



# Test–Retest Reliability, Agreement and Responsiveness of Productivity Loss (iPCQ-VR) and Healthcare Utilization (TiCP-VR) Questionnaires for Sick Workers with Chronic Musculoskeletal Pain

Timo T. Beemster<sup>1,2,3</sup> · Judith M. van Velzen<sup>2,3</sup> · Coen A. M. van Bennekom<sup>2,3</sup> · Michiel F. Reneman<sup>1</sup> · Monique H. W. Frings-Dresen<sup>3</sup>

Published online: 16 March 2018  
© The Author(s) 2018

## Abstract

**Purpose** The purpose of this study was to assess test–retest reliability, agreement, and responsiveness of questionnaires on productivity loss (iPCQ-VR) and healthcare utilization (TiCP-VR) for sick-listed workers with chronic musculoskeletal pain who were referred to vocational rehabilitation. **Methods** Test–retest reliability and agreement was assessed with a 2-week interval. Responsiveness was assessed at discharge after a 15-week vocational rehabilitation (VR) program. Data was obtained from six Dutch VR centers. Test–retest reliability was determined with intraclass correlation coefficient (ICC) and Cohen’s kappa. Agreement was determined by Standard Error of Measurement (SEM), smallest detectable changes (on group and individual level), and percentage observed, positive and negative agreement. Responsiveness was determined with area under the curve (AUC) obtained from receiver operation characteristic (ROC). **Results** A sample of 52 participants on test–retest reliability and agreement, and a sample of 223 on responsiveness were included in the analysis. Productivity loss (iPCQ-VR): ICCs ranged from 0.52 to 0.90, kappa ranged from 0.42 to 0.96, and AUC ranged from 0.55 to 0.86. Healthcare utilization (TiCP-VR): ICC was 0.81, and kappa values of the single healthcare utilization items ranged from 0.11 to 1.00. **Conclusions** The iPCQ-VR showed good measurement properties on working status, number of hours working per week and long-term sick leave, and low measurement properties on short-term sick leave and presenteeism. The TiCP-VR showed adequate reliability on all healthcare utilization items together and medication use, but showed low measurement properties on the single healthcare utilization items.

**Keywords** Productivity loss · Healthcare utilization · Vocational rehabilitation · Cost-effectiveness · Measurement properties

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s10926-018-9767-9>) contains supplementary material, which is available to authorized users.

---

✉ Timo T. Beemster  
t.beemster@amc.nl

- <sup>1</sup> Department of Rehabilitation Medicine, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
- <sup>2</sup> Department of Research and Development, Heliomare Rehabilitation Center, Wijk aan Zee, The Netherlands
- <sup>3</sup> Academic Medical Center, Department Coronel Institute of Occupational Health, Amsterdam Public Health Research Institute, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

## Introduction

Chronic musculoskeletal pain (CMP) is a common condition that results in major disability and substantial healthcare costs [1, 2]. CMP has a negative impact on performing work, resulting in productivity loss from work; reflected by absenteeism (sick off work) or presenteeism (productivity loss while at work) [3]. Productivity loss is labeled in cost-effectiveness studies as indirect healthcare costs [4]. Direct health costs are intervention costs, traveling costs and healthcare utilization costs. Vocation rehabilitation (VR) showed (cost-)effective in improving absenteeism and presenteeism and the reduction of healthcare utilization [5–7].

For clinical practice and research purposes, data about the (cost-)effectiveness of VR interventions are often collected with patient-reported outcome measures (PROMS). PROMS

are standardized, validated questionnaires that are completed by patients to measure their perceptions of their functional status and wellbeing [8]. To give reliable statements on the (cost-)effectiveness of VR, PROMS on productivity loss and healthcare utilization must show adequate measurement properties [3, 8].

However, currently there are no gold standards available for the assessment of productivity loss [9–12]. Evidence on retest reliability and responsiveness on PROMS on absenteeism is scarce [13] and shows mixed results [11]. Research on retest reliability of five presenteeism questionnaires showed moderate to sufficient retest reliability in a sample with rheumatic diseases (ICCs 0.59–0.78) [10], and low to moderate responsiveness in a sample with rheumatoid arthritis or osteoarthritis [14]. However, some issues with presenteeism questionnaires are prominent; they have different recall periods, different outcome scales (0–10 or 1–7), are developed for different populations (general or sickness-specific, for example rheumatic diseases), and they measure different concepts of presenteeism, for example productivity, performance or ability [10]. As a consequence, the correlation between global measures of presenteeism is low, which complicates comparison [10].

Two Dutch questionnaires on the assessment of productivity loss and healthcare utilization have recently been developed. These questionnaires are recommended by the Dutch guideline for health economic evaluations [4]. The questionnaire on the measurement of productivity loss is called the iMTA Productivity Cost Questionnaire (iPCQ) [11, 15–17] and the questionnaire on the assessment of healthcare utilization is called the Trimbos iMTA questionnaire for measuring Costs of Psychiatric Illnesses (TiC-P, part I) [18]. In addition, the TiC-P consists of two parts, a healthcare usage part (part I) and a productivity loss part (part II). Part II has been further developed for the general population and resulted in the iPCQ. In a sample with mental problems, the TiC-P (parts I and II) showed sufficient feasibility and construct validity, and low to sufficient retest reliability [18]. In another study, the feasibility and face validity of the iPCQ was confirmed [15].

However, the iPCQ and TiC-P questionnaires are not fully applicable for sick workers with CMP who are referred to VR. For example, a large portion of sick workers referred to VR are on part-time sick leave and thus part-time at work. The iPCQ, however, does not measure part-time work/sick leave. Furthermore, the TiC-P questionnaire contains many items about mental healthcare but, for example, no items about workplace adaptations or visits of reintegration specialists. Therefore, we modified the iPCQ and TiCP questionnaires to enhance feasibility and usefulness. We called these modified versions the TiCP-VR and the iPCQ-VR. The aim of this study is to assess the test–retest reliability, agreement and responsiveness of the iPCQ-VR and TiCP-VR in

workers with chronic musculoskeletal pain and referred to VR in the Netherlands.

## Methods

The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist was applied in the design of the study [19].

## Procedures

For this study we used two study samples. The first study sample was used to perform the retest reliability and agreement analysis, the second study sample was used to perform the responsiveness analysis. Participants of the first sample were recruited from six VR centers in the Netherlands (Rijndam, MRC Doorn, Klimmendaal, Trappenberg, UMCG CvR and Heliomare). At baseline (T0), patients completed the iPCQ-VR, TiCP-VR and other web-based questionnaires at home as part of care as usual [20]. After a multidisciplinary screening, eligible patients were informed about the study by a member of the multidisciplinary screening team and written information describing the study was provided. 2 weeks after T0, respondents received the iPCQ-VR and TiCP-VR for the second time (T1). If T0 was more than 2 weeks before granted informed consent, the T0 and T1 questionnaires were sent with 2 weeks in between. If participants did not complete the T0 or T1 questionnaires within a week, they received a reminder email. If the questionnaires were not completed after this reminder, participants were phoned by the first author TB. Data of study sample 2 was derived from routinely collected data from six Dutch rehabilitation centers (Heliomare, Roessingh, Adelante, Libra, Klimmendaal, Trappenberg), all offering a multidisciplinary VR program (15-week duration) for workers with chronic musculoskeletal pain. We used baseline (T0) and discharge data (T2). The T2 questionnaires were automatically sent 14 weeks after the start of the VR program. Figure 1 shows the measurement points of samples 1 and 2.

## Participants

The inclusion criteria were: (1) being of working age (18–65 years); (2) suffering from subacute (6–12 weeks) or chronic (> 12 weeks) nonspecific musculoskeletal pain such as back, neck, shoulder, widespread pain, Whiplash Associated Disorder (WAD I or II), or fibromyalgia; (3) having paid work (employed or self-employed) for at least 12 h per week; (4) having sick leave (part-time or full-time); (5) being able to complete questionnaires in Dutch; (6) having an email address; and (7) having granted informed consent. The exclusion criterion was having comorbidities that were the



**Fig. 1** Measurement points of this study. VR vocational rehabilitation, Sample 1: assessment of test–retest reliability and agreement, Sample 2: assessment of responsiveness

primary reason for sick leave, such as acute or specific medical problems, clinical depression or burnout, severe asthmatic symptoms, diagnosed chronic fatigue, and neuropathy. The Medical Ethical Committee of the Academic Medical Center, Amsterdam, the Netherlands, authorized this study and decided that a full application was not required. Participation in the study was voluntary, all participants provided informed consent and answers were processed anonymously.

## Measurements

### Patient Characteristics

Several demographic and clinical variables were assessed at baseline: age, gender, education, pain features (location, duration and intensity), work features (status, contract), and level of disability.

### iPCQ-VR

The iPCQ-VR is a modified version of the iPCQ [11, 15, 17, 18], and is used by six VR centers in the Netherlands. The iPCQ-VR adopted the absenteeism and presenteeism modules of the original iPCQ [17], and two extra modules were added: working status and pain-specific sick leave. We pilot-tested preliminary versions within our research team and four patients pilot-tested the pre-final version of the questionnaire. All items of the iPCQ-VR and the corresponding rating scales are shown in Online Appendix 1.

### TiCP-VR

The original TiC-P assesses the visits and consultation of several healthcare providers, and medication use [18]. The utilization of each healthcare provider is assessed with a yes/

no item and if patients answer ‘yes’, the number of visits/consultations is assessed. A recall period of 4 weeks is used in the original questionnaire, which we adopted in the TiCP-VR version. In the TiCP-VR version, we removed five items that were specific to psychiatric patients, but not for our population. Furthermore, we added pain-specific items to allow differentiation between pain-related and other healthcare utilization. Finally, we removed non pain-related medication use. This was due to feasibility reasons and it was expected that medication use other than pain-related was marginal when translated to costs. Also, it was expected that this adaptation would prevent missing data on medication use, as this was prominent in the original TiC-P validation study [18]. We pilot-tested preliminary versions within our research team and four patients pilot-tested the pre-final version of the questionnaire. All items of the TiCP-VR and the corresponding rating scales are shown in Online Appendix 2.

### Global Perceived Effect

One global perceived effect (GPE) item (‘How much did the vocational rehabilitation program change your work functioning compared to pre-treatment level?’) was assessed at T2 and was used as the external criterion (anchor) in the responsiveness analysis in this study. GPE was measured with a 7-point Likert scale ranging from 1 to 7 (1; ‘extremely worsened’, 2; ‘much worsened’, 3; ‘little worsened’, 4; ‘unchanged’, 5; ‘little improved’, 6; and ‘much improved’, 7; ‘completely improved’).

## Statistical Analysis

### Reliability

Test–retest reliability of the *continuous items* of the iPCQ-VR were performed with intraclass correlation coefficient (ICC random, single, and on absolute agreement) [21]. To allow comparison with other studies, in particular the original iPCQ study by Bouwmans et al. [18], we performed sensitivity analyses with ICC random, average, and on absolute agreement. One overall ICC of all healthcare visits/consultations of TiCP-VR together was calculated because the single continuous items were expected to be underpowered [18]. We considered an ICC of  $>0.70$  sufficient for use at group level and an ICC of  $>0.90$  sufficient for use at individual level [22].

Reliability of *dichotomous items* of iPCQ-VR and TiCP-VR were studied using Cohen's kappa analyses [ $k = P_o - P_c / 1 - P_c$ ] where  $P_o$  is the proportion of observed agreements and  $P_c$  is the proportion of agreements expected by chance [23]. The range of possible values of kappa is from  $-1$  to  $1$  [23]. We interpreted kappa values as follows: slight (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80) and almost perfect (0.81–1.00) [24]. The pain-specific items of the TiCP-VR were expected to be underpowered and were blended to one  $2 \times 2$  contingency table.

Reliability of *categorical variables* was performed with linear weighted kappa coefficients [25, 26].

### Agreement

Agreement of *continuous variables* was analyzed by the standard error of measurement [ $SEM = SD\sqrt{1 - ICC}$ ], where SD is the SD of the scores from all participants, which were determined from an ANOVA analysis with the formula [ $\sqrt{SS_{total} / (n - 1)}$ ], and ICC is the retest reliability coefficient [21]. The SEM was converted into the smallest detectable changes on individual level [ $SDC_{individual} = 1.96 \times \sqrt{2 \times SEM}$ ]. This number reflects the smallest within-person change in a score that can be considered to be a real change above any measurement error within one individual. The SDC individual was converted into the SDC for a group (SDC group) by dividing the SDC individual by  $\sqrt{n}$ . We proposed a positive rating for agreement if the absolute measurement error (SDC individual for change within individuals and SDC group for change between groups) is smaller than the minimal important change (MIC, see responsiveness) [27, 28].

Agreement of *dichotomous variables* was analyzed by the percentage observed agreement [ $P_o = (a + d)/n$ ], the percentage positive agreement [ $PA = 2a/2a + b + c$ ], and the

percentage negative agreement [ $NA = 2d/2d + b + c$ ] [29]. PA is known as the specific agreement on a positive rating and NA is known as the specific agreement on a negative rating [29]. All  $2 \times 2$  contingency tables will be provided in Online Appendices 3 and 4. *Categorical variables* were analyzed by the percentage observed agreement.

### Responsiveness

Responsiveness in this study was defined as the ability of the iPCQ-VR to detect clinically relevant changes over time [27]. We assessed the responsiveness on four continuous items: the number of sick leave days in the preceding 4 weeks (for participants with short-term sick leave at T0), the number of working hours per week (for participants with 100% sick leave at T0), the number of presenteeism days in the preceding 4 weeks and the presenteeism score (0–10) (for participants who scored 'yes' on presenteeism at T0). Various statistics were applied to calculate responsiveness [30]. Mean changes and 95% confidence intervals of mean changes were calculated. Sensitivity and specificity for change plotted by receiver operating characteristic (ROC) curve and area under the curves (AUCs) were calculated [31]. The AUC is the probability of correctly discriminating between improved and nonimproved patients. When the AUC was more than 0.70, responsiveness was considered sufficient [27]. MIC was measured by determining the optimal cut-off point (OCP). This is the point of the ROC curve where the sum of sensitivity and 1-specificity is maximal. Sensitivity and specificity of the OCP were computed. Sensitivity and specificity range from 0 to 1.00, where higher numbers reflect higher sensitivity or specificity. Because the objective of the responsiveness analysis was to differentiate between improved and unchanged samples of participants, the GPE score was dichotomized into a subgroup with GPE score "improved" (little improved, much improved and completely improved) and a subgroup with the GPE score "unchanged". The GPE group "worsened" was not included in the analyses [30].

### Stability

The ICC, kappa, and agreement analyses were performed on a stable sample that completed the questionnaire twice in similar conditions, with a 2-week interval. To perform this, we added external anchor items at T1 (external anchor item: 'In relation to question x, did something change in the preceding 2 weeks, compared to the weeks before?'). To allow comparison with other studies, results of both stable and unstable (i.e. total sample) retest samples will be reported.

We applied an online calculation tool to calculate kappa and linear weighted kappa [32]. All other analyses were performed using SPSS 23 for Windows (SPSS Inc., Chicago, USA). The demographic data of the individuals were described by means and standard deviations (SD), or interquartile range in the case of no normal distribution. The assumption of normal data distribution was visually verified using histograms and QQ-plots.

### Power

Fifty patients are needed to obtain a reasonable  $2 \times 2$  contingency table to determine the kappa and to obtain a confidence interval ranging from 0.70 to 0.90 around an ICC of 0.80 [12, 24, 27]. 50 to 99 patients are needed to obtain reasonable responsiveness scores [33].

### Results

A total of 52 participants completed the retest questionnaires (response rate retest 71%). Reasons for non-response were technical problems ( $n = 7$ ), withdrawal consent ( $n = 3$ ), no telephone number ( $n = 2$ ), or unknown ( $n = 9$ ). The retest was submitted on average 19.6 days (SD 5.8) after submission of the initial questionnaires. A sample of 223 participants completed baseline and discharge responsiveness questionnaires. Response rates of this sample were unknown. The responsiveness questionnaires were submitted on average 14.5 weeks (SD 1.0) after T0. Table 1 shows the characteristics of both study samples.

### Reliability

The ICCs of the iPCQ-VR ranged from 0.52 to 0.90 (Table 2). Number of working hours per week scored 0.90, number of short-term sick leave days scored 0.54, presenteeism score scored 0.56, and number of presenteeism days scored 0.52. The ICC of total healthcare utilization was 0.81. Sensitivity analysis with average measures of ICC showed the following ICCs: number of working hours (0.95), presenteeism score (0.72), number of presenteeism days (0.68), number of sick leave days (0.70), and total healthcare utilization (0.89).

Cohen's kappa of the iPCQ-VR ranged from 0.42 to 0.96 (Table 2). In the total (both stable and unstable participants) sample, long-term pain-specific sick leave scored a kappa of 1.00 (Table 3). Cohen's kappa items of the healthcare utilization items of the TiCP-VR ranged from 0.11 to 1 (Table 4). Medication use showed substantial kappa (0.78) and total pain-specific healthcare utilization showed fair kappa (0.35). Table 5 shows kappa and agreement measures of the total

**Table 1** Characteristics of the study populations

	Reliability and agreement (n = 52) Mean (SD) or %	Responsiveness (n = 223) Mean (SD) or %
Age (years)	44.6 (10.3)	47.4 (10.9)
Gender (% female)	78.8	59.9
Education		
Low	17.0	15.3
Medium	55.3	55.4
High	27.7	24.3
Work (% yes)	96.2	96.0
Work situation		
Employed	90.2	87.5
Self-employed	2.0	2.7
Disability benefit	3.9	1.4
Housewife/houseman	2.0	1.4
Incapacitated	2.0	0.9
Contract (hours/week)	32.8 (7.0)	31.3 (8.8)
Contract (days/week)	4.3 (1.1)	4.2 (1.0)
Work status		
Working full-time	14.0	8.5
Working part-time	58.0	52.6
100% sick leave	28.0	39.0
Sick leave short (% yes)	50	62
Sick leave long (% yes)	36	49.8
Presenteeism (% yes)	74	63.8
Pain location		
Spine (% yes)	69.2	76.7
Lower extremities (% yes)	32.7	35.0
Upper extremities (% yes)	42.3	45.3
Headache/burnout	30.8	40.8
Number of pain locations (IQR)	2 (1–4)	2 (1–4)
Pain duration		
1–3 months	9.6	8.6
3–6 months	15.4	15.3
0.5–1 year	28.8	23.0
1–2 years	15.4	22.1
2–5 years	17.3	13.1
More than 5 years	13.5	18.0
Pain mean (0–10) <sup>a</sup>	6.0 (1.9)	5.5 (2.3)
Pain worse (0–10) <sup>a</sup>	7.4 (1.9)	6.9 (2.5)
Pain disability index (0–70) <sup>b</sup>	37 (12.1)	34.3 (11.6)

SD standard deviation, IQR interquartile range

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

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

sample on the TiCP-VR items. Online Appendix 3 (iPCQ-VR) and Online Appendix 4 (TiCP-VR) show all  $2 \times 2$  contingency tables of both stable and unstable (total) samples.

**Table 2** ICC, kappa and agreement of the iPCQ-VR in a stable group of participants

	Observed agreement (%)	Positive agreement (%)	Negative agreement (%)	Cohen's kappa (95% CI)	Yes (N)	ICC single (95% CI)
<b>Work status</b>						
Working status	98	NA	NA	0.96 (0.88–1) <sup>a</sup>	41	–
Number of hours working per week <sup>b</sup>	–	–	–	–	16	0.90 (0.74–0.96)
<b>Sick leave short (&lt;4 weeks)</b>						
Sick leave during last 4 weeks	72	71	73	0.45 (0.19–0.70)	16	–
Number of sick leave days	–	–	–	–	16	0.54 (0.05–0.81) <sup>c</sup>
Sick leave pain-specific	94	96	80	0.76 (0.32–1)	13	–
<b>Sick leave long (&gt;4 weeks)</b>						
Sick leave longer than 4 weeks	88	83	91	0.74 (0.54–0.94)	15	–
Sick leave pain-specific	NA	NA	NA	NA	14	–
<b>Presenteeism (&lt;4 weeks)</b>						
Presenteeism during last 4 weeks	74	80	62	0.42 (0.13–0.72)	22	–
Number of presenteeism days	–	–	–	–	23	0.52 (0.14–0.76)
Score of 0–10	–	–	–	–	23	0.56 (0.21–0.79)
Healthcare usage <sup>d</sup>	NA	NA	NA	NA	37	0.81 (0.66–0.89)

NA not applicable

<sup>a</sup>Linear weighted kappa

<sup>b</sup>In the case of part-time working (one of the three 'working status' options)

<sup>c</sup>One extreme outlier (value 49) was excluded from the analysis

<sup>d</sup>All TiCP-VR healthcare usage items (01–14) together

**Table 3** ICC, kappa and agreement of the iPCQ-VR in total sample (stable and unstable participants)

	Observed agreement (%)	Positive agreement (%)	Negative agreement (%)	Cohen's kappa (95% CI)	Yes (N)	ICC single (95% CI)
<b>Work status</b>						
Working status	86	NA	NA	0.80 (0.65–0.94) <sup>a</sup>	50	–
Working part-time: number of hours working per week <sup>b</sup>	–	–	–	–	24	0.76 (0.52–0.89)
<b>Sick leave short (&lt;4 weeks)</b>						
Sick leave during last 4 weeks	72	71	73	0.44 (0.19–0.69)	17	–
Number of sick leave days	–	–	–	–	16	0.54 (0.05–0.81) <sup>c</sup>
Sick leave pain-specific	94	80	97	0.77 (0.34–1.00)	14	–
<b>Sick leave long (&gt;4 weeks)</b>						
Sick leave longer than 4 weeks	88	83	91	0.74 (0.54–0.94)	15	–
Sick leave pain-specific	100	100	100	1.00 (1.00–1.00)	14	–
<b>Presenteeism (&lt;4 weeks)</b>						
Presenteeism during last 4 weeks	76	83	60	0.43 (0.17–0.70)	29	–
Number of presenteeism days	–	–	–	–	29	0.51 (0.18–0.74)
Score of 0–10	–	–	–	–	29	0.54 (0.22–0.75)

NA not applicable

<sup>a</sup>Linear weighted kappa

<sup>b</sup>In the case of part-time working (one of the three 'working status' options)

<sup>c</sup>One extreme outlier (value 49) was excluded from the analysis

**Table 4** Kappa and agreement of the TiCP-VR in a stable group of participants

	N	Use of service (N)	No use of service (N)	Observed agreement (%)	Positive agreement (%)	Negative agreement (%)	Cohen's kappa (95% CI)
01 General practitioner	46	9	17	56	57	63	0.13 (0–0.42)
02 PT, ET, OT, MT	42	16	16	76	86	76	0.55 (0.31–0.79)
03 Occupational physician	43	18	6	56	65	39	0.11 (0–0.41)
04 Reintegration specialist	44	6	32	86	67	91	0.59 (0.28–0.89)
05 Insurance physician	49	0	47	96	NA	98	NA
06 Workplace adaptations	48	6	35	85	63	91	0.54 (0.23–0.86)
07 Medical specialist	46	6	31	80	57	87	0.45 (0.12–0.77)
08 Stay in healthcare setting	50	0	50	100	NA	100	NA
09 Psychiatrist or psychologist	47	5	31	76	48	85	0.35 (0–0.68)
10 Social worker	49	0	48	98	NA	99	NA
11 Dietician	49	1	48	100	100	100	1
12 Home care	49	0	48	98	NA	99	NA
13 Alternative care	48	3	41	92	60	95	0.56 (0.14–0.97)
14 Medication use	38	21	13	89	91	87	0.78 (0.58–0.98)
Medication use pain-specific	27	14	2	59	NA	NA	0.31 (0–0.66) <sup>a</sup>
Healthcare usage pain-specific <sup>b</sup>	50	70	3	89	94	40	0.35 (0–0.75)

NA not applicable, PT, ET, OT, MT physical therapist, exercise therapist, occupational therapist, manual therapist

<sup>a</sup>Linear weighted kappa

<sup>b</sup>The sum of all pain-specific healthcare usage items

**Table 5** Kappa and agreement of the TiCP-VR in total sample (stable and unstable participants)

	N	Use of service (N)	No use of service (N)	Observed agreement (%)	Positive agreement (%)	Negative agreement (%)	Cohen's kappa (95% CI)
01 General practitioner	50	12	18	60	55	64	0.21 (0–0.48)
02 PT, ET, OT, MT	50	23	16	78	81	74	0.57 (0.35–0.80)
03 Occupational physician	50	19	7	52	61	59	0.04 (0–0.32)
04 Reintegration specialist	49	9	33	86	72	90	0.63 (0.37–0.88)
05 Insurance physician	50	0	47	94	na	97	NA
06 Workplace adaptations	50	7	35	84	64	90	0.53 (0.24–0.83)
07 Medical specialist	50	8	33	82	44	88	0.52 (0.24–0.80)
08 Stay in health care setting	50	0	50	100	na	100	NA
09 Psychiatrist or psychologist	50	7	31	76	54	84	0.40 (0.10–0.69)
10 Social worker	49	0	48	98	na	99	NA
11 Dietician	49	1	48	100	100	100	1
12 Home care	49	0	48	98	na	99	NA
13 Alternative care	49	4	41	92	67	95	0.62 (0.27–0.98)
14 Medication use	48	30	13	90	92	84	0.76 (0.56–0.96)
Medication use pain specific	30	17	2	63	NA	NA	0.36 (0.05–0.68) <sup>a</sup>
Health care usage pain specific <sup>b</sup>	50	74	4	82	90	32	0.23 (0–0.56)

NA not applicable, PT, ET, OT, MT physical therapist, exercise therapist, occupational therapist, manual therapist

<sup>a</sup>Linear weighted kappa

<sup>b</sup>The sum of all pain-specific healthcare usage items

**Table 6** Responsiveness iPCQ-VR

Variable	N anchor	AUC (CI)	MIC	Sens	Spec	SEM <sup>a</sup>	SDCind <sup>a</sup>	SDCgrp <sup>a</sup>
Working hours <sup>b</sup>	49 positive 22 negative	0.86 (0.77–0.94)	–1	0.796	0.818	0.83	2.31	0.58
Sick leave days <sup>c</sup>	76 positive 31 negative	0.66 (0.55–0.78)	5.5	0.671	0.581	3.64	10.09	2.52
Presenteeism days <sup>d</sup>	84 positive 28 negative	0.55 (0.44–0.67)	4.5	0.548	0.571	2.84	7.87	1.64
Presenteeism score <sup>d</sup>	89 positive 29 negative	0.60 (0.48–0.72)	–0.5 to –1.5	0.708 0.584	0.414 0.655	0.71	1.98	0.41

<sup>a</sup>Available from the SDs and ICCs of sample I

<sup>b</sup>Selection of participants who were on 100% sick leave at T0 and who reported stable or improved work functioning at T2

<sup>c</sup>Selection of participants who were on short-term sick leave at T0 and who reported stable or improved work functioning at T2

<sup>d</sup>Selection of participants who scored ‘yes’ on presenteeism at T0 and who reported stable or improved work functioning at T2

## Agreement

For the continuous items of the iPCQ-VR, the SEM, SDCind and SDCgrp were respectively 0.8, 2.3, 0.6 (number of working hours per week), 3.6, 10.1, 2.5 (number of sick leave days), 2.8, 7.9, 1.6 (number of presenteeism days), 0.7, 2.0, 0.4 (presenteeism score) (Table 6).

For the dichotomous items, observed agreement of the iPCQ-VR ranged from 72 to 98%, positive agreement ranged from 71 to 96% and negative agreement ranged from 62 to 91% (Table 2). Observed agreement (OA) of the healthcare items of the TiCP-VR ranged from 56 to 100%, positive agreement (PA) ranged from 48 to 100%, and negative agreement (NA) ranged from 39 to 100% (Table 4). Medication use scored OA: 89%, PA: 91%, NA: 87%. Pain-specific medication use (categorical item) scored OA: 59%. All pain-specific healthcare items together scored OA: 89%, PA: 94%, NA: 40%.

## Responsiveness

The AUC, MIC, sensitivity and specificity of the iPCQ-VR are presented in Table 6 and the ROC curves are shown in Fig. 2. The AUCs ranged from 0.55 to 0.86. The number of working hours per week showed adequate responsiveness for the participants who were on 100% sick leave at baseline (AUC 0.86, MIC = –1). Sick leave days in the preceding 4 weeks showed moderate responsiveness (AUC 0.66, MIC = 5.5). Presenteeism days in the preceding 4 weeks showed poor responsiveness (AUC 0.55, MIC = 4.5). Presenteeism score showed moderate responsiveness (AUC 0.60, MIC = –0.5 to –1.5). Table 7 shows the mean change scores of the iPCQ-VR.

## Discussion

In this study, the retest reliability, agreement and responsiveness of two modified questionnaires on productivity loss (iPCQ-VR) and healthcare utilization (TiCP-VR) for workers on sick leave due to chronic musculoskeletal pain and referred to VR was assessed.

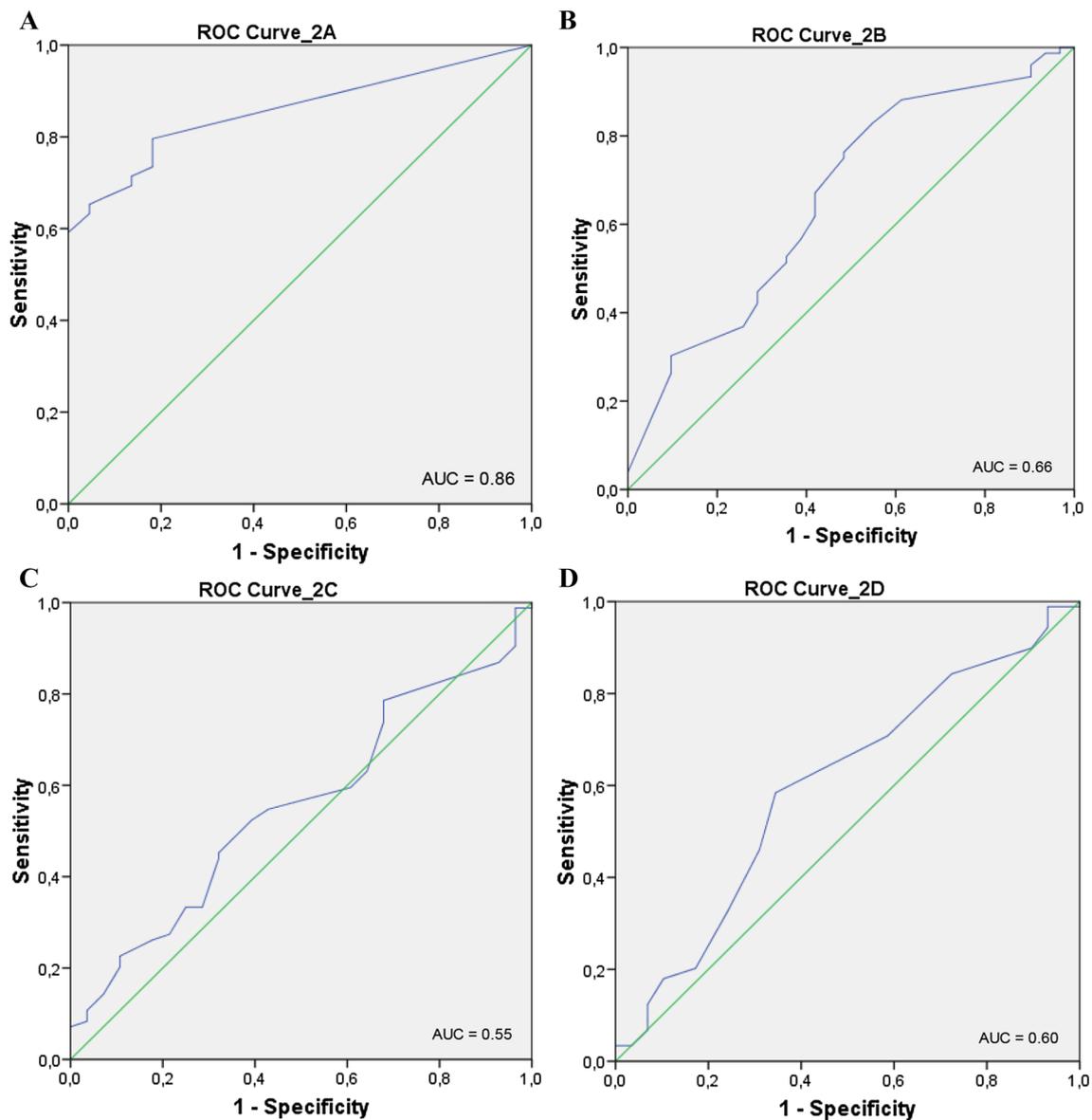
### iPCQ-VR

The working status and number of working hours per week items scored high on retest reliability, agreement, and responsiveness. These items can be used at the group and individual levels as well as for evaluative purposes. Long-term sick leave scored sufficient retest reliability and agreement and can be used at group level. Short-term sick leave and presenteeism scored low retest reliability, agreement and responsiveness, and can therefore not be used at the group or individual level, or for evaluative purposes.

### Reliability

Comparing the retest reliability of the absenteeism items of the current study with the original study [18] is complicated, because the original study used average measures ICC,<sup>1</sup> which results in higher ICCs. In our opinion, single measures ICC is the appropriate ICC to answer the research question on retest reliability because in clinical practice patients complete the iPCQ-VR once per measurement point (i.e. at baseline, discharge, follow-up). Furthermore, the original study measured short-term sick leave with a recall period of 2 weeks, whereas we applied 4 weeks. Finally, the original study did not select a stable group of participants.

<sup>1</sup> The type of ICC is not clearly stated in the article. This information was known after e-mailing with the last author LHvR.



**Fig. 2** ROC curves of the iPCQ-VR. **a** Number of working hours per week, ROC curve of participants who were on 100% sick leave at T0 and who reported stable or improved work functioning at T2 ( $n=71$ ). **b** Number of sick leave days in preceding 4 weeks, ROC curve of participants who were on short-term sick leave at T0 and who reported stable or improved work functioning at T2 ( $n=107$ ). **c** Number of presenteeism days in preceding 4 weeks, ROC curve of participants

who scored 'yes' on presenteeism at T0 and who reported stable or improved work functioning at T2 ( $n=112$ ). **d** Presenteeism score (0–10) of preceding 4 weeks, ROC curve of participants who scored 'yes' on presenteeism at T0 and who reported stable or improved work functioning at T2 ( $n=118$ ). ROC receiver operating characteristic, AUC area under the curve

In a recent systematic review, the psychometric properties of eleven work productivity questionnaires were examined [11]. Data on the retest reliability of absenteeism was available for only four questionnaires. However, we cannot compare our results with these questionnaires for several reasons: no ICC or kappa performed [34–36], type of ICC unknown [37, 38], or a different recall period (3 months) and calculation of kappa (absenteeism 0 vs. > 0 days) [39].

Despite the importance of absenteeism data as a return to work outcome and as a resource for economic evaluations, the evidence on the reliability of absenteeism measures is remarkably scarce. A possible explanation for this is that in several countries researchers can obtain sick leave data from social security databases [40], which is a feasible and reliable alternative [13]. However, such databases are not available for all countries, and another disadvantage is that the accuracy of sick leave data from electronic databases is

**Table 7** Mean change scores of iPCQ-VR

	Working hours	Sick leave days	Presenteeism days <sup>a</sup>	Presenteeism score <sup>b</sup>
Baseline T0				
Mean (SD)	0	13.1 (7.0)	11.8 (6.6)	5.3 (2.3)
Minimum–maximum	0–0	0–31	2–28	0–10
N	83	123	130	136
Discharge T2				
Mean (SD)	8.7 (10.7)	4.1 (7.4)	5.2 (6.0)	7.3 (2.5)
Minimum–maximum	0–40	0–31	0–20	0–10
N	81	127	129	131
Mean change (SD)	–8.7 (10.7)	9.2 (9.0)	6.4 (7.4)	–2.02 (3.2)
95% CI of mean change	–11.08 to –6.35	7.58–10.85	5.08–7.72	–2.58 to –1.46
N	81	119	124	131

<sup>a</sup>If presenteeism in the preceding 4 weeks was answered *no* on T2, presenteeism days was set at zero (0)

<sup>b</sup>If presenteeism in the preceding 4 weeks was answered *no* on T2, presenteeism score was set at ten (10)

low for short recall periods (i.e. “acute” sick leave) [12, 13, 41]. Because the reliability of short term sick leave was also low in the present study, this measure warrants improvement in future studies.

The ICCs ranging from 0.52 to 0.56 of the presenteeism items of the current study are somewhat lower compared with a review on the reliability of five at work productivity loss questionnaires in patients with rheumatic diseases, with single measures ICCs ranging from 0.59 to 0.78 ( $n=62-65$ ) [10]. The higher ICCs of other studies can be explained by the low power ( $n=23$ ) and longer recall period (four weeks) of the present study. A power of  $\geq 50$  and a recall period of 1 week is advocated [12].

### Agreement

The observed agreement of the current study was somewhat lower compared with the original study (short-term sick leave: 72 vs. 87%, long-term sick leave: 88 vs. 93%, and presenteeism: 74 vs. 81%) [18]. This difference can be explained through a difference in power ( $n=50$  vs.  $n=79$ ). Unfortunately, the original study did not calculate the positive and/or negative agreement. There is one study known which also calculated observed agreement [39], but comparison with this study is not possible due to a different calculation of kappa (0 vs.  $>0$  h of absenteeism, presenteeism). As there are currently no cut-off scores available for the interpretation of positive and negative agreement, the information from the  $2 \times 2$  contingency tables (Online Appendix 3) can be used by the reader to judge the uptake of a questionnaire or a particular item.

### Responsiveness

The responsiveness analyses showed that a minimal important change of  $\geq 1$  working hours per week at discharge of

VR can be used for evaluative purposes for patients who are on full sick leave at baseline. A minimal important change of 5.5 sick leave days per month can be considered for evaluative purposes for patients who are on full sick leave at baseline. However, this warrants caution because the moderate AUC value of 0.66 is below the adequate level of 0.7.

The number of presenteeism days and the presenteeism score cannot be used for evaluative purposes because the AUCs were too low (0.55 and 0.60). One study assessed the responsiveness of five presenteeism scales (ranging from 0 to 10 or 1–7) [14]. In this study, ROCs and AUCs were assessed (and no MICs). The AUCs in this study ranged from 0.52 to 0.66, which is similar to that of the current study.

### TiCP-VR

The sum of all healthcare visits of the TiCP-VR showed sufficient retest reliability and agreement, and can be used at group level. However, the single healthcare items of the TiCP-VR showed low kappa values and moderate agreement, which can be explained by uneven distributions of the  $2 \times 2$  contingency tables (Online Appendix 4). This negatively affects the kappa and agreement values [23]. Furthermore, of four healthcare items (stay in a healthcare setting, social worker, insurance physician, home care) it was not possible to calculate kappa and agreement measures as none of the participants used these services. These items may be deleted to increase feasibility.

Medication use showed substantial retest reliability and adequate agreement. This item can be used at group level. In contrast, pain-specific medication use scored poor retest reliability and agreement, and this item cannot be used at group level and needs to be refined. Unfortunately, due to a technical error we were not able to assess the dosage, frequency and name of the consumed pain medications.

The observed agreement of the current study is in line with the observed agreement from the original study [18]. Comparison on retest reliability (ICC values) with the original study is not possible as they used a different type of ICC.

### Strengths and Limitations

A strength of this study is that we included a sample of patients with chronic musculoskeletal pain who were referred to six VR centers in the Netherlands. This increases the clinical utility of this study. Second, we have extensively investigated both PROMS and we provided all  $2 \times 2$  contingency tables (Online Appendices 3, 4), as recommended [29].

Our results should be generalized cautiously as our study has some limitations that must be addressed. First, an inclusion criteria for this study was that participants should be on sick leave (part-time or full-time) at baseline. However, 14% of study sample one and 8.5% of study sample two were not at sick leave at baseline but full-time at work. This has resulted in lower samples for the performed analyses, which probably negatively affected the results on sick leave and presenteeism. Second, we applied anchor items at measurement 2 to detect stable and unstable (i.e. changed) samples of participants. For working status and the number of hours working per week, this resulted in better results on retest reliability in the stable group of participants. However, for the other items of the iPCQ-VR, such as short- and long-term sick leave and presenteeism, the results remained the same. Remarkably, the healthcare items of the TiCP-VR showed in general lower retest reliability (lower kappa values) in the stable sample compared with the unstable sample. Therefore, the anchor items applied in this study warrant refinement.

Third, we assessed presenteeism with a time interval of 2 weeks. This is in line with similar studies [10]. Presenteeism may be unstable; it can fluctuate between days and weeks. Sim et al. [23] stated that for the time interval in retest reliability studies ‘the stability of the attribute being rated is crucial to the period between repeated ratings’. We advise using a shorter time interval (for example 2 days) with control for stability to increase retest reliability in future studies.

The fourth and final limitation is the second measurement point in the responsiveness analysis (Fig. 1). Due to feasibility/technical reasons, patients received these questionnaires 14-weeks after the start of their 15-week VR program. In clinical practice, this is 1 week before the real discharge date and in some patients, this might even be worse if they were on holiday during the intervention period or had an extension of their training period. We suppose that this flaw yields an underestimation on the responsiveness measures in this study, because when people are in rehabilitation they cannot be at work.

### Clinical Recommendations

We recommend using the working status and number of working hours per week items of the iPCQ-VR to provide an estimation of short-term sick leave, which is in line with the majority of the return to work intervention studies, which use an estimate of lost time from work as their primary RTW outcome [42, 43]. A minimal important change of  $\geq 1$  working hours per week can be used for evaluative purposes for patients who are on full sick leave at baseline. Furthermore, a minimal important change of 5.5 sick leave days per month can be considered for patients who are on full sick leave at baseline. However, this warrants caution due to the moderate AUC of 0.66. The items of the iPCQ-VR should not be used for the assessment of presenteeism.

The sum of all healthcare utilization items of the TiCP-VR can be used at group level, but the single items needs further investigation. The generic item on medication use can be used at group level, but the pain-specific medication use item warrants improvement.

### Conclusion

The iPCQ-VR showed good measurement properties on working status, number of hours working per week and long-term sick leave, and low measurement properties on short-term sick leave and presenteeism. The TiCP-VR showed adequate reliability on total healthcare utilization and medication use, but showed low measurement properties on the single healthcare utilization items.

**Funding** No commercial sponsorship was involved in the design and conduction of the study.

### Compliance with Ethical Standards

**Conflict of interest** Author TB, author JvV, author CvB, author MR, and author MFD declare that they have no conflict of interest.

**Ethical Approval** All procedures performed were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The Medical Ethical Committee of the Academic Medical Center, Amsterdam, the Netherlands, authorized this study and decided that a full application was not required.

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

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate

credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Breivik H, Collett B, Ventafridda V, Cohen R, Gallacher D. Survey of chronic pain in Europe: prevalence, impact on daily life, and treatment. *Eur J Pain*. 2006;10(4):287–333.
- de Vroome EM, Uegaki K, van der Ploeg CP, Treutlein DB, Steenbeek R, de Weerd M, et al. Burden of sickness absence due to chronic disease in the Dutch workforce from 2007 to 2011. *J Occup Rehabil*. 2015;25(4):675–684.
- van Dongen JM, van Wier MF, Tompa E, Bongers PM, van der Beek AJ, van Tulder MW, et al. Trial-based economic evaluations in occupational health: principles, methods, and recommendations. *J Occup Environ Med*. 2014;56(6):563–572.
- Hakkaart-van Roijen LTS, Bouwmans C. Manual for cost research. Methods and standard cost prices for economic evaluations in health care. Rotterdam; 2010.
- Airaksinen O, Brox JI, Cedraschi C, Hildebrandt J, Klaber-Moffett J, Kovacs F, et al. Chapter 4. European guidelines for the management of chronic nonspecific low back pain. *Eur Spine J*. 2006;15(Suppl 2):S192–S300.
- Escorpizo R, Brage S, Homa D, Stucki G. Handbook of vocational rehabilitation and disability evaluation. Cham: Springer; 2014. pp. 3–10
- Waddell G, Burton AK, Kendall NAS. Vocational rehabilitation. What works, for whom, and when? 2013.
- Dawson J, Doll H, Fitzpatrick R, Jenkinson C, Carr AJ. The routine use of patient reported outcome measures in healthcare settings. *BMJ*. 2010;340:c186. <https://doi.org/10.1136/bmj.c186>.
- Krol M, Brouwer W, Rutten F. Productivity costs in economic evaluations: past, present, future. *Pharmacoeconomics*. 2013;31(7):537–549.
- Leggett S, van der Zee-Neuen A, Boonen A, Beaton DE, Bojinca M, Bosworth A, et al. Test–retest reliability and correlations of 5 global measures addressing at-work productivity loss in patients with rheumatic diseases. *J Rheumatol*. 2016;43(2):433–439.
- Tang K. Estimating productivity costs in health economic evaluations: a review of instruments and psychometric evidence. *Pharmacoeconomics*. 2015;33(1):31–48.
- Zhang W, Bansback N, Anis AH. Measuring and valuing productivity loss due to poor health: a critical review. *Soc Sci Med*. 2011;72(2):185–192.
- Ostelo RW, de Vet HC. Clinically important outcomes in low back pain. *Best Pract Res Clin Rheumatol*. 2005;19(4):593–607.
- Beaton DE, Tang K, Gignac MA, Lacaille D, Badley EM, Anis AH, et al. Reliability, validity, and responsiveness of five at-work productivity measures in patients with rheumatoid arthritis or osteoarthritis. *Arthritis Care Res*. 2010;62(1):28–37.
- Bouwmans C, Krol M, Severens H, Koopmanschap M, Brouwer W, Hakkaart-van Roijen L. The iMTA productivity cost questionnaire: a standardized instrument for measuring and valuing health-related productivity losses. *Value Health*. 2015;18(6):753–758.
- Krol M, Brouwer W. How to estimate productivity costs in economic evaluations. *Pharmacoeconomics*. 2014;32(4):335–344.
- Bouwmans C, Hakkaart-van Roijen L, Koopmanschap M, Krol M, Severens H, Brouwer W. Manual of the iMTA productivity cost questionnaire (iPCQ). Rotterdam: iMTA, Erasmus University Rotterdam; 2013.
- Bouwmans C, De Jong K, Timman R, Zijlstra-Vlasveld M, Van der Feltz-Cornelis C, Tan Swan S, et al. Feasibility, reliability and validity of a questionnaire on healthcare consumption and productivity loss in patients with a psychiatric disorder (TiC-P). *BMC Health Serv Res*. 2013;13:217. <https://doi.org/10.1186/1472-6963-13-217>.
- Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol 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):539–549.
- 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):576–584.
- Weir JP. Quantifying test–retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res*. 2005;19(1):231–240.
- Nunnally JC. Psychometric theory. 2nd ed. New York: McGraw-Hill; 1978.
- Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther*. 2005;85(3):257–268.
- De Vet HCW, Terwee CB, Knol DL, Mokkink LB. Measurement in medicine. Cambridge: Cambridge University Press 2011.
- Warrens M. Weighted Kappas for 3 × 3 Tables. *J Probab Stat*. 2013;2013:325831. <https://doi.org/10.1155/2013/325831>.
- Vanbelle S. A new interpretation of the weighted Kappa coefficients. *Psychometrika*. 2016;81(2):399–410.
- Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60(1):34–42.
- Kievit AJ, Kuijer PP, Kievit RA, Sierevelt IN, Blankevoort L, Frings-Dresen MH. A reliable, valid and responsive questionnaire to score the impact of knee complaints on work following total knee arthroplasty: the WORQ. *J Arthroplasty*. 2014;29(6):1169–1175.
- de Vet HC, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen’s kappa. *BMJ*. 2013;346:f2125. <https://doi.org/10.1136/bmj.f2125>.
- Soer R, Reneman MF, Vroomen PC, Stegeman P, Coppes MH. Responsiveness and minimal clinically important change of the pain disability index in patients with chronic back pain. *Spine*. 2012;37(8):711–715.
- Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *J Chronic Dis*. 1986;39(11):897–906.
- Accessed 20 Jun 2017; Available from: <http://vassarstats.net/kappa.html>.
- Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res*. 2012;21(4):651–657.
- Kessler RC, Ames M, Hymel PA, Loeppke R, McKenas DK, Richling DE, et al. Using the World Health Organization Health and work performance questionnaire (HPQ) to evaluate the indirect workplace costs of illness. *J Occup Environ Med*. 2004;46(6 Suppl):S23–S37.
- Goetzl RZ, Hawkins K, Ozminkowski RJ, Wang S. The health and productivity cost burden of the “top 10” physical and mental health conditions affecting six large U.S. employers in 1999. *J Occup Environ Med*. 2003;45(1):5–14.
- Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. *Pharmacoeconomics*. 1993;4(5):353–365.

37. Ariza-Ariza R, Hernandez-Cruz B, Navarro-Compan V, Leyva Pardo C, Juanola X, Navarro-Sarabia F. A comparison of telephone and paper self-completed questionnaires of main patient-related outcome measures in patients with ankylosing spondylitis and psoriatic arthritis. *Rheumatol Int*. 2013;33(11):2731–2736.
38. Bushnell DM, Reilly MC, Galani C, Martin ML, Ricci JF, Patrick DL, et al. Validation of electronic data capture of the irritable bowel syndrome–quality of life measure, the work productivity and activity impairment questionnaire for irritable bowel syndrome and the EuroQol. *Value Health*. 2006;9(2):98–105.
39. Zhang W, Bansback N, Kopec J, Anis AH. Measuring time input loss among patients with rheumatoid arthritis: validity and reliability of the valuation of lost productivity questionnaire. *J Occup Environ Med*. 2011;53(5):530–536.
40. Anema JR, Schellart AJ, Cassidy JD, Loisel P, Veerman TJ, van der Beek AJ. Can cross country differences in return-to-work after chronic occupational back pain be explained? An exploratory analysis on disability policies in a six country cohort study. *J Occup Rehabil*. 2009;19(4):419–426.
41. Pole JD, Franche RL, Hogg-Johnson S, Vidmar M, Krause N. Duration of work disability: a comparison of self-report and administrative data. *Am J Ind Med*. 2006;49(5):394–401.
42. Cullen KL, Irvin E, Collie A, Clay F, Gensby U, Jennings PA, et al. Effectiveness of workplace interventions in return-to-work for musculoskeletal, pain-related and mental health conditions: an update of the evidence and messages for practitioners. *J Occup Rehabil*. 2018;28(1):1–15. <https://doi.org/10.1007/s10926-016-9690-x>.
43. Vogel AP, Barker SJ, Young AE, Ruseckaite R, Collie A. What is return to work? An investigation into the quantification of return to work. *Int Arch Occup Environ Health*. 2011;84(6):675–682.