European Spine Journal (2022) 31:943–952
https://doi.org/10.1007/s00586-022-07109-x

ORIGINAL ARTICLE

Validation of the work ability index—single item and the pain disability index—work item in patients with chronic low back pain

Imke Boekel[1] · Alisa L. Dutmer[2] · Henrica R. Schiphorst Preuper[2] · Michiel F. Reneman[2]

Received: 7 July 2021 / Revised: 17 December 2021 / Accepted: 4 January 2022 / Published online: 23 January 2022
© The Author(s) 2022

Abstract

Purpose A cross-sectional and longitudinal study was conducted to analyse construct validity, responsiveness, and Minimal Clinically Important Change (MCIC) in the Work Ability Score (WAS) and Pain Disability Index Work item (PDI-W) in patients with Chronic Low Back Pain (CLBP).

Method Construct validity was assessed by testing predefined hypotheses. Responsiveness and MCIC were measured with an anchor-based method. The area under the receiver Operating Characteristic Curve (AUC) and the optimal cut-off point were calculated. Smallest Detectable Change (SDC) was calculated to determine measurement error.

Results In total, 1502 patients (age 18–65 years) with CLBP were included. For validity of the WAS and PDI-W, respectively, seven and six out of 10 hypotheses were not rejected. The WAS ($n = 355$) was responsive to change with an AUC of 0.70. MCIC was 1.5 point, $\text{SDC}_{\text{individual}}$ 4.9, and $\text{SDC}_{\text{group}}$ 0.3. MCICs were 4.5, 1.5, and − 0.5 points for, respectively, low, middle, and high scoring baseline groups. The PDI-W ($n = 297$) was responsive to change with an AUC of 0.80. MCIC was − 2.5 points, $\text{SDC}_{\text{individual}}$ 5.2, and $\text{SDC}_{\text{group}}$ 0.3. MCICs were − 0.5, − 2.5, and − 4.5 points for, respectively, low, middle, and high scoring baseline groups.

Conclusion Construct validity of the WAS and PDI-W was insufficient in this patient sample. The WAS and PDI-W are responsive to change. On average, improvements of 1.5 point (WAS) and − 2.5 points (PDI-W) were interpreted as clinically important. However, MCICs are also baseline dependent. Due to a risk of measurement error, at the individual level change scores should be interpreted with caution.

Keywords Construct validity · Responsiveness · Minimal clinically important difference · Smallest detectable change · Longitudinal validity

Introduction

The prevalence of disability due to Low Back Pain (LBP) increases from the third decade of life on, peaking between the age of 35 and 55 years [1]. LBP causes a large number of absenteeism and work productivity losses [2]. This makes LBP the most common health problem in the European workforce. In the Netherlands, costs of LBP have been estimated at 1.7% of its Gross National Product [3].

The greatest potential for cost reduction is decreasing work absenteeism and disability due to LBP [4]. Absenteeism and disability at work are influenced by the work ability of a person [5]. Higher work ability is associated with less disability and pain, and higher quality of life [6]. The Work Ability Index (WAI) was developed as a measure for self-reported work ability. The Work Ability Score (WAS) is an item of the WAI and compares current work ability with lifetime best [7]. It is an acceptable brief alternative for the WAI in determining work ability [8]. Convergent validity between the WAI and WAS is sufficient [9]. Measurement properties are sufficient in a secondary vocational rehabilitation setting [10], but have not been analysed in a secondary and tertiary spine care. The interference of chronic pain

✉ Michiel F. Reneman
m.f.reneman@umcg.nl

[1] Faculty of Medicine, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

[2] Department of Rehabilitation Medicine, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

with daily activities can be assessed by the Pain Disability Index (PDI). The PDI has been validated in patients with chronic pain [11]. The PDI Work item measures interference of chronic pain with the ability to engage in occupational activities. The PDI-W has also not been validated yet.

The WAS and PDI-W are Patient Reported Outcome Measures (PROMs). PROMs are highly recommended in clinical guidelines to assess the quality of care, treatment effects and change in health status from the patient's perspective. Selection of PROMs should be based on the strength of relevant measurement characteristics (i.e. validity, responsiveness) [12]. To meet conditions for construct validity, a measurement instrument should be consistent with hypotheses regarding relationships with other measures. The ability to detect changes in health status within individuals over time (responsiveness) and interpretation of change scores are important characteristics of PROMs [13]. Minimal Clinically Important Change (MCIC) and measurement error (Smallest Detectable Change, SDC) can be used to interpret change scores. The MCIC is useful as this change score is perceived as beneficial and meaningful to patients [14].

Despite the usefulness of PROMs, these measurements can be a burden for patients and caregivers. Length of time to fill out the questionnaires, difficulty in completing them independently, and length of time to analyse the results were the most frequently mentioned reasons for not using the measurements [15]. Therefore, if measurement characteristics are sufficient, the WAS and PDI-W single items may be used in routine care instead of lengthy questionnaires. The aim of the present study was to assess construct validity, responsiveness, and MCIC of the WAS and PDI-W in patients with CLBP.

Materials and methods

Study design

Data collection took place in the Groningen Spine Center, a university-based multidisciplinary secondary and tertiary care center in the Netherlands, as part of a research project called Groningen Spine Cohort (GSC) [16]. Patients were included from July 2015 to September 2018. Patients digitally filled out a set of questionnaires at baseline (T0) and 6 months follow-up (T1), including the WAS, PDI-W, a modified version of the short Copenhagen Psychosocial Questionnaire (COPSOQ II), the EuroQol-5D (EQ5D), the National Institutes of Health (NIH) Impact Stratification, and a Global Perceived Effect (GPE) scale. The Medical Ethical Committee of the University Medical Center Groningen provided a waiver (M15.169472) for the data collection of the GSC with respect to medical ethical permission. Informed consent was signed after being informed on the purpose of the study. Handling of the data was done in accordance with the guideline for Good Research Practice [17].

Patients and setting

Patients admitted to the Groningen Spine Center between 18 and 65 years old, experiencing lower back and/or leg pain for more than 12 weeks were included. Patients with insufficient understanding of Dutch language or no Internet access were excluded. Patients who did not respond to follow-up questionnaires were excluded for assessment of longitudinal validity. Retired and permanent disabled patients, and stay at home parents were also excluded for analyses, because improvement of work ability was not a treatment goal for these patients. Care as usual was provided between baseline and follow-up. Treatment options were minimal intervention (pain education and information), referral for primary care treatment (e.g. physiotherapy), multidisciplinary rehabilitation, surgery, pain anaesthesiology treatment, and/or other if needed (e.g. referral to an additional medical specialist) [16].

Measurements

Primary measures

The WAS measures current work ability compared to lifetime best. Scores range from 0 (completely unable to work) to 10 (work ability at best period). Scores are classified as poor (0–5 points), moderate (6–7 points), good (8–9 points), and excellent (10 points) work ability. The WAS has been demonstrated to be a good brief alternative for the lengthy WAI in determining work ability [8]. Convergent validity is acceptable between the WAI and WAS [9]. Measurement properties in a secondary rehabilitation setting were sufficient [10].

The PDI-W measures self-reported work disability, including paying and non-paying jobs. The wording of PDI-W is "This category refers to activities that are part of or directly related to one's job. This includes non-paying jobs as well, such as that of a housewife or volunteer". Patients rate their level of pain interference on each item on a scale ranging from 0 (no pain interference) to 10 (total pain interference). Patients can also choose the option 'not applicable'. The PDI-W is an item of the PDI, which consists of 7 items. The PDI is internally consistent (Cronbach's alpha = 0.87) with good convergent validity. Test–retest reliability is good (Intraclass Correlation Coefficient [ICC] = 0.76) [11]. Measurement properties of the PDI-W are unknown.

Secondary measures

Absenteeism was measured for employed patients by calculating the percentage missed hours of work in the past month (0 to 100%). Missing data on weekly work hours (in 14% of employed patients) and weekly work days (in 7% of employed patients) were handled with multiple imputation. Constraints were set on the minimum and maximum number of work hours (between 2 and 40 h) and work days (between 1 and 5) per week.

Quality of life was measured with the EQ5D consisting of five dimensions: mobility, self-care, activities of daily living, pain/discomfort, and anxiety/depression. These dimensions are measured on three levels (no problems, some problems, extreme problems) [18]. EQ5D scores can be transformed into a utility value by means of a validated algorithm, ranging from -0.33 to 1.00 [19]. Reliability and validity of the ED5Q are acceptable [20].

Psychosocial work environment was measured with the COPSOQ II which consists of 25 items covering domains work demands, work organization and job contents, leadership and interpersonal relationships, work-individual interface, and values at the workplace. Test–retest reliability is adequate to good [21].The scale considering work demands was used in the present study (6 questions, 0–4, 4 being highest self-reported work demands), consisting of items concerning work pace, quantitative and emotional demands.

The NIH Impact Stratification measures self-reported impact of pain on functioning. The Impact Stratification total score is realized by nine items: a Numeric Rating Scale (NRS) of pain intensity (range 0–10), four items on physical functioning (range 1–5), and four items on pain interference (range 1–5). The total score ranges from 8 (least impact) to 50 (most impact). The impact is classified as mild (8–27 points), moderate (28–35 points), or severe (≥ 35 points) [22]. Test–retest reliability is sufficient and longitudinal validity is adequate [23, 24]. The NRS pain score was also used as single item in the present study.

For work status, patients answered the question whether they were currently employed. If yes, patients were asked about the status of the employment (working, sick leave, partial sick leave). In addition, an item of the NIH minimal dataset concerning permanent disablement was used [22].

The GPE was used as external criterion. The GPE measures patient-rated assessment of treatment outcome [25]. The question "How much did your treated complaints change compared with pre-treatment level?" was answered at 6 months follow-up (T1). Responses range from 0 to 6 on a 7-point Likert scale: 0, extremely worsened; 1, much worsened; 2, little worsened; 3, unchanged; 4, little improved; 5, much improved; and 6, completely improved. Strong correlations are reported between GPE scores and changes in disability and pain [26]. There is limited evidence for validity of the GPE scale [27]. Nevertheless, the GPE is recommended and considered a suitable anchor [28].

Data analyses

Patient characteristics were reported using descriptive statistics. Continuous data are presented as means and standard deviations or medians and Interquartile Range (IQR), depending on data distribution. Categorical data are reported as frequencies with percentages. $P < 0.05$ was considered statistically significant. Statistics were performed using IBM SPSS for Windows (version 24.0, Armonk, NY: IBM Corp.).

Construct validity was assessed using elaborate hypothesis testing by calculating Pearson or Spearman's rank correlation coefficients between the WAS and PDI-W on the one hand, and absenteeism, EQ5D, COPSOQ II work demands, NIH Impact Stratification, NRS pain score, and items concerning work status on the other hand. The strength of the association is expressed in the absolute value of the correlation coefficient. Correlation coefficients were interpreted as follows: 0.00–0.25 little if any, 0.26–0.49 weak, 0.50–0.69 moderate, 0.70–0.89 strong, and 0.90–1.00 very strong correlation [29]. Hypotheses were formulated based on previous studies. Weak correlations of the WAS and PDI-W with work demands [30], partial sick leave and permanent disablement [31, 32] were expected. We expected weak to moderate correlations of the WAS and PDI-W with absenteeism [5], NRS pain score [32], working [33], and sick leave [31, 32]. Finally, we expected moderate correlations of the WAS and PDI-W with quality of life [34] and impact of pain on functioning [35]. Correlations of the WAS and PDI-W with absenteeism and work demands were calculated for the 'employed' patient group ($n = 901$). Because a floor or ceiling effect of maximally 15% is considered acceptable [36], this hypothesis was not rejected if < 15% of patients achieved the lowest or highest possible score (0–10, respectively). Support for construct validity was considered when ≥ 80% of the predefined hypotheses were not rejected [36].

For hypothesis testing, we used the COSMIN Study Design checklist for Patient reported outcome measurement instruments [37]. The section 'Hypotheses testing for construct validity' includes the following requirement: "Perform the analysis in a sample with an appropriate number of patients (taking into account expected number of missing values)". A patient sample of ≥ 100 patients is considered "very good". The patient samples for hypothesis testing in the present study varied from 884 to 1502 patients. Therefore, we did not include a sample size and power calculation.

Responsiveness and MCIC were calculated in accordance with the Consensus-based Standards for the Selection of Health Measurement Instruments criteria (COSMIN) [13, 37]. Two categories were defined: improved (i.e. *much improved* and *completely improved*) and unimproved (all other scores). The area under the Receiver Operating Characteristics (ROC) curve (AUC) was calculated with a 95% Confidence Interval (CI). AUCs were interpreted as excellent 1.0–0.90, good 0.90–0.80, fair 0.80–0.70, poor 0.70–0.60, and failed 0.60–0.50 [38]. Therefore, an AUC ≥ 0.70 was considered responsive. The Optimal Cut-off Point (OCP) of the AUC was determined to calculate the MCIC. This is the point closest to the top-left corner of the ROC curve, which maximizes the proportion of correctly classified patients according to the GPE. The sum of squares of 1-sensitivity and 1-specificity is minimal at this point [39]. Cut-off values depend on the baseline level of a measurement [40]. To take the effect of baseline scores into account, secondary analyses were performed measuring responsiveness and MCICs for different baseline-score groups. Three subgroups were formed based on tertiles of baseline scores on the WAS and PDI-W. In addition, a percentage change score was calculated for the PDI-W to take the effect of baseline scores into account. For the WAS, an inverted percentage change score was calculated since calculation of a percentage change score resulted in a high frequency of missing values due to division by zero. The inverted percentage change score was calculated by the following equation:

$$\text{Inverted percentage change score} = \frac{\text{change score}}{\text{maximum score} - \text{baseline score}} \times 100 \tag{1}$$

The Standard Error of Measurement (SEM) and Smallest Detectable Change (SCD) were calculated to determine measurement error. The SEM represents the standard deviation of repeated measurements in one patient. The SEM was calculated by the square root of the variability between baseline and 6 months follow-up added to the variability due to random error (Eq. 2) in patients reporting *unimproved* scores on the GPE [41]. The SPSS VARCOMP analysis command was used to retrieve the variance component values.

$$\mathrm{SEM} = \sqrt{\sigma^2_{time} + \sigma^2_{error}} \tag{2}$$

The SDC represents the minimal change in score that a patient has to show to ensure that the observed change is real and not caused by measurement error. To distinguish clinically important change from measurement error, SDC should be smaller than the MCIC. The SDC can be calculated by the following equations on individual (Eq. 3) and group level (Eq. 4), in which 1.96 refers to a 95% CI and $\sqrt{2}$ is a correction for repeated measurements [41, 42]:

$$\mathrm{SDC}_{\mathrm{individual}} = 1.96 \times \sqrt{2} \times SEM \tag{3}$$

$$\mathrm{SDC}_{\mathrm{group}} = \frac{\mathrm{SDC}_{\mathrm{individual}}}{\sqrt{n}} \tag{4}$$

Results

Patients

Baseline data were available for 1502 patients. Demographic and clinical variables are presented in Table 1. The median of the WAS was 4.0 (IQR 1.0–6.0) points, and the median of the PDI-W was 7.0 (IQR 5.0–8.0) points.

Hypothesis testing

For the WAS, 7 of 10 hypotheses were not rejected (70%) (Table 2). For the PDI-W, 6 of 10 hypotheses were not rejected (60%). Therefore, construct validity of the WAS and PDI-W was not supported.

Responsiveness and minimal clinically important change

The WAS showed fair responsiveness with an AUC of 0.70 (0.63–0.76). The PDI-W showed a good responsiveness with an AUC of 0.80 (0.74–0.87) (Table 3; Fig. 1). The MCIC was 1.5 point for the WAS with a sensitivity of 0.67 and specificity of 0.61. The MCIC was − 2.5 points for the PDI-W with sensitivity of 0.74 and specificity of 0.75.

Baseline-score groups

For the PDI-W, MCICs were − 0.5, − 2.5, and − 4.5 points for, respectively, low, middle, and high scoring baseline groups (Table 4). The MCIC of the percentage change score was 56%. For the WAS, MCICs were 4.5, 1.5, and -0.5 points for, respectively, low, middle, and high scoring baseline groups (Table 5). The MCIC of the inverted percentage change score was 38%.

Measurement error

The SEM was 1.9 points for the PDI-W with an $\mathrm{SDC}_{\mathrm{individual}}$ of 5.2 and an $\mathrm{SDC}_{\mathrm{group}}$ of 0.3 points. The SEM for the WAS was 1.8 points with an $\mathrm{SDC}_{\mathrm{individual}}$ of 4.9 and an $\mathrm{SDC}_{\mathrm{group}}$ of 0.3 points.

Table 1 Patient characteristics at baseline ($n = 1502$)

Characteristic	Patients ($n = 1502$)
Age, y, mean ± *SD*	46.3 ± 12.8
Sex, n (%)	
Female	857 (57)
Level of education, n (%)	
No education	29 (2)
Low	522 (35)
Middle	487 (32)
High	343 (23)
Other	121 (8)
Medical history	
Duration LBP, *n* (%)	
< 3 mo	40 (3)
3 mo-1 y	240 (16)
1–5 y	527 (35)
> 5 y	695 (46)
Work	
Work status, n (%)	
Not working	601 (40)
Permanently disabled	253 (17)
Employed	901 (60)
Working	409 (27)
Partial sick leave	260 (17)
Sick leave	232 (15)
WAS (1–10), median (IQR)	4.0 (1.0–6.0)
PDI-W (0–10), median (IQR)	7.0 (5.0–8.0)
COPSOQ Work demands (0–4), mean ± *SD*[1]	2.6 ± 0.8
Absenteeism (0–100%), median (IQR)[2]	0.0 (0.0–25.0)
Quality of life	
EQ5D: utility value (-0.33–1.00), median (IQR)	0.39 (0.17–0.72)
Pain and functioning	
PDI total (0–70), mean ± *SD*	38.2 ± 14.0
NIH minimal dataset Impact Stratification (8–50), mean ± *SD*	35.2 ± 7.5
Mild (8–27), *n* (%)	232 (16)
Moderate (28–34), *n* (%)	402 (26)
Severe (≥ 35), *n* (%)	868 (58)
NRS score back pain (0–10), median (IQR)	7.0 (6.0–8.0)

LBP, low back pain; WAS, work ability score; IQR, interquartile range: quartile 1 to 3; PDI-W, pain disability index work item; COPSOQ, Copenhagen Psychosocial Questionnaire; EQ5D, Euroqol-5D; PDI, pain disability index; NIH, National Institutes of Health; NRS, numerical rating scale

[1] $n = 899$, [2] $n = 807$

Discussion

The aim of this study was to assess construct validity, responsiveness, and MCIC of the WAS and PDI-W in patients with CLBP. For the WAS and PDI-W, respectively, 70% and 60% of predefined hypotheses were not rejected, which is lower than the threshold that was set at ≥ 80%. Therefore, construct validity was not supported. The WAS and PDI-W are responsive to change. MCICs of 1.5 point (WAS) and -2.5 points (PDI-W) were found. Nevertheless, clinically important change could not be distinguished from measurement error, since MCICs were smaller than SDC values. Individual change scores up to 5 points should be interpreted with caution.

For construct validity, rejection of more hypotheses than expected can be explained by different reasons. For the WAS, measurement scales of reference instruments might have contributed to the rejection of hypotheses. The WAS asks to compare current work ability to lifetime best, whereas reference instruments only ask for current functioning. Consequently, loss of functioning might have been scored differently resulting in lower correlations. Additionally, the work demands hypothesis focused on work pace, emotional and quantitative demands (i.e. evaluation of time available to finish work), because these questions were classified as work demands by the COPSOQII questionnaire. In retrospect, this construct could have been expanded by inclusion of physical demands, commitment to work, and job satisfaction. These factors are considered important factors in predicting work (dis)ability [30]. Hypotheses on partial sick leave were rejected. Contrary to the scores on the WAS and PDI-W for patients on sick leave or fully working that were heavily skewed to the lower or higher end of the scale, scores for patients on partial sick leave were normally distributed with a high variance. Therefore we observed lower correlations than a priori hypothesized. Finally, in hindsight, permanent disablement was an insufficient reference test. The majority of permanently disabled patients scored the PDI-W as 'not applicable', because this item was irrelevant to these patients.

Floor (WAS; 25%) and ceiling (PDI-W; 15%) effects were also observed, both indicating most severe interference of LBP. These effects might have affected correlations with reference tests. Data were collected from patients receiving secondary and tertiary multispecialty care. Consumption of medical care and the influence of LBP on work ability are higher in this patient sample, compared to patients receiving primary level care [16]. Therefore, the WAS and PDI-W might not be adequate instruments for distinguishing work ability levels in patients with severe CLBP. Further research should investigate the validity of these items in patients receiving primary care level.

Regarding longitudinal validity, measurement error should be considered in decision-making in individual patients. The $\text{SDC}_{\text{individual}}$ for the WAS (4.9 points) and PDI-W (5.2 points) both exceeded the MCIC values (respectively, 1.5 and -2.5 points). This corresponds with results of previous research on PROMs in back pain [11, 43].

Table 2 Correlations between the WAS ($n = 1502$) and PDI-W ($n = 1381$) and reference tests, and floor and ceiling effects

	Reference test	Criteria for non-rejection	WAS	Rejected	PDI-W	Rejected
1	Absenteeism	$0.25<\|r\|<0.70$	– 0.62[a]	No	0.52[a]	No
2	EQ5D	$0.50\leq\|r\|<0.70$	0.50	No	– 0.52	No
3	COPSOQ II work demands	$0.25<\|r\|<0.50$	– 0.07*[a]	Yes	0.11[a]	Yes
4	NIH Impact Stratification	$0.50\leq\|r\|<0.70$	– 0.59	No	0.61	No
5	NRS pain score	$0.25<\|r\|<0.70$	–0.33	No	0.37	No
6	Working	$0.25<\|r\|<0.70$	0.58	No	– 0.45	No
7	Partial sick leave	$0.25<\|r\|<0.50$	0.18	Yes	– 0.10	Yes
8	Sick leave	$0.25<\|r\|<0.70$	–0.38	No	0.38	No
9	Permanent disablement	$0.25<\|r\|<0.50$	– 0.35	No	0.17	Yes
10	Floor and ceiling effect	$<15\%$	24.1	Yes	15.1	Yes

r = correlation coefficient, absolute value. The direction of the association is depending on the scoring of the reference test

WAS, work ability score; n, number of patients; PDI-W, pain disability index work item; EQ5D, Euroqol-5D; COPSOQ II, Copenhagen Psychosocial Questionnaire modified version; NIH, National Institutes of Health; NRS, numeric rating scale

* *p = .04. All other correlations were significant (p < .01).* [a] *n = 884–901*

Table 3 Responsiveness and Minimal Clinically Important Change in the WAS ($n = 335$) and PDI-W ($n = 297$)

	WAS	PDI-W
Improved patients, *n* (%)	66 (19.7)	58 (19.5)
Scores		
Score T0, mean ± *SD*	4.3 ± 3.0	6.3 ± 2.7
Min–max	0–10	0–10
Score T1, mean ± *SD*	5.3 ± 2.9	4.9 ± 3.1
Min–max	0–10	0–10
Mean change ± *SD*	1.0 ± 2.7*	− 1.4 ± 3.0*
95% CI of mean change	0.75; 1.3	− 1.0; − 1.7
Change (%)	23.3	22.2
Responsiveness		
AUC	0.70	0.80
95% CI	0.63–0.76	0.74–0.87
Agreement with GPE (%)[a]	56	72
MCIC		
OCP	1.5	− 2.5
Sensitivity (%); specificity (%)	67; 61	74; 75

*Significant change between baseline and 6 months follow-up ($p < 0.001$)

[a]Agreement with GPE: percentage of patients with 1) a GPE 'improved' in combination with a clinically important change in score, or 2) a GPE 'improved' in combination with no clinically important change in score

WAS, work ability score; n, number of patients; PDI-W, pain disability index work item; SD, standard deviation; min, minimum; max, maximum; CI, confidence interval; AUC, area under the curve; GPE, global perceived effect; MCIC, minimal clinically important change; OCP, optimal cut-off point

Individual change scores larger than the MCIC but smaller than the $\mathrm{SDC}_{\mathrm{individual}}$ should be interpreted with caution. These scores fall within the measurement error, which results in the risk of incorrect classification of patients as improved. The WAS and PDI-W are better at detecting changes at a group level, as the $\mathrm{SDC}_{\mathrm{group}}$ was smaller than the $\mathrm{SDC}_{\mathrm{individual}}$. Results of the present study indicate that small changes in work ability can be considered important by CLBP patients. Because CLBP is very disabling [7, 8], small improvement can have meaningful effect on well-being of patients.

For interpretation of individual change scores, the effect of baseline scores should be taken into account [40]. Higher (PDI-W) or lower (WAS) baseline values (both indicating worse work ability) require higher MCIC values, because there is a greater potential for improvement [44]. The results of the present study confirm that MCICs for the WAS and PDI-W are baseline dependent. This is supported by the (inverted) percentage change scores, which were 39% (WAS) and 56% (PDI-W).

Patient burden is an important consideration in selecting measurement instruments. If patient burden is decreased by using single items instead of lengthy questionnaires, then slightly less sufficient measurement characteristics might be acceptable. For example, when patients have to fill out multiple questionnaires or in frequent evaluations (e.g. daily or weekly) assessing work ability trends. In addition, the WAS can be considered on group level and large-scale surveys [45]. The WAS is also suitable for systematic application during medical examinations in occupational health care or in public health surveys [9].

A methodological consideration is the dichotomization of the external criterion into improved and unimproved patient groups. The improved group consisted of patients reporting to be *much improved* and *completely improved*.

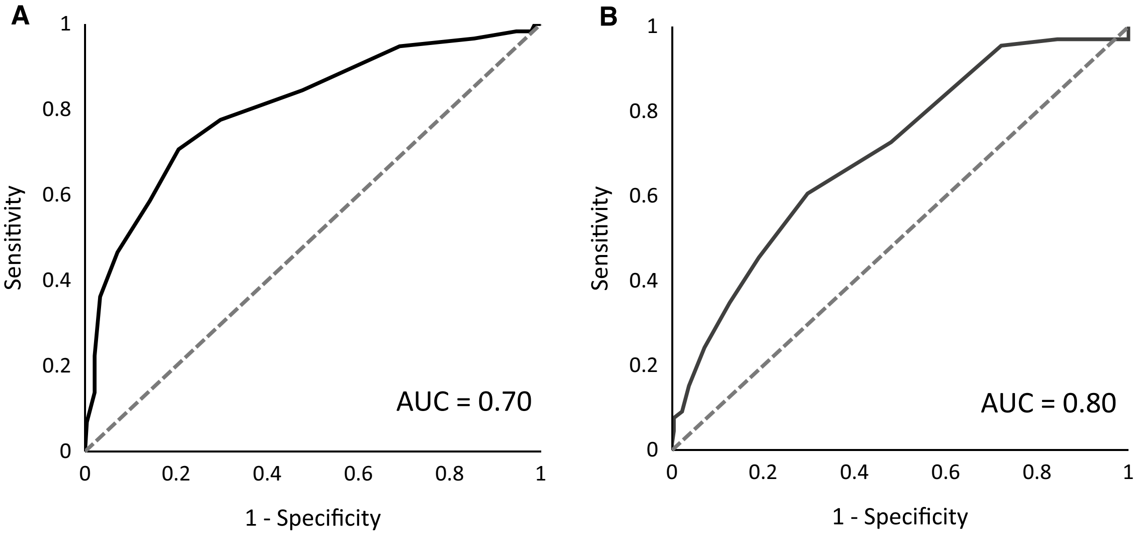

Fig. 1 Receiver Operating Characteristic curves (ROC) of the Work Ability Score (WAS) and Pain Disability Index – Work item (PDI-W). AUC, area under the curve. A. ROC-curve of WAS (*n* = 335). B. ROC-curve of PDI-W (*n* = 297)

Table 4 Responsiveness and Minimal Clinically Important Change in the PDI-W (*n* = 297) for baseline-score groups

	PDI– W			
	Baseline Tertile 1	Baseline Tertile 2	Baseline Tertile 3	Percentage change score
Patients, *n*	102	86	109	285
Improved patients, *n* (%)	17 (16.7)	21 (24.4)	20 (18.3)	55 (19.3)
Scores				
Score T0, mean ± *SD*	3.2 ± 1.8	6.7 ± 0.47	8.9 ± 0.86	
Min–max	0–5	6–7	8–10	
Score T1, mean ± *SD*	3.3 ± 2.7	4.9 ± 2.4	6.4 ± 3.2	
Min–max	0–10	0–10	0–10	
Mean change ± *SD*	0.2 ± 2.4*	− 1.8 ± 2.3*	− 2.5 ± 3.3*	− 14.2 ± 67.1
95% CI of mean change	− 0.3; 0.6	− 2.3; − 1.3	− 3.1; − 1.9	
Change (%)	6.3	26.9	28.1	
Responsiveness				
AUC	0.75	0.83	0.87	0.83
95% CI	0.63–0.88	0.74–0.93	0.76–0.98	0.76–0.90
Agreement with GPE (%)[a]	69	70	81	84
MCIC				
OCP	− 0.5	− 2.5	− 4.5	56%
Sensitivity (%); specificity (%)	59; 78	79; 70	90; 83	76; 86

*Significant change between T0 and T1 ($p < .001$)

[a] Agreement with GPE: percentage of patients with 1) a GPE 'improved' in combination with a clinically important change in score, or 2) a GPE 'improved' in combination with no clinically important change in score

PDI-W, pain disability index work item; n, number of patients; SD, standard deviation; min, minimum; max, maximum; CI, confidence interval; AUC, area under the curve; GPE; global perceived effect; MCIC, minimal clinically important change; OCP, optimal cut-off point

Only 20% of patients was classified as improved. Previous research stated that *little improved* patients can be added to the improved group [46]. However, other research stated that little improvement is in the range of natural fluctuation

Table 5 Responsiveness and Minimal Clinically Important Change of the WAS (*n* = 335) for baseline-score groups

	WAS			
	Baseline Tertile 1	Baseline Tertile 2	Baseline Tertile 3	Inverted percentage change score**
Patients, *n*	105	132	98	332
Improved patients, *n* (%)	17 (16.2)	29 (22.0)	20 (20.4)	66 (19.9)
Scores				
Score T0, mean ± *SD*	0.63 ± 0.85	4.6 ± 1.1	7.8 ± 0.78	
Min–max	0–2	3–6	7–10	
Score T1, mean ± *SD*	3.3 ± 3.0	5.5 ± 2.4	7.3 ± 1.9	
Min–max	0–10	0–10	0–10	
Mean change ± *SD*	2.6 ± 2.9*	0.93 ± 2.3*	−0.53 ± 1.8*	7.1 ± 62.4
95% CI of mean change	2.1; 3.2	0.54; 1.3	−0.17; −0.89	
Change (%)	412.7	20.2	6.8	
Responsiveness				
AUC	0.74	0.79	0.72	0.75
95% CI	0.58–0.89	0.69–0.88	0.59–0.86	0.68–0.82
Agreement with GPE (%)[a]	68	70	58	72
MCIC				
OCP	4.5	1.5	−0.5	39%
Sensitivity (%); specificity (%)	68; 73	66; 76	70; 65	63; 74

*Significant change between T0 and T1 ($p < .001$)

**Inverted percentage change score: (change score / (max score – baseline score)) × 100

[a] Agreement with GPE: percentage of patients with 1) a GPE 'improved' in combination with a clinically important change in score, or 2) a GPE 'improved' in combination with no clinically important change in score

WAS, work ability score; n, number of patients; SD, standard deviation; min, minimum; max, maximum; CI, confidence interval; AUC, area under the curve; GPE; global perceived effect; MCIC, minimal clinically important change; OCP, optimal cut-off point

[47]. When *little improved* patients are considered improved, accuracy to differentiate between improved and unimproved patients decreases [44]. In order to better reflect the concept of meaningful improvement, *little improved* patients were not classified as improved.

In addition, the patient sample was chosen based on relevance, because increasing work ability is not a treatment goal for all patients. Therefore, retired and permanently disabled patients, and stay at home parents were excluded. We included these patients in a sensitivity analysis to test the accuracy of the applied relevance criterion. Results reported the same MCICs for total group and baseline score groups as reported for the patient sample selected on relevance. Only the PDI-W percentage change score differed, which was 41% instead of 56%. The PDI-W takes unpaid work into account, including that of housework or volunteer. This is also carried out by the excluded patients. It is possible that small improvements in interference of pain with unpaid work are considered important, resulting in a lower percentage change score.

The effect of treatment should also be included in future research. For measurement of responsiveness, it is required to ensure that a proportion of patients is likely to change [37]. Not knowing for how many patients, when, what type of treatment takes place means that normally it would be difficult to predict whether a proportion of patients is likely to change within the utilized time interval of one year between baseline and follow-up. However, previous studies on patients from the GSC have shown that approximately a third shows clinically relevant improvement on measures of disability and impact of LBP one year after baseline measurement [16, 48]. Therefore, we expected a similar proportion of our patient sample to improve on work ability during follow-up.

Conclusion

Based on predefined hypotheses, construct validity of the WAS and PDI-W was not supported in this patient sample. The WAS and PDI-W are responsive to change in CLBP patients. Overall, an improvement of 1.5 point (WAS) and − 2.5 points (PDI-W) can be interpreted as clinically important change. However, MCIC values are baseline dependent. Clinically important change cannot be distinguished from measurement error, since MCICs were smaller than SDC values. Therefore, cautious interpretation is necessary for individual change scores up to 5 points.

Authors' contributions All authors contributed to the study conception and design. The literature search was performed by I Boekel, and data analysis was performed by I Boekel and AL Dutmer. The first draft of the manuscript was written by I Boekel, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding No funding was received for conducting this study.

Declarations

Conflicts of interest The authors have no relevant financial or non-financial interests to disclose.

Ethics approval The Medical Ethical Committee of the University Medical Center Groningen provided a waiver (M15.169472) for the data collection of the GSC with respect to medical ethical permission.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Informed consent Informed consent was obtained from all individual participants included in the study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Andersson GBJ (1997) The epidemiology of spinal disorders. In: Frymoyer JW (ed) The adult spine: principles and practice. Lippincott-Raven, Philadelphia, pp 93–141
2. Werner EL, Côté P (2009) Low back pain and determinants of sickness absence. Eur J Gen Pract 15(2):74–79. https://doi.org/10.1080/13814780903051866
3. van Tulder MW, Koes BW, Bouter LM (1995) A cost-of-illness study of back pain in The Netherlands. Pain 62(2):233–240. https://doi.org/10.1016/0304-3959(94)00272-G
4. Lambeek LC, van Tulder MW, Swinkels IC et al (2011) The trend in total cost of back pain in the Netherlands in the period 2002 to 2007. Spine (Phila Pa 1976) 36(13):1050–8. https://doi.org/10.1097/BRS.0b013e3181e70488
5. Collins JJ, Baase CM, Sharda CE et al (2005) The assessment of chronic health conditions on work performance, absence, and total economic impact for employers. J Occup and Environ Med 47(6):547–557. https://doi.org/10.1097/01.jom.0000166864.58664.29
6. Nordstoga AL, Vasseljen O, Meisingset I et al (2019) Improvement in work ability, psychological distress and pain sites in relation to low back pain prognosis: A longitudinal observational study in primary care. Spine (Phila Pa 1976) 44(7):E423-9. https://doi.org/10.1097/BRS.0000000000002860
7. Tuomi K, Ilmarinen J, Jahkola A, et al (1994) Respect for the aging: Work ability index. Helsinki: Institute of Occupational Health
8. Ahlstrom L, Grimby-Ekman A, Hagberg M et al (2010) The work ability index and single-item question: associations with sick leave, symptoms, and health - a prospective study of women on long-term sick leave. Scand J Work Environ Health 36(5):404–412. https://doi.org/10.5271/sjweh.2917
9. El Fassi M, Bocquet V, Majery N et al (2013) Work ability assessment in a worker population: comparison and determinants of work ability index and work ability score. BMJ Public Health 13:305. https://doi.org/10.1186/1471-2458-13-305
10. Stienstra M, Edelaar MJA, Fritz B et al (2021) Measurement properties of the work ability score in sick-listed workers with chronic musculoskeletal pain. J Occup Rehabil. https://doi.org/10.1007/s10926-021-09982-7
11. Soer R, Köke AJ, Vroomen PC et al (2013) Extensive validation of the pain disability index in 3 groups of patients with musculoskeletal pain. Spine (Phila Pa 1976) 38(9):E562-68. https://doi.org/10.1097/BRS.0b013e31828af21f
12. Kyte DG, Calvert M, van der Wees PJ et al (2015) An introduction to patient-reported outcome measures (PROMs) in physiotherapy. Physiotherapy 101(2):119–125. https://doi.org/10.1016/j.physio.2014.11.003
13. Mokkink LB, Terwee CB, Knol DL et al (2010) The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 10:22. https://doi.org/10.1186/1471-2288-10-22
14. Jaeschke R, Singer J, Guyatt GH (1989) Measurement of health status. Control Clin Trials 10(4):407–415. https://doi.org/10.1016/0197-2456(89)90005-6
15. Jette DU, Halbert J, Iverson C et al (2009) Use of standardized outcome measures in physical therapist practice: perceptions and applications. Phys Ther 89(2):125–135. https://doi.org/10.2522/ptj.20080234
16. Dutmer A, Schiphorst Preuper HR, Soer R et al (2019) Personal and societal impact of low back pain: the groningen spine cohort. Spine (Phila Pa 1976) 44(24):E1443-51. https://doi.org/10.1097/BRS.0000000000003174
17. Internationaal richtsnoer voor 'good clinical practice' voor het onderzoek met geneesmiddelen; vertaling naar de Nederlandse praktijk. rev version ed. Den Haag: GCP Begeleidingscommissie; 2003
18. EuroQol Group (1990) EuroQol-a new facility for the measurement of health-related quality of life. Health Policy 16(3):199–208. https://doi.org/10.1016/0168-8510(90)90421-9
19. Lamers LM, McDonnell J, Stalmeier PF et al (2006) The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ 15(10):1121–1132. https://doi.org/10.1002/hec.1124
20. Soer R, Reneman MF, Speijer BL et al (2012) Clinimetric properties of the EuroQol-5D in patients with chronic low back pain. Spine J 12(11):1035–1039. https://doi.org/10.1016/j.spinee.2012.10.030

21. Thorsen SV, Bjorner JB (2010) Reliability of the copenhagen psychosocial questionnaire. Scand J Public Health 38(Suppl 3):25–32. https://doi.org/10.1177/1403494809349859
22. Deyo RA, Dworkin SF, Amtmann D et al (2015) Report of the NIH task force on research standards for chronic low back pain. Phys Ther 95(2):E1-18. https://doi.org/10.2522/ptj.2015.95.2.e1
23. Boer A, Dutmer AL, Schiphorst Preuper HR et al (2017) Measurement properties of the NIH-minimal dataset dutch language version in patients. Spine (Phila Pa 1976) 42(19):1472–7. https://doi.org/10.1097/BRS.0000000000002134
24. Dutmer AL, Reneman MF, Schiphorst Preuper HR et al (2019) The NIH minimal dataset for chronic low back pain: responsiveness and minimal clinically important change. Spine (Phila Pa 1976) 44(20):E1211-8. https://doi.org/10.1097/BRS.0000000000003107
25. Kamper SJ, Ostelo RW, Knol DL et al (2010) Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. J Clin Epidemiol 63(7):760–766. https://doi.org/10.1016/j.jclinepi.2009.09.009
26. Pengel LH, Refshauge KM, Maher CG (2004) Responsiveness of pain, disability, and physical impairment outcomes in patients. Spine (Phila Pa 1976) 29(8):879–83. https://doi.org/10.1097/00007632-200404150-00011
27. Ostelo RW, de Vet HC (2005) Clinically important outcomes in low back pain. Best Pract Res Clin Rheumatol 19(4):593–607. https://doi.org/10.1016/j.berh.2005.03.003
28. McLeod LD, Coon CD, Martin SA et al (2011) Interpreting patient reported outcome results: US FDA guidance and emerging methods. Expert Rev Pharm Out 11(2):163–169. https://doi.org/10.1586/erp.11.12
29. Hazard Munro B (1986) Statistical methods for health care research. J. B. Lippincott, Philadelphia
30. Hoogendoorn WE, Bongers PM, de Vet HC et al (2002) High physical work load and low job satisfaction increase the risk of sickness absence due to low back pain: results of a prospective cohort study. Occup Environ Med 59(5):323–328. https://doi.org/10.1136/oem.59.5.323
31. Zirkzee EJ, Sneep AC, de Buck PD et al (2008) Sick leave and work disability in patients with early arthritis. Clin Rheumatol 27:11–19. https://doi.org/10.1007/s10067-007-0642-x
32. Schumacher H-L, Baumann N (2003) Temporary working disability among younger patients with back pain and its interaction with severity of pain, functional capacity, social factors and personality styles. Rehabilitation 42(4):245–252. https://doi.org/10.1055/s-2003-41651
33. Ledbe N, Burns J, Mackey M et al (2020) Normative reference values and physical factors associated with work ability: a cross-sectional observational study. Occup Environ Med 77(4):231–237. https://doi.org/10.1136/oemed-2019-106248
34. Hansson E, Hansson T, Jonsson R (2006) Predictors for work ability and disability in men and women with low-back or neck problems. Eur Spine J 15:780–793. https://doi.org/10.1007/s00586-004-0863-5
35. Steenstra IA, Munhall C, Irvin E et al (2017) Systematic review of prognostic factors for return to work in workers with sub acute and chronic low back pain. J Occup Rehab 27(3):369–381. https://doi.org/10.1007/s10926-016-9666-x
36. Terwee CB, Bot SD, de Boer MR et al (2007) Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60(1):34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012
37. Mokkink LB, Prinsen CAC, Patrick DL, et al (2019) COSMIN Study Design checklist for Patient-reported outcome measurement instruments. Retrieved 2021 from https://www.cosmin.nl
38. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36. https://doi.org/10.1148/radiology.143.1.7063747
39. Froud Robert, Abel Gary (2014) Using ROC curves to choose minimally important change thresholds when sensitivity and specificity are valued equally: the forgotten lesson of Pythagoras. Theoretical considerations and an example application of change in health status. PLoS ONE 9(12):e114468. https://doi.org/10.1371/journal.pone.0114468
40. de Vet HC, Foumani M, Scholten MA et al (2015) Minimally important change values of a measurement instrument depend more on baseline values than on the type of intervention. J Clin Epidemiol 68(5):518–524. https://doi.org/10.1016/j.jclinepi.2014.07.008
41. Terwee CB, Roorda LD, Knol DL et al (2009) Linking measurement error to minimal important change of patient-reported outcomes. J Clin Epidemiol 62(10):1062–1067. https://doi.org/10.1016/j.jclinepi.2008.10.011
42. van Kampen DA, Willems WJ, van Beers LW et al (2013) Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). J Orthop Surg Res 8:40. https://doi.org/10.1186/1749-799X-8-40
43. Demoulin C, Ostelo R, Knottnerus JA et al (2010) Quebec Back Pain Disability Scale was responsive and showed reasonable interpretability after a multidisciplinary treatment. J Clin Epidemiol 63(11):1249–1255. https://doi.org/10.1016/j.jclinepi.2009.08.029
44. de Vet HC, Ostelo RW, Terwee CB et al (2007) Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual Life Res 16(1):131–142. https://doi.org/10.1007/s11136-006-9109-9
45. Roelen CAM, van Rhenen W, Groothoff JW et al (2014) Work ability as prognostic marker of disability pension: single-item work ability score versus multi-item work ability index. Scand J Work Environ Health 40(4):428–431. https://doi.org/10.5271/sjweh.3428
46. Terwee CB, Roorda LD, Dekker J et al (2010) Mind the MIC: large variation among population and methods. J Clin Epidemiol 63(5):524–534. https://doi.org/10.1016/j.jclinepi.2009.08.010
47. Ostelo RW, Deyo RA, Stratford P et al (2008) Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. Spine (Phila Pa 1976) 33(1):90–4. https://doi.org/10.1097/BRS.0b013e31815e3a10
48. Dutmer AL, Schiphorst Preuper HR, Stewart RE et al (2020) Trajectories of disability and low back pain impact. Spine 45(23):1649–1660. https://doi.org/10.1097/BRS.0000000000003647

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.