



# Measurement Properties of the Work Ability Score in Sick-Listed Workers with Chronic Musculoskeletal Pain

M. Stienstra<sup>1</sup> · M. J. A. Edelaar<sup>1</sup> · B. Fritz<sup>1</sup> · M. F. Reneman<sup>2</sup>

Accepted: 26 April 2021 / Published online: 26 May 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

## Abstract

**Purpose** Chronic musculoskeletal pain can have a major impact on ability to work. The work ability score is a commonly used single-item question to assess work ability but has not been fully validated yet. The aim of the present study was to evaluate test–retest reliability, agreement, construct validity, and responsiveness of the work ability score among sick-listed workers with chronic musculoskeletal pain. **Methods** Data of sick-listed workers with chronic musculoskeletal pain was routinely collected at seven rehabilitation centres in the Netherlands. Assessments included a set of questionnaires, administered at admission and discharge from a fifteen-week vocational rehabilitation program. Test–retest reliability was determined with the intraclass correlation coefficient. For agreement, the standard error of measurement and smallest detectable changes were calculated. Construct validity was assessed by testing hypotheses regarding Spearman rank correlation coefficient. Area under the curve obtained from the receiver operating characteristic curve and minimal clinically important change were determined for the total sample and work ability score baseline tertile groups to assess responsiveness. **Results** In total, 34 workers were analyzed for reliability and agreement, 1291 workers for construct validity, and 590 responded to the responsiveness questionnaire. Reliability reached an intraclass correlation coefficient of 0.89; 95% CI 0.77–0.94, a standard error of measurement of 0.69 points, and the smallest detectable change of 1.92 points. For construct validity, six of the seven predefined hypotheses were not refuted. The area under the curve was 0.76 (95% CI 0.71–0.81) allowing for discrimination between stable and improved workers, with a minimal clinically important change of 2.0 points for the total sample. **Conclusion** The work ability score showed good measurement properties among sick-listed workers with chronic musculoskeletal pain.

**Keywords** Chronic pain · Vocational rehabilitation · Self-reported outcome · Reproducibility · Responsiveness

## Introduction

Chronic musculoskeletal pain (CMP) affects approximately 20% of the adult European population [1, 2]. Pain is considered chronic when it persists for three months or longer [3]. Because CMP can impact work ability (WA), CMP is a major reason for reduced work participation [4, 5]. WA is defined as the ability of workers to do their work according

to the demands of the job contextual to their health and mental resources [6]. It is a comprehensive concept composed of different aspects that are presented as the ‘house of WA’. The foundation of this model is the health aspect that consists of the amalgam of mental and physical health and social functioning [7]. To measure WA from the perspective of the worker, self-reported outcome measures are used. Self-reported outcome measures need to have adequate measurement properties to justify their use in the clinic or in research [8].

The work ability index (WAI) is worldwide the most commonly used WA questionnaire in occupational health care, clinical practice, and research [9]. This questionnaire correlates moderate to strong with self-rated general health questionnaires and is therefore considered as a valid instrument to estimate WA among healthy workers ( $r=0.44–0.79$ ) [10–12]. The WAI is a 10-item questionnaire that has been translated and validated into several languages, including

✉ M. Stienstra  
Manouk.Stienstra@outlook.com

<sup>1</sup> Department of Research and Development, and Institute of Vocational Assessment and Education, Heliomare Rehabilitation Center, Relweg 51, 1949, EC Wijk aan Zee, The Netherlands

<sup>2</sup> Department of Rehabilitation Medicine, Center for Rehabilitation, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Dutch [10, 13–16]. The first question of the WAI (“current WA compared with lifetime best WA”), is also known as the work ability score (WAS). This single item was strongly related to the total WAI for assessing the current level and progression of WA among general workers and those who are on long-term sick leave ( $R_s = 0.63–0.87$ ) [17–19]. Because of its brevity, the WAS may be a good alternative for the WAI in research and clinically useful for routine evaluation and interpretation of patient outcomes [20].

Despite the widespread use of the WAS, its test–retest reliability, agreement, construct validity, and responsiveness has not been studied in sick-listed workers with CMP. The research questions for the present study were:

1. What is the test–retest reliability and agreement of the WAS in sick-listed workers with CMP?
2. Is the construct validity of the WAS adequate in sick-listed workers with CMP?
3. What is the responsiveness and minimal clinically important change of the WAS in sick-listed workers with CMP?

## Methods

The Consensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist was applied when designing this study [8, 21, 22].

### Study Design and Ethics

A retrospective observational cohort study was used to evaluate the measurement properties of the WAS. Data were derived from electronic health records from seven vocational rehabilitation (VR) centres in the Netherlands, collected between November 2014 and October 2019. For the construct validity, a cross-sectional study design was used and a repeated-measurement design with a fifteen-week interval was used to enable test–retest reliability, agreement, and responsiveness.

In the Netherlands, no permission is required from a medical ethics committee for the evaluation of outcomes of care solely based on anonymous data derived from the medical records. All data security and privacy regulations were adhered to. Informed consent was obtained from all workers being included in the study.

### Study Sample

The study samples consisted of sick-listed workers with CMP admitted to a fifteen-week multidisciplinary VR program, provided in one of the participating VR centres. The program involved an individualized exercise program,

cognitive behaviour therapy, group education, relaxation, and work-related guidance, delivered by a team of healthcare providers [23]. The inclusion criteria for the program were: being of working age (18–65 years), suffering from subacute (6–12 weeks) or chronic (> 12 weeks) musculoskeletal pain, and having decreased work participation (part-time or full-time sick leave or reduced work productivity) [23]. When essential baseline or discharge data was missing, data were excluded from analysis. Workers were excluded from this study if they have comorbidities other than CMP as a primary reason for sick leave or if they have no paid work.

### Procedure

At baseline, before the start of the VR program, personal characteristics were collected and all workers filled out a set of questionnaires, as part of usual care [24]. Questionnaires were sent by mail to be completed individually by the workers at home. The workers received the set of questionnaires at discharge from the VR program for the second time and also completed the global perceived effect scale.

### Measurements

#### Personal Characteristics

The personal characteristics collected in this study were: age (years), sex (male, female), educational level (low, medium, high), work status at baseline (full-time, part-time, 100% sick leave), extent of contract (hours/days), number of pain locations, and duration of pain (months, years).

#### Work Ability Score (WAS)

WA was assessed using the WAS, which is the first item of the WAI: ‘What is your current WA compared to your lifetime best WA?’ The question is scored on an 11-point Likert scale, where 0 represents ‘completely unable to work’ and 10 represents ‘WA perceived as lifetime best’. WAS and WAI are strongly related and are good indicators of WA [17].

#### iMTA Productivity Cost Questionnaire (iPCQ)

Work productivity is determined by the worker’s presence and performance at work. The first phenomenon is known as sickness absence, while the second phenomenon is called presenteeism [25]. Sickness absence and presenteeism were assessed with the iMTA Productivity Cost Questionnaire (iPCQ). Long-term sickness absence related to the reason for which workers came to VR was reported as the number of calendar days between the date of reporting going on sick leave and date of filling out baseline questionnaires. For the workers on short term sick leave, the number of days on sick

leave in the past 4 weeks was reported. The presenteeism score from workers who were partly or completely at work and experienced presenteeism was used. The score ranges from 0 (I couldn't do anything) to 10 (I could do the same as normal) [26].

### Pain Disability Index (PDI)

Self-reported disability related to pain was assessed using the PDI. This questionnaire covers seven areas of activities and participation: family and home responsibilities, recreation, social activity, occupation, sexual behaviour, self-care, and life-support activity. Each area has one question, which is scored on an 11-point rating scale where 0 means no disability and 10 represents maximum disability. The total score ranges from 0 to 70 points, with higher scores indicating more disability [27].

### RAND-36 Physical Functioning

Physical functioning was assessed using the physical functioning scale of the RAND-36 Health Survey. This scale consists of 10 questions with three levels of response ('yes, strongly limited', 'yes, a bit limited', and 'no, not limited'). The total score ranges from 0 to 100, with higher scores indicating better physical functioning [28, 29].

### EuroQol 5D (EQ-5D)

Health-related quality of life was assessed using the first part of the EQ-5D. This part covers five dimensions of health: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has one question with three levels of response (no problems, some problems, and severe problems). Answers can be transformed into an index score ranging from 0 to 1, with higher scores indicating better overall health [30].

### Numeric Pain Rating Scale (NPRS)

The pain intensity was assessed using the NPRS, requiring workers to rate their average and worst level of pain over the past seven days. The questions were scored on an 11-point rating scale, where 0 referred to no pain and 10 to worst possible pain [31].

### Global Perceived Effect (GPE)

Evaluation of how much the rehabilitation program changed the work functioning of the worker compared with pre-rehabilitation level was assessed using one item of the global perceived effect (GPE) ('How much did the VR program change

your work functioning compared to pre-treatment?'). The question was scored on a 7-point Likert scale, ranging from 1 (extremely worsened) to 7 (completely improved), while 4 is unchanged [32]. Based on the GPE score the workers were classified as improved (score 5–7), stable (score 4) or worsened (score 1–3). This instrument was used as the anchor (external criterion) in the responsiveness analysis, to compare the changes over time as derived from the WAS.

### Statistical Analysis

Data were analyzed with SPSS version 24.0 statistical software for Windows (SPSS Inc., Chicago, IL, USA). For baseline characteristics, the distribution was assessed by skewness, kurtosis, and histograms. A skewness or kurtosis between -1.0 and 1.0 indicated a normal distribution for large sample sizes (> 300 participants). For smaller sample sizes a z-score less than 1.96 is accepted for a normal distribution [33]. Mean value and standard deviation (SD) were presented for continuous normally distributed data, and median and interquartile range (IQR) were used for non-normally distributed data.  $p < 0.05$  was interpreted as statistical significance for all analyses.

### Reliability

The test–retest reliability was derived from workers who were stable over a fifteen-week period. Stability was defined based on four criteria. 1. The GPE score was 4. 2. The change on question number 4 of the PDI was not greater than 1 point from baseline till discharge. 3. The change on the presenteeism score was not greater than 1 point for workers who were partly or completely at work and experienced presenteeism at baseline. 4. The difference on short term sick leave was less than 5.5 days and workers who were on long-term sick leave at baseline were stable on this criterium if they were still with 100% sick leave at discharge [27, 34]. The intraclass correlation coefficient (ICC) was calculated to assess test–retest reliability, based on a 2-way mixed-effects model for absolute agreement, with corresponding 95% confidence intervals (CI) [35]. An  $ICC \geq 0.7$  with a value  $> 0.5$  for the lower bound of the 95% CI is generally considered to be acceptable for test–retest reliability [35, 36].

### Agreement

The standard error of measurement (SEM) was calculated to assess the absolute amount of measurement error ( $SEM = SD\sqrt{1 - ICC}$ ), where SD is the standard deviation of the WAS scores obtained from all workers and ICC is the test–retest reliability coefficient. The SEM was also used to determine the smallest detectable change (SDC) for an individual ( $SDC_{individual} = SEM \times 1.96 \times \sqrt{2}$ ) and the total

sample ( $SDC_{sample} = \frac{SDC_{individual}}{\sqrt{n}}$ ). The agreement of the WAS is considered as good if the absolute measurement error is smaller than the minimal clinically important change [36, 37].

### Floor and Ceiling Effects

The presence of significant floor and ceiling effects were considered if more than 15% of the workers from the construct validity sample achieved the minimum (0) or maximum (10) possible WAS score at baseline [36].

### Construct Validity

The datasets which contain the complete required baseline measurements from the medical records were used for construct validity analysis. WAS construct validity was examined based on seven hypotheses. Spearman rank correlation coefficient ( $\rho$ ) was used to measure associations. The construct validity was considered sufficient if at least 75% of the predefined hypotheses were not refuted [36].

1. WAS correlates moderately ( $r > 0.5$ ) with the work productivity measures. Work productivity is the result of the workers' capacities and abilities, thus both instruments are related to the assessment of a worker's capability to carry out work [38–41].
  - 1.1 WAS correlates moderately negative ( $r > -0.5$ ) with days of sickness absence.
  - 1.2. WAS correlates moderately positive ( $r > 0.5$ ) with the presenteeism score.
2. WAS correlates weakly to moderate negative ( $-0.2 < r < -0.5$ ) with the total PDI score.
  - 2.1 WAS correlates moderately negative ( $r > -0.5$ ) with question number 4 of the PDI. Question 4 is the specific work-related question, which captures most specific the construct of WA.
3. WAS correlates weakly to moderate positive ( $0.2 < r < 0.5$ ) with the RAND-36 physical functioning. The instrument measures the three primary domains of physical health that are key components to consider when evaluating physical functioning in the context of work [42]. Physical functioning is part of the foundation and relevant for WA and daily life functioning but is mainly related to WA in workers with high physical demands [43].

4. WAS correlates weakly to moderate positive ( $0.2 < r < 0.5$ ) with the EQ-5D. Quality of life is a generic dimension of health, which is less directly related to WA.
5. WAS correlates weakly negative ( $r < -0.3$ ) with the NPRS because pain is a comprehensive multidimensional construct that possibly represents only a fraction of WA.

### Responsiveness

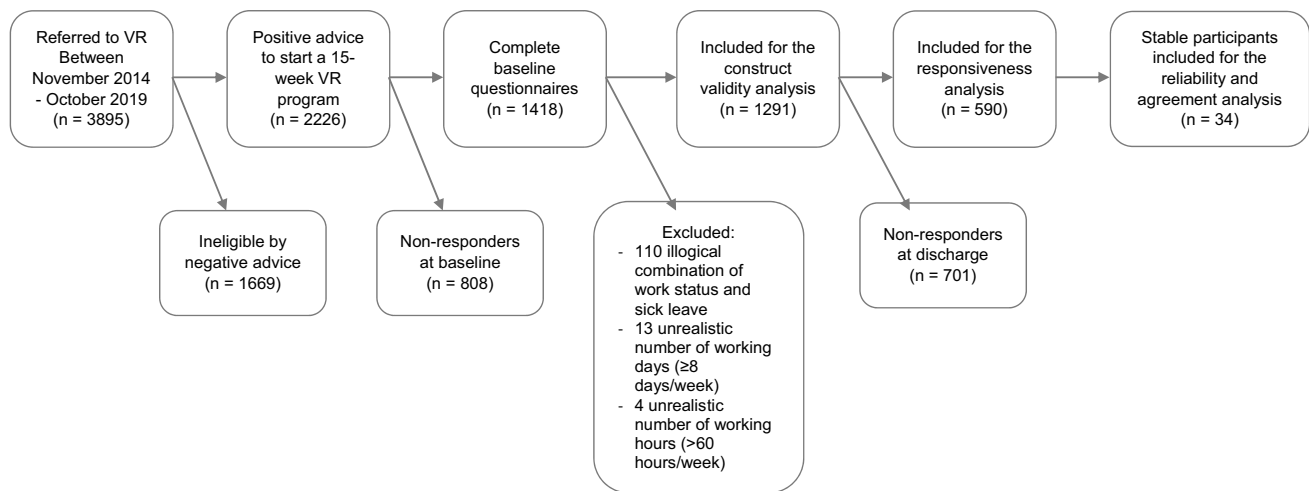
The GPE was used to classify workers as 'improved', 'stable' or 'worsened'. The group 'worsened' was not included in the analysis. Based on this classification, a receiver operating characteristic (ROC) curve of the absolute change score was created by plotting the false positive rate (1-specificity) against the true positive rate (sensitivity). The minimal clinically important change was determined by the optimal cut-off point of the ROC curve of the change scores [44]. Additional responsiveness analyses were performed for the change scores in which the total sample was stratified by the baseline WAS tertile scores, because minimal clinically important change is likely to be influenced by baseline scores [45]. Area under the curve and 95% CI were used for describing the ability of the WAS to distinguish improved workers from not improved workers. Area under the curve  $> 0.9$  indicates excellent discrimination, good discrimination by  $0.7-0.9$ , moderate discrimination by  $0.5-0.7$ , and discrimination fails if area under the curve  $\leq 0.5$  [46].

### Sample Size

According to the COSMIN checklist, a sample size of 50–99 participants is considered adequate to obtain reasonable results for determining test–retest reliability, agreement, validity, and responsiveness. Furthermore, a sample size of  $\geq 100$  participants is assessed as excellent [22].

### Results

A total of 34 workers were eligible for the reliability and agreement analysis because they met the four criteria of the operational definition for being considered stable. In total, 1291 eligible workers filled out the complete baseline questionnaires and were available for the construct validity analysis. The baseline and discharge responsiveness questionnaires were completed by 590 workers. WAS at baseline was not significantly different between responders and non-responders ( $p = 0.413$ ). A flowchart of the inclusion of workers is shown in Fig. 1. The baseline characteristics of the workers for the study samples are shown in Table 1.



**Fig. 1** Flowchart of recruitment, evaluation, and exclusion

## Reliability and Agreement

Test–retest reliability was  $ICC = 0.89$  (95% CI 0.77–0.94), mean WAS score for test and retest were respectively 2.9 (SD 2.1) and 3.1 (SD 2.2),  $p = 0.386$ . The SEM, SDC individual, and SDC for the total sample were respectively 0.69, 1.92, and 0.33.

## Floor and Ceiling Effects

At baseline, 8.0% of the workers scored 0 and 1.4% of the workers scored 10 (valid  $n = 1291$ ). The percentages did not exceed 15%, therefore significant floor and ceiling effects were not present.

## Construct Validity

Results of the construct validity are shown in Table 2. Six of the seven (85,7%) predefined hypotheses on the magnitude of the relationship between WA and the other constructs were supported. The correlation between WAS and the number of sickness absence days was refuted, the observed correlation was slightly weaker than hypothesized.

## Responsiveness

Based on the GPE classification, 48 out of the 590 workers worsened, 117 workers were stable and 425 workers improved. WAS at baseline did not significantly differ between the stable and improved group ( $p = 0.120$ ).

The mean scores, area under the curve, minimal clinically important change, sensitivity, and specificity of the WAS for the total sample and the baseline tertiles are presented

in Table 3 and the ROC curves are shown in Fig. 2. The discriminative ability in the total sample between the stable and improved group was an area under the curve of 0.76 (95% CI 0.71–0.81), with a corresponding minimal clinically important change of 2.0 points.

## Discussion

This is the first study that assesses the measurement properties of the WAS in sick-listed workers with CMP. The test–retest reliability analysis resulted in an  $ICC = 0.89$  which is considered adequate. Floor and ceiling effects were not present. Six of the seven predefined hypotheses were not refuted, supporting the construct validity of the WAS, the minimal clinically important change for the total sample was 2.0 points with a good discriminative ability. In summary, the WAS demonstrated acceptable reliability, construct validity, and responsiveness in this study sample.

The test–retest reliability in the present study was similar to a study among Iranian workers ( $ICC = 0.83$ ) [47] and comparable with the total WAI among healthy nurses and healthcare workers ( $ICC = 0.92$ ) [12]. Direct comparison is difficult because of differences in study samples (healthy versus CMP), and the time interval between test and retest.

As expected, the strongest correlation with the WAS was seen between the presenteeism score ( $r = 0.64$ ), followed by PDI question 4 ('How would you rate the level of disability you typically experience during occupational activities?') ( $r = -0.52$ ) and sickness absence ( $r = -0.40$ ), indicating that these measurement instruments were best related to the construct of WA. The correlation indicates that when the presenteeism score decreased, or when the score on PDI question 4 or days of sickness absence increased, perceived WA

**Table 1** Baseline characteristics of the sick-listed workers for different study samples

	Reliability and agreement n=34	Construct validity n=1291	Responsiveness n=590
Age (years)	50.9 ± 9.8	45.9 ± 10.6	46.4 ± 10.6
Sex (female)	22 (64.7%)	795 (61.6%)	356 (60.3%)
Educational level			
Low	9 (26.5%)	300 (23.2%)	135 (22.9%)
Medium	18 (52.9%)	600 (46.5%)	264 (44.7%)
High	7 (20.6%)	391 (30.3%)	191 (32.4%)
Work contract			
Hours/week	31.0 (22.4–38.5)	31.1 ± 8.5	30.8 ± 8.9
Days/week	4.2 ± 1.0	4.2 ± 1.0	4.2 ± 1.0
Work status			
Working full-time	2 (5.9%)	151 (11.7%)	56 (9.5%)
Working part-time	14 (41.2%)	624 (48.3%)	299 (50.7%)
100% sick leave	18 (52.9%)	516 (40.0%)	235 (39.8%)
Sick leave short (yes)	5 (14.7%)	167 (12.9%)	86 (14.6%)
Sick leave long (yes)	24 (70.6%)	742 (57.5%)	337 (57.1%)
Presenteeism (yes)	15 (44.1%)	689 (53.4%)	311 (52.7%)
Presenteeism score <sup>a</sup>	4.9 ± 2.1	5.5 ± 2.2	5.5 ± 2.2
Headache/burnout (yes)	18 (52.9%)	535 (41.4%)	244 (41.4%)
Number of pain locations	2 (1–4)	2 (1–4)	2 (1–4)
Pain duration			
1–3 months	–	72 (5.6%)	33 (5.6%)
3–6 months	6 (17.6%)	213 (16.5%)	101 (17.1%)
6–12 months	7 (20.6%)	326 (25.3%)	157 (26.6%)
1–2 years	5 (14.7%)	275 (21.3%)	134 (22.7%)
2–5 years	8 (23.5%)	198 (15.3%)	73 (12.4%)
More than 5 years	8 (23.5%)	207 (16.0%)	92 (15.6%)
NPRS <sup>b</sup>			
Average level of pain	6.1 ± 1.4	5.4 ± 2.3	5.3 ± 2.3
Worst level of pain	8 (7–8)	8 (6–8)	8 (6–8)
WAS <sup>c</sup>	2.9 ± 2.1	3.8 ± 2.4	3.7 ± 2.3
PDI <sup>d</sup>	38.0 ± 10.7	34.8 ± 12.2	34.5 ± 12.0
RAND-36 physical functioning <sup>e</sup>	55.6 ± 20.9	58.6 ± 20.1	59.9 ± 19.9
EQ-5D <sup>f</sup>	0.67 (0.22–0.78)	0.59 ± 0.25	0.62 ± 0.23

Depending on the type of data and the distribution, variables are presented as numbers (%), median (interquartile range) or mean ± standard deviation

EQ-5D = EuroQol 5D; n = number; NPRS = numeric pain rating scale; PDI = pain disability index; WAS = work ability score

<sup>a</sup>0 = couldn't do anything, 10 = could do the same as normal

<sup>b</sup>0 = no pain, 10 = worst possible pain

<sup>c</sup>0 = unable to work, 10 = lifetime best work ability

<sup>d</sup>0 = no disability, 70 = maximum disability

<sup>e</sup>0 = worst state of physical functioning, 100 = best state of physical functioning

<sup>f</sup>0 = worst overall health, 1 = best overall health

decreased. The correlation between the WAS and sickness absence was weaker than expected. Stronger correlations were present within healthy samples ( $r = -0.44$  to  $-0.62$ ) [39], indicating that days of sickness absence capture WA better among samples of healthy individuals than among

those with CMP and a relatively high rate of long-term sick leave (57.5%). The correlation between the WAS and presenteeism score was comparable with the result of another study ( $r = 0.69$ ) [40], supporting validity for the WAS. Construct validity of the WAI was better supported by physical

**Table 2** Hypothesized and observed Spearman rank correlation coefficient between the baseline work ability score (WAS) and other measurement instruments

Construct	Measure	n	Hypothesis	Observed correlation	p-value	Hypothesis refuted
Work-related	Days of sickness absence	1160	$r > -0.5$	-0.40	<0.01	Yes
	Presenteeism score	689	$r > 0.5$	0.64	<0.01	No
	PDI Question 4	1291	$r > -0.5$	-0.52	<0.01	No
Physical functioning / disability	Total PDI score	1291	$-0.2 < r < -0.5$	-0.35	<0.01	No
	RAND-36 physical functioning	1291	$0.2 < r < 0.5$	0.22	<0.01	No
Quality of life	EQ-5D	1291	$0.2 < r < 0.5$	0.26	<0.01	No
Pain intensity	NPRS average	1291	$r < -0.3$	-0.05	<0.01	No
	NPRS worst	1291	$r < -0.3$	-0.11	<0.01	No

*EQ-5D* EuroQol 5D, *n* number, *NPRS* numeric pain rating scale, *PDI* pain disability index, *r* Spearman's rho, *WAS* work ability score

**Table 3** Mean baseline and change scores, standard deviations, and responsiveness of the work ability score (WAS)

	Baseline Mean $\pm$ SD	Change score Mean $\pm$ SD	AUC (95% CI)	MCIC	Sensitivity	Specificity
Total sample						
Stable (n = 117)	3.4 $\pm$ 2.4	0.2 $\pm$ 2.1	0.76 (0.71–0.81)	2.0	0.664	0.778
Improved (n = 425)	3.8 $\pm$ 2.3	2.4 $\pm$ 2.4				
Baseline WAS tertile 1 (score 0–2)						
Stable (n = 48)	1.2 $\pm$ 0.7	0.9 $\pm$ 1.6	0.90 (0.85–0.95)	3.0	0.811	0.833
Improved (n = 132)	1.2 $\pm$ 0.8	4.3 $\pm$ 2.0				
Baseline WAS tertile 2 (score 3–4)						
Stable (n = 34)	3.5 $\pm$ 0.6	0.2 $\pm$ 1.9	0.85 (0.78–0.93)	2.0	0.842	0.765
Improved (n = 133)	3.5 $\pm$ 0.5	2.8 $\pm$ 1.6				
Baseline WAS tertile 3 (score 5–10)						
Stable (n = 35)	6.3 $\pm$ 1.7	-0.8 $\pm$ 2.5	0.68 (0.58–0.77)	1.0	0.531	0.771
Improved (n = 160)	6.2 $\pm$ 1.3	0.5 $\pm$ 1.8				

*AUC* area under the curve, *CI* Confidence interval, *MCIC* minimal clinically important change, *n* number, *SD* standard deviation, *WAS* work ability score

functioning ( $r = 0.38$ – $0.40$ ) [14, 16] compared with the result for the WAS in the present study ( $r = 0.22$ ). This difference could be explained because the WAI is a more comprehensive measurement instrument, and previous studies included workers who primarily worked in physically demanding jobs influencing perceived WA and physical functioning [9, 48].

In the present study, the area under the curve of 0.76 provides evidence that the WAS is a responsive instrument for detecting clinically relevant changes at discharge from VR. The discriminative ability of the WAS was good within the first (area under the curve = 0.90) and second (area under the curve = 0.85) tertile, and moderate for the third tertile (area under the curve = 0.68). Responsiveness of the WAS has to our knowledge not previously been assessed in another study, therefore the results cannot be compared.

The results of this study support the WAS as a valid, reliable, and responsive instrument. Consequently, the WAS is

suitable for WA assessments at a group and at an individual level consisting of workers with heterogeneity concerning work types, and suitable for monitoring progress in VR. To decide whether an improvement in the WAS is clinically important and is not due to measurement error, minimal clinically important change values should be interpreted in relation to the SDC. The results of this study indicate a total sample SDC of 1.92 and an anchor-based minimal clinically important change of 2.0 points for the WAS of the total sample. The minimal clinically important change for the first and second tertile was respectively 3 and 2 points. Because the minimal clinically important change is larger than the SDC, the minimal clinically important change should be used as the cut-off value. In contrast, the minimal clinically important change of the third tertile was only 1 point and cannot be distinguished from the measurement error, therefore the SDC should be used as the cut-off value.

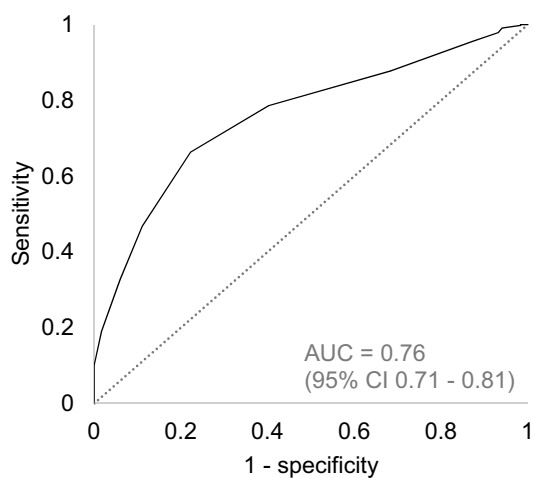


Figure 2a

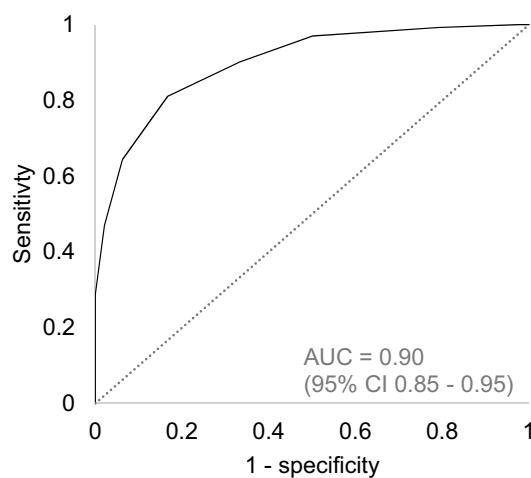


Figure 2b

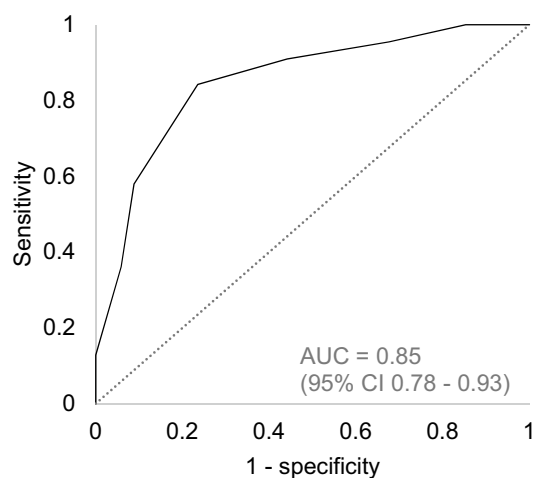


Figure 2c

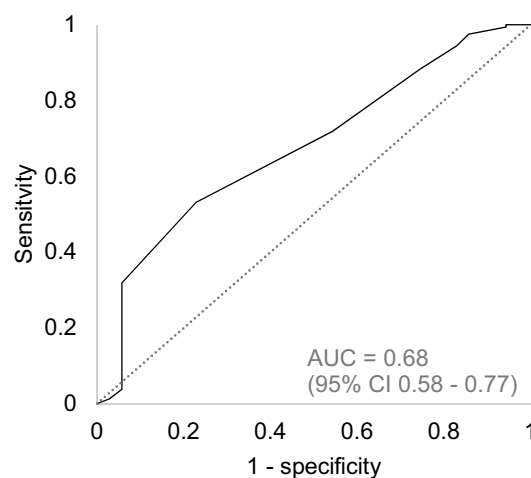


Figure 2d

**Fig. 2** Receiver operating characteristics (ROC) curves of the work ability score (WAS). **a** ROC-curve of the total study sample (n=542). **b** ROC-curve of the sample with baseline WAS tertile 1 score (n=180). **c** ROC-curve of the sample with baseline WAS tertile 2

score (n=167). **d** ROC-curve of the sample with baseline WAS tertile 3 score (n=195). AUC=area under the curve; CI=confidence interval

By the interpretation of changes for an individual, it is recommended to account for baseline scores to avoid misclassification bias [49]. Sick-listed workers with a baseline score of  $\leq 2$  (first tertile) should increase minimal 3.0 points and sick-listed workers with a baseline score  $\geq 3$  (second and third tertile) should increase minimal 2.0 points to conclude that a relevant and measurable change has occurred. This information should be useful for the clinicians in the VR setting and researchers using the WAS as an outcome measure to help determine whether a clinically meaningful change has occurred as a consequence of the VR program.

### Strengths and Limitations

A general strength of this study was the use of data from usual care collected in seven different VR centres in the Netherlands. Because the study was performed in a setting and sample that is representative of the daily clinical practice, the results are broader generalizable. There was also a sufficient sample size for the construct validity and responsiveness analysis. The study sample consists of sick-listed workers with different CMP complaints, a broad range of work professions, working hours, educational level, sex, and



age. This makes the WAS suitable for a wide population of workers with CMP.

Despite the strengths, the present study does have some limitations that primarily impacts reliability and agreement. First, a traditional test–retest design could not be used, because there was no earlier measurement moment at which WAS had been measured. The time between test and retest assessment was 15 weeks during which a VR program was followed. To ensure that the sample was stable between the two assessment moments, which is a prerequisite for test–retest analyses, a strict operational definition was used. This strict operational definition resulted in a sample of  $n=34$ , which is lower than recommended ( $n=50$ ) [22]. To investigate the extent to which the operational definition and small sample size affected the results, post hoc two sensitivity analysis were conducted. For the first sensitivity analysis the new threshold for the PDI question 4 change score was 2 points ( $n=53$ ), which equals the minimal clinically important change determined from data in this study. The results of these analyses are  $ICC=0.83$  (95% 0.70–0.90),  $SEM=0.82$ ,  $SDC_{individual}=2.27$  and  $SDC_{group}=0.31$ . In the second analysis, besides the extension of the PDI change score, the GPE score was broadened to include 3 and 5 ( $n=111$ ). The results of the second analysis were respectively  $ICC=0.76$  (95% 0.65–0.83),  $SEM=1.04$ ,  $SDC_{individual}=2.88$  and  $SDC_{group}=0.27$ . By loosening the operational definition of stability, a slightly less stable sample was created, resulting in larger samples and a slight decrease of the ICC, and an increase in the SEM. Given these results, it is unlikely that the strict definition of stable affected the test–retest reliability and agreement of the WAS.

The second limitation of the present study is the potential selection bias in the reliability and agreement sample. The sample was not completely representative of the total study sample. The included workers are on average older, a longer duration of pain, higher pain scores, and worse WAS and PDI scores. A further limitation is that the results of the current study are limited to the care as usual population in the Dutch VR setting. Future research should reveal whether these findings can be replicated and generalized to other samples.

## Conclusion

The current study provides support for use of the WAS for assessing and evaluating the WA in workers with CMP in vocational rehabilitation. Apart from adequate measurement properties, it is easy to administer, simple to interpret, and not time-consuming for the worker to complete. An group change of 2.0 points, and a change score of 3.0 and 2.0 points for individuals with a baseline score  $\leq 2.0$  and  $\geq 3.0$

respectively can be used for evaluation purposes to assess the effectiveness of the VR program.

**Acknowledgements** The authors thank the following vocational rehabilitation centers for their help with the data collection: Adelante (Hoensbroek), Heliomare (Wijk aan Zee), Klimmendaal – Zintens (Arnhem), Libra revalidatie en audiologie (Eindhoven & Tilburg), Merem medische revalidatie (Almere), Roessingh, centrum voor revalidatie (Enschede), and UMCG centrum voor revalidatie, Beatrixoord (Groningen). The following researchers and clinicians have actively contributed to the research and agreed to be acknowledged for their contribution: Timo Beemster, Judith van Velzen, Coen van Bennekom, Janneke Nachtegaal, Lucienne van der Meer, Marlique Langkamp, Noelle Muller, Tristan Erens, Margaret Lam, Levijn Romp, Saskia Koning, Clement Roos, Karin Karstens, Sanny Rakers, Simone Akkerman, Gabrielle Rovers, Sabrine du Bois, Barbara Strang and Arne Martens.

## Declarations

**Conflict of interest** All authors declare that they have no conflict of interest.

**Ethical Approval** This retrospective study involving human participants was in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki Declarations and its later amendments or comparable ethical standards.

**Consent to Participate** Informed consent was obtained from all workers being included in the study.

## References

1. Cimmino MA, Ferrone C, Cutolo M. Epidemiology of chronic musculoskeletal pain. *Best Pract Res Clin Rheumatol.* 2011;25(2):173–83.
2. van Hecke O, Torrance N, Smith BH. Chronic pain epidemiology and its clinical relevance. *Br J Anaesth.* 2013;111(1):13–8.
3. Treede RD, Rief W, Barke A, Aziz Q, Bennett MI, Benoliel R, et al. Chronic pain as a symptom or a disease: the IASP Classification of Chronic Pain for the International Classification of Diseases (ICD-11). *Pain.* 2019;160(1):19–27.
4. Bevan S. Economic impact of musculoskeletal disorders (MSDs) on work in Europe. *Best Pract Res Clin Rheumatol.* 2015;29(3):356–73.
5. Mather L, Ropponen A, Mittendorfer-Rutz E, Narusyte J, Svedberg P. Health, work and demographic factors associated with a lower risk of work disability and unemployment in employees with lower back, neck and shoulder pain. *BMC Musculoskelet Disord.* 2019;20(1):622.
6. Tengland PA. The concept of work ability. *J Occup Rehabil.* 2011;21(2):275–85.
7. Ilmarinen J, Tuomi K, Seitsamo J. New dimensions of work ability. *Int Congr Ser.* 2005;1280:3–7.
8. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res.* 2010;19(4):539–49.
9. van den Berg TI, Elders LA, de Zwart BC, Burdorf A. The effects of work-related and individual factors on the Work Ability Index: a systematic review. *Occup Environ Med.* 2009;66(4):211–20.

10. Radkiewicz P, Widerszal-Bazyl M. Psychometric properties of Work Ability Index in the light of comparative survey study. *Int Congr Ser.* 2005;1280:304–9.
11. Ebener M, Hasselhorn HM. Validation of short measures of work ability for research and employee surveys. *Int J Environ Res Public Health.* 2019;16(18):3386.
12. Abdolalizadeh M, Arastoo AA, Ghsemzadeh R, Montazeri A, Ahmadi K, Azizi A. The psychometric properties of an Iranian translation of the Work Ability Index (WAI) questionnaire. *J Occup Rehabil.* 2012;22(3):401–8.
13. Blik op werk. Available from: <https://www.blikopwerk.nl/werkgever/duurzame-inzetbaarheid/work-ability-index/the-wai-in-the-netherlands> [Accessed: May 2020].
14. Bascour-Sandoval C, Soto-Rodríguez F, Muñoz-Poblete C, Marzuca-Nassr GN. Psychometric properties of the Spanish version of the Work Ability Index in working individuals. *J Occup Rehabil.* 2020;30(2):288–97.
15. Kaewboonchoo O, Ratanasiripong P. Psychometric properties of the Thai version of the work ability index (Thai WAI). *J Occup Health.* 2015;57(4):371–7.
16. Alexopoulos E. Work Ability Index: validation of the Greek version and descriptive data in heavy industry employees. *Br J Med Res.* 2013;3(3):608–21.
17. Ahlstrom L, Grimby-Ekman A, Hagberg M, Dellve L. 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.* 2010;36(5):404–12.
18. El Fassi M, Bocquet V, Majery N, Lair ML, Couffignal S, Mairiaux P. Work ability assessment in a worker population: comparison and determinants of Work Ability Index and Work Ability score. *BMC Public Health.* 2013;13:305.
19. Torgén M. Experiences of WAI in a random sample of the Swedish working population. *Int Congr Ser.* 2005;1280:328–32.
20. Lundin A, Kjellberg K, Leijon O, Punnett L, Hemmingsson T. The association between self-assessed future work ability and long-term sickness absence, disability pension and unemployment in a general working population: A 7-year follow-up study. *J Occup Rehabil.* 2016;26(2):195–203.
21. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol.* 2010;63(7):737–45.
22. Mokkink LB, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, de Vet HCW, et al. COSMIN Study Design checklist for patient-reported outcome measurement instruments [PDF file]. Department of Epidemiology and Biostatistics: Amsterdam University Medical Centers, Location VUmc; 2019. p. 1–32.
23. Beemster TT, van Velzen JM, van Bennekom CA, Frings-Dresen MH, Reneman MF. Cost-effectiveness of 40-hour versus 100-hour vocational rehabilitation on work participation for workers on sick leave due to subacute or chronic musculoskeletal pain: study protocol for a randomized controlled trial. *Trials.* 2015;16:317.
24. Reneman MF, Beemster TT, Edelaar MJ, van Velzen JM, van Bennekom C, Escorpizo R. Towards an ICF- and IMMPACT-based pain vocational rehabilitation core set in the Netherlands. *J Occup Rehabil.* 2013;23(4):576–84.
25. van den Heuvel SG, Geuskens GA, Hooftman WE, Koppes LL, van den Bossche SN. Productivity loss at work; health-related and work-related factors. *J Occup Rehabil.* 2010;20(3):331–9.
26. Bouwmans C, Krol M, Severens H, Koopmanschap M, Brouwer W, Hakkaart-van RL. The iMTA productivity cost questionnaire: a standardized instrument for measuring and valuing health-related productivity losses. *Value Health.* 2015;18(6):753–8.
27. Beemster T, van Bennekom C, van Velzen J, Reneman M, Frings-Dresen M. The interpretation of change score of the pain disability index after vocational rehabilitation is baseline dependent. *Health Qual Life Outcomes.* 2018;16(1):182.
28. Krops LA, Wolthuisen L, Dijkstra PU, Jaarsma EA, Geertzen JHB, Dekker R. Reliability of translation of the RAND 36-item health survey in a post-rehabilitation population. *Int J Rehabil Res.* 2018;41(2):128–37.
29. Van der Zee KI, Sanderma R. *Het meten van de algemene gezondheidstoestand met de RAND-36, een handleiding.* 2nd ed. Research Institute SHARE; 2012.
30. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. *Ann Med.* 2001;33(5):337–43.
31. Farrar JT, Young JP Jr, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain.* 2001;94(2):149–58.
32. Kamper SJ, Ostelo RWJG, Knol DL, Maher CG, de Vet HCW, Hancock MJ. 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.* 2010;63(7):760–6.
33. Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. *Ann Card Anaesth.* 2019;22(1):67–72.
34. Beemster TT, van Velzen JM, van Bennekom CAM, Reneman MF, Frings-Dresen MHW. Test-retest reliability, agreement and responsiveness of productivity loss (iPCQ-VR) and healthcare utilization (TiCP-VR) questionnaires for sick workers with chronic musculoskeletal pain. *J Occup Rehabil.* 2019;29(1):91–103.
35. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016;15(2):155–63.
36. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 2007;60(1):34–42.
37. Scholtes VA, Terwee CB, Poolman RW. What makes a measurement instrument valid and reliable? *Injury.* 2011;42(3):236–40.
38. van Schaaijk A, Nieuwenhuijsen K, Frings-Dresen MHW, Sluiter JK. Work ability and work functioning: measuring change in individuals recently returned to work. *Int Arch Occup Environ Health.* 2019;92(3):423–33.
39. Voltmer JB, Deller J. Measuring work ability with its antecedents: evaluation of the Work Ability Survey. *J Occup Rehabil.* 2018;28(2):307–21.
40. Karlsson ML, Busch H, Aboagye E, Jensen I. Validation of a measure of health-related production loss: construct validity and responsiveness - a cohort study. *BMC Public Health.* 2015;15:1148.
41. Vänni K, Virtanen P, Luukkaala T, Nygård CH. Relationship between perceived work ability and productivity loss. *Int J Occup Saf Ergon.* 2012;18(3):299–309.
42. Marfeo EE, Haley SM, Jette AM, Eisen SC, et al. Conceptual foundation for measures of physical function and behavioral health function for Social Security work disability evaluation. *Arch Phys Med Rehabil.* 2013;94(9):1645–52.
43. Mänty M, Kouvonen A, Lallukka T, Lahti J, Lahelma E, Rahkonen O. Changes in working conditions and physical health functioning among midlife and ageing employees. *Scand J Work Environ Health.* 2015;41(6):511–8.
44. Froud R, Abel G. 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.* 2014;9(12):e114468.
45. Olsen MF, Bjerre E, Hansen MD, Tendal B, Hilden J, Hróbjartsson A. Minimum clinically important differences in chronic pain vary considerably by baseline pain and methodological factors systematic review of empirical studies. *J Clin Epidemiol.* 2018;101:87–106.

46. Akobeng AK. Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta Paediatr.* 2007;96(5):644–7.
47. Adel M, Akbar R, Ehsan G. Validity and reliability of Work Ability Index (WAI) questionnaire among Iranian workers; a study in petrochemical and car manufacturing industries. *J Occup Health.* 2019;61(2):165–74.
48. Sell L, Lund HL, Holtermann A, Sjøgaard K. The interactions between pain, pain-related fear of movement and productivity. *Occup Med (Lond).* 2014;64(5):376–81.
49. Escobar A, Riddle DL. Concordance between important change and acceptable symptom state following knee arthroplasty: the role of baseline scores. *Osteoarthritis Cartilage.* 2014;22(8):1107–10.

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