Lancet 2018; 391(10118):329−38.

[4] Nguyen TH, Han HR, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. Patient 2014;7(1):23−35.

[5] Vanier A, Oort FJ, McClimans L, Ow N, Gulek BG, Böhnke JR, et al. Response shift in patient-reported outcomes: definition, theory, and a revised model. Qual Life Res 2021;30:3323−4.

[6] Howard J, Mattacola C, Howell D, Lattermann C. Response shift theory: an application for health-related quality of life in rehabilitation research and practice 2004. Available at https://core.ac.uk/display/345086799?utm_source=pdf&utm_medium=banner&utm_campaign=pdf-decoration-v1. Accessed December 14, 2022.

[7] Teresi JA, Ramirez M, Lai JS, Silver S. Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: description of DIF methods, and review of measures of depression, quality of life and general health. Psychol Sci Q 2008;50(4):538.

[8] Schwartz CE, Huang IC, Rohde G, Skolasky RL. Listening to the elephant in the room: response-shift effects in clinical trials research. J Patient Rep Outcomes 2022;6(1):105.

[9] Gorter R, Fox JP, Twisk JWR. Why item response theory should be used for longitudinal questionnaire data analysis in medical research. BMC Med Res Methodol 2015;15:55.

[10] Gorter R, Fox JP, Apeldoorn A, Twisk J. Measurement model choice influenced randomized controlled trial results. J Clin Epidemiol 2016;79:140−9.

[11] Gorter R, Fox JP, Riet GT, Heymans M, Twisk J. Latent growth modeling of IRT versus CTT measured longitudinal latent variables. Stat Methods Med Res 2020;29(4):962−86.

[12] Harrison CJ, Plessen CY, Liegl G, Rodrigues JN, Sabah SA, Cook JA, et al. Item response theory may account for unequal item weighting and individual-level measurement error in trials that use PROMs: a psychometric sensitivity analysis of the TOPKAT trial. J Clin Epidemiol 2023;158:62−9.

[13] Hudak PL, Amadio PC, Bombardier C, Beaton D, Cole D, Davis A, et al. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder, and hand). Am J Ind Med 1996;29(6):602−8.

[14] Insititute for Work and Health. Scoring the DASH 2006. Available at https://dash.iwh.on.ca/sites/dash/files/downloads/dash_scoring_2010.pdf. Accessed December 14, 2022.

[15] Andrich D. Rating scales and Rasch measurement. Expert Rev Pharmacoecon Outcomes Res 2011;11(5):571−85.

[16] Fayad F, Lefevre-Colau MM, Macé Y, Fermanian J, Mayoux-Benhamou A, Roren A, et al. Validation of the French version of the disability of the arm, shoulder and hand questionnaire (F-DASH). Joint Bone Spine 2008;75(2):195−200.

[17] Franchignoni F, Giordano A, Sartorio F, Vercelli S, Pascariello B, Ferriero G. Suggestions for refinement of the disabilities of the arm, shoulder and hand outcome measure (DASH): a factor analysis and Rasch validation study. Arch Phys Med Rehabil 2010;91: 1370−7.

[18] Rodrigues J, Zhang W, Scammell B, Russell P, Chakrabarti I, Fullilove S, et al. Validity of the disabilities of the arm, shoulder and hand patient-reported outcome measure (DASH) and the quick-dash when used in dupuytren's disease. J Hand Surg Eur Vol 2016; 41(6):589−99.

[19] Stirling PHC, McEachan JE, Rodrigues JN, Harrison CJ. Quick-DASH questionnaire items behave as 2 distinct subscales rather than one scale in Dupuytren's disease. J Hand Ther 2021;36:228−33.

[20] Dixon D, Johnston M, McQueen M, Court-Brown C. The disabilities of the arm, shoulder and hand questionnaire (DASH) can measure the impairment, activity limitations and participation restriction constructs from the international classification of functioning, disability and health (ICF). BMC Musculoskelet Disord 2008;9(1): 114.

[21] Schreiber JB, Nora A, Stage FK, Barlow EA, King J. Reporting structural equation modeling and confirmatory factor analysis results: a review. J Educ Res 2006;99(6):323−38.

[22] Samejima F. Graded response model. In: van der Linden WJ, Hambleton RK, editors. Handbook of modern item response theory. New York: Springer; 1997:85−100.

[23] Chapman R. Expected a posteriori scoring in PROMIS®. J Patient Rep Outcomes 2022;6(1):59.

[24] Choi SW, Gibbons LE, Crane PK. Lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. J Stat Softw 2011;39(8):1−30.

[25] Harrison C, Clelland AD, Davis TRC, Scammell BE, Zhang W, Russell P, et al. A comparative analysis of multidimensional computerized adaptive testing for the DASH and QuickDASH scores in Dupuytren's disease. J Hand Surg Eur Vol 2022;47(7):750−4.

[26] Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of P-hacking in science. PLoS Biol 2015;13(3): e1002106.

[27] Fiero MH, Pe M, Weinstock C, King-Kallimanis BL, Komo S, Klepin HD, et al. Demystifying the estimand framework: a case study using patient-reported outcomes in oncology. Lancet Oncol 2020; 21(10):e488−94.

[28] Cook JA, Julious SA, Sones W, Hapson LV, Hewitt C, Berlin JA, et al. DELTA2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. BMJ 2018;363:k3750.

[29] Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. J Man Manip Ther 2009;17(3):163−70.

[30] Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003;41:582−92.

[31] Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010;10:22.