Enhancing return to work of cancer patients

Tamminga, S.J.

Publication date
2012

Chapter 4.

Reproducibility, validity, and responsiveness of the Work Limitation Questionnaire (WLQ) among cancer survivors

S.J. Tamminga, J.H.A.M. Verbeek, M.H.W. Frings-Dresen, and A.G.E.M. de Boer

Submitted

Abstract

Objective

To determine reproducibility, validity, and responsiveness of the Work Limitation Questionnaire (WLQ) among cancer survivors.

Study design and Setting

A cohort of 53 cancer survivors completed the WLQ and other questionnaires at baseline, 4 weeks, and at 6 months follow-up, we assessed internal consistency, Intraclass Correlation Coefficient (ICC), Standard Error of Measurement (SEM), floor- and ceiling effects, and compared the WLQ with other constructs. For responsiveness, we assessed the following anchor-based measures: Minimal Important Change (MIC) versus Smallest Detectable Change (SDC) and Area Under the Curve (AUC) of Receiver Operation Characteristic (ROC).

Results

We found sufficient reproducibility at the group level but not at the individual level. There was no indication of systematic bias or proportional bias. Internal consistency and construct validity for the WLQ and its subscales were sufficient or slightly less than sufficient. There was a floor effect for one subscale but there were no ceiling effects. Responsiveness was sufficient.

Conclusion

The WLQ is reproducible, valid, and responsive for the use at group level but it is not sufficiently reproducible to be used at an individual level among cancer survivors.

Introduction

Work disability refers to the condition of a partial or total inability to work, which may lead to unemployment, (partly) disability pension, absenteeism, or lower levels of work functioning. Several studies have addressed work disability of cancer survivors by studying the unemployment risk (e.g. de Boer et al[1]) or the time to return to work (e.g. Spelten et al[2]) but only few studies have addressed work functioning of cancer survivors.

Studies that addressed work functioning of cancer survivors found that cancer survivors had a significant lower level of work functioning compared to non-cancer controls,[3-5] or compared to patients with chronic illnesses.[6] Impaired work functioning of cancer survivors may increase the risk of lower economic well being.[7] In addition, lower levels of work functioning will increase the costs for the society and the employer.

With continuing advances in cancer treatment, it is expected that employment rates will also significantly improve. This will lead to an increased importance of studying work functioning to better understand the full impact of a cancer diagnosis on work disability. Knowledge about the measurement properties of work-functioning questionnaires is essential for the use of these questionnaires as outcome measures in studies on the effectiveness of interventions, for the use as an indicator of the economic burden of work disability due to cancer, and for the development of specific interventions aimed at improving work functioning in cancer survivors.

A commonly used measure of impaired work functioning due to ill health is the Work Limitation Questionnaire (WLQ).[8] This questionnaire measures a person's functional limitation (i.e. health status) in relation to the demands of a person's physical, psychological, and social work environment.[8] The measurement properties of the English WLQ have shown moderate to good reliability and validity in various chronic health conditions[8-11] but two reviews on the measurement properties of work-functioning questionnaires have pointed out that the measurement error (reproducibility) of the WLQ has not been determined previously.[9] [10] However, the measurement error is an important measurement property of a questionnaire that is designed for evaluating interventions.[12] The measurement error and the minimal important change of a questionnaire enable the interpretation of a change score over

time. Only change scores larger than the measurement error can be seen as a real change while change scores smaller than the measurement error cannot.[12]

To be able to use the WLQ for evaluating an intervention, the measurement error of the WLQ needs to be determined both at the group and at the individual level. A questionnaire can be reproducible for the use at group level and be useful for research projects but not at the individual level to be used for measuring changes in clinical practice. Furthermore, the WLQ has recently been translated into Dutch. In order to use the Dutch translation of the WLQ among cancer survivors, the measurement properties of the WLQ need to be determined in this specific population. Therefore, the aim of this study is to determine reproducibility, validity, and responsiveness of the Dutch translation of the WLQ among cancer survivors.

Methods

We studied a cohort of Dutch cancer survivors at baseline, 4 weeks, and at 6 months follow-up.

Participants and recruitment

In this study, we refer to cancer survivors as to individuals who have been diagnosed with cancer and were recurrence free at the time of inclusion in the study. Cancer survivors were recruited via websites of cancer patient organisations, via a database of the Academic Medical Center's surgery outpatient clinic, and via the department of gynaecology. To be included, cancer survivors had to be employed, had to be working in the past two weeks, be able to read or write Dutch, and had to be without severe co-morbidity.

If a cancer survivor was eligible and willing to participate, the first questionnaire and informed consent form were sent by mail to the cancer survivors' home with a free return envelope enclosed. The follow-up questionnaires were also sent by mail. The researcher sent reminders by mail and contacted non-responders by telephone to encourage returning questionnaires. We recruited cancer survivors from January 2010 until September 2010 via websites of cancer patient organisations and during March 2011 via the database.

Ethical approval for this study was sought from the Medical Ethics Committee of the Academic Medical Center, who judged that ethical approval was not required. Participants signed informed consent forms before they filled in the first questionnaire.

Study design

A prospective cohort study with measurements at the study entry (baseline), 4 weeks, and 6 months follow-up was conducted. Data measured at baseline were used to determine distribution of the WLQ, internal consistency, construct validity, and floor and ceiling effects (Table 1). To determine reproducibility we used the data measured at baseline and at four weeks follow-up in a population that reported no change (i.e. test-retest reproducibility and level of agreement) (Table 1). The time span of 4 weeks was chosen to prevent recall bias and cancer survivors were expected to be stable. To determine responsiveness, we used data measured at baseline and at 6 months (Table 1). The time span of 6 months was chosen to increase the chance of a clinical change. Participants did not receive any intervention as part of the study in the interim period.

Measurements

Demographic variables

The following demographic variables were assessed: age, gender, marital status, education level, cancer diagnosis, cancer treatment, time since cancer diagnosis, co-morbidity, breadwinner position, type of occupation, and time since work resumption. Based on type of occupation we divided work demands in mainly mentally demanding work and mainly physically demanding work.

Table 1. Measurement properties.

Measurement property *A. Definition* *B. Purpose (evaluation, discrimination, both)* *C. Dependency of a measurement property on the characteristics of a population*	**Prerequisite** *A. Prerequisite* *B. Determination method* *C. Adequateness*	**Method** *A. Measurement point* *B. Data analysis*	***A priori* formulated criteria** *A. Measurement property* *B. Adequateness of prerequisite*
Internal consistency A. The interrelatedness among items in a scale B. Both C. Little	NA	A. Baseline B. Chronbach's alpha	A. 0.70 - 0.95 B. NA
Reproducibility - test-retest reproducibility A. The ability to distinguish participants despite measurement error B. Discrimination C. Large	A. Stable subpopulation B. Single item external anchor C. Comparing stable subgroup with unstable subgroup on time since work resumption and WLQ	A. Baseline and 4 week follow-up B. Single measures ICC_agremeent: $\sigma^2_{participants} / (\sigma^2_{participants} + \sigma^2_{time} + \sigma^2_{residual})$	A. ICC_agreement > 0.90 individual level ICC_agreement > 0.70 group level B. Stable participants have resumed work a longer time ago and less work limitations in comparison to unstable participants
Reproducibility - level of agreement A. The ability to measure the same scores on repeated measurements B. Evaluation C. Little		A. Baseline and 4 week follow-up B. SEM_agreement: $\sqrt{(\sigma^2_{time} + \sigma^2_{residual})}$ SDC_ind: 1.96 x √2 x SEM_agreement LoA: 1.96 x $SD_{differences} \pm mean_{differences}$ Systematic bias: one-sample t-test mean_differences Proportional bias: correlation between means and differences	A. Large SDC or LoA values indicate that the questionnaire is not able to measure small changes ≠ systematic bias ≠ proportional bias B. See test-retest reproducibility

Table 1. *(Continued).*

Validity - construct validity A. Ability to relate questionnaire to other related constructs B. Both C. Large	NA	A. Baseline B. Correlation between the WLQ and work- and disease-related constructs	A. > 75% of the a priori formulated hypotheses were confirmed B. NA
Validity - floor and ceiling effects A. The ability to measure the highest and lowest score B. Both C. Large	NA	A. Baseline B. Percentage of participants that had the highest and the lowest score	A. < 15% of the population had the highest or lowest score B. NA
Responsiveness A. The ability to measure changes over time despite measurement error B. Evaluation C. Large	A. Stable and improved subpopulation B. Single item external anchor C. Correlation between external anchor and change scores	A. Baseline and 6 months follow-up B. Compare the MIC to SDC_ind and SDC_group MIC based on two methods; mean change method and ROC curve method	A. SDC_ind < MIC individual level SDC_group < MIC group level B. Correlation coefficient > 0.4
		B. The AUC of a ROC-curve	A. AUC > 0.70 B. Correlation coefficient > 0.4

Abbreviations: NA = not applicable; WLQ = Work Limitation Questionnaire; ICC = Intraclass correlation coefficient; SEM = Standard Error of Measurement; SDC = Smallest Detectable Change; LoA = Limits of Agreement of Bland and Altman plot; SD = Standard deviation; MIC = Minimal Important Change; AUC = Area Under the Curve; ROC = Receiver Operating Characteristic.

The WLQ

The WLQ consists of 25 items divided into 4 different subscales, including time management demands (5 items), physical demands (6 items), mental-interpersonal demands (9 items), and output demands (5 items).[8] Time management demands address scheduling demands, physical demands address job tasks that require bodily strength, mental-interpersonal demands address job tasks that require cognitive strength and the interaction with people on-the-job, and output demands address overall work productivity. The possible responses are: 'all of the time (100%)', 'most of the time', 'half of the time (about 50%)', 'a slight bit of the time', 'none of the time (0%)', and 'does not apply to my job'. The WLQ refers to the past two weeks. Scale responses were scored from 1 to 5 and 'does not apply to my job' was scored as missing value.[8] All subscales except physical demands were reversed and all subscale responses were normalised to scores ranging from 0 (no limitations) to 100 (highest limitations). Missing values were imputed with the personal scale mean if at least 50% of the items of a subscale were known.[8]

Other instruments

To measure construct validity, we compared the WLQ scores to the scores on the following questionnaires: overall work functioning measured on a Visual Analogue Scale (VAS), Work Ability Index (WAI), overall quality of life measured on a VAS, and the Rotterdam Symptom Checklist (RSCL). VAS overall work functioning ranged from 0 (worst possible work functioning) to 100 (highest possible work functioning) and refers to the past two weeks. VAS scales have proved valid and reliable.[13] We assessed work ability with the first three questions of the WAI. The first question evaluates current work ability compared to the life time best, the second question evaluates current physical work ability in relation to the physical job demands, and the third question evaluates current mental work ability in relation to the mental job demands. Acceptable measures for reliability and validity have been determined.[14] [15] The VAS overall quality of life ranged from 0 (worst possible quality of life) to 100 (highest possible quality of life) and refers to the past week. The single item VAS overall quality of life has shown good validity and reliability.[13] The RSCL consists of the following four subscales, physical symptom distress (23 items), psychological distress (7 items), activity

level (8 items), and overall valuation of life (1 item)[16] and refers to the past week. The RSCL proved reliable and valid in assessing quality of life of cancer patients.[16]

Stability and change

We used a single item external anchor to assess stability in work functioning of participants between baseline and 4 weeks follow-up (‘*to rate work functioning compared to 4 weeks ago’)* and to assess change in work functioning of participants between baseline and 6 months follow-up (‘*to rate work functioning compared to 6 months ago’).* Both questions were assessed on a 5-point Likert scale. We considered participants stable if they reported neither having improved nor deteriorated and the remainder of participants was considered as changed. We assessed the adequateness of the external anchor by comparing the group that reported being stable to the group that reported being unstable on time since work resumption and WLQ. We assumed that the subgroup that reported being stable has resumed work a longer time ago and had better work functioning measured on the WLQ compared to the subgroup that reported being changed.

Data analysis

Data entry was verified by means of a 20% double data entry. PASW version 18 was used for all statistical analysis. For correlation coefficients, we first tested whether variables were normally distributed with the Kolmogorov-Smirnov test of normality (cut-off p-value ≤ 0.05). We used a Pearson correlation coefficient if both variables were normally distributed and a Spearman correlation coefficient otherwise.

Internal consistency

Internal consistency was defined as the interrelatedness among items in a (sub)scale (Table 1). We determined the Cronbach’s alpha for the WLQ and its subscales and we considered a Cronbach’s alpha between 0.70-0.95 sufficient.[17] Since the factor structure and unidimensionality of the subscales were determined previously[8] we did not perform (confirmatory) factor analysis.[17]

Reproducibility

Reproducibility was determined based on the baseline and four weeks follow-up data of stable participants (Table 1).

Test-retest reproducibility: Test-retest reproducibility determines how well participants can be distinguished from each other despite measurement error (Table 1).[12] We calculated test-retest reproducibility with the single measures Intraclass Correlation Coefficient (ICC) including systematic difference, so called ICC_agreement.[12] We considered an ICC of > 0.90 sufficient for the use at individual level and we considered an ICC of > 0.70 sufficient for the use at group level.[17]

Level of agreement: Level of agreement determines the agreement between repeated measurements (Table 1).[12] We measured level of agreement with the Standard Error of Measurement (SEM) including variance between measurement points, so-called systematic differences between baseline and 4 weeks follow-up (SEM_agreement).[12] The SEM_agreement is expressed on the same scale as the questionnaire (0-100). To calculate 95% confidence interval of the SEM_agreement, the SEM_agreement was converted to the Smallest Detectable Changes (SDC). The SDC is used at an individual level (i.e. SDC_ind). For the use of the SDC at group level, the SDC_ind needs to be dived by $\sqrt{n}$. No prior values that were considered sufficient for the SDC can be proposed since it is expressed on the same scale as the questionnaire. However, large SDC_ind or SDC_group values indicate respectively that the questionnaire is not able to distinguish small changes from measurement error at an individual level or at a group level.

We constructed a so-called Bland and Altman plot to establish Limits of Agreement (LoA) (Table 1).[18] We plotted the means at baseline and at 4 weeks follow-up and the differences between these measurement points as well as the 95% LoA. Change scores outside the LoA can be considered a real change and change scores that fall within the LoA cannot be distinguished from measurement error. We checked if there was a systematic bias between baseline and 4 weeks follow-up by testing with a one-sample Student's t-test if the mean differences at these time points were statistically different from zero.[19] We also checked if there was proportional bias meaning that the measurement error varies across the range of the scores by testing if the correlation

between the means at baseline and four weeks follow-up and the differences between baseline and 4 weeks follow-up was ≤ 0.4[19] (Table 1).

Validity

No 'gold standard' was available to determine validity of the WLQ. Therefore, we assessed validity of the WLQ by means of comparing the WLQ with a reference standard, so called construct validity.[20]

Construct validity: Construct validity measures the degree to which a questionnaire demonstrates a logical relation to related constructs (Table 1).[17] To determine the construct validity we assessed correlations between the WLQ and various work-oriented constructs (i.e. VAS overall work functioning, WAI) and disease-oriented constructs (i.e. VAS overall quality of life, RSCL) and formulated *a priori* 4 hypotheses based on theoretical grounding. First, we hypothesised negative correlations of > 0.8 between the WLQ and the VAS overall work functioning. Second, we hypothesised negative correlations of > 0.6 between the WLQ and current work ability (first question of the WAI), between the WLQ physical demands subscale and physical work ability (second question of the WAI), and between the WLQ mental-interpersonal demands and mental work ability (third question of the WAI). Third, we hypothesised a negative correlation of 0.40-0.60 between the WLQ and the VAS overall quality of life. Fourth, we considered a positive correlation of 0.40-0.60 sufficient between the WLQ physical demands subscale and RSCL physical symptom distress subscale and between the WLQ mental-interpersonal subscale and the RSCL psychological distress subscale.

Floor (lowest limitations) and ceiling (highest limitations) effects: Floor and ceiling effects were determined based on the percentage of participants who had the lowest score (i.e. no work limitations) and the highest score (i.e. highest work limitations) (Table 1). A percentage of < 15% of the population who had the lowest or highest score of each (sub)scale was considered sufficient.[17]

Responsiveness

Responsiveness refers to the ability of a questionnaire to detect clinically important changes over time despite measurement error (Table 1).[17] To determine responsiveness,

we used the data at baseline and at six months follow-up of improved and stable participants only, because only 6 participants indicated that they had deteriorated on work functioning. Based on the single item external anchor, three groups were distinguished: those who reported having slightly improved or improved, those who reported being stable, and those who reported having slightly deteriorated or deteriorated.

We determined responsiveness both by comparing the SDC (see level of agreement) with the Minimal Important Change (MIC) as well as by the Area Under the Curve (AUC) of a Receiver Operating Characteristics (ROC) curve. Responsiveness of the WLQ is considered sufficient if the SDC is smaller than the MIC or if the AUC value is > 0.7.[17]

To assess the MIC, we first used the mean change method. We calculated the mean change on the WLQ between baseline and 6 months follow-up as the differences in mean change of those who reported being improved and those who reported being stable.[21] We also used the ROC-curve method to assess responsiveness and the MIC because large variations between MIC values within one questionnaire have been reported depending on the method used.[21] We plotted a ROC curve with sensitivity and 1-specificity for each change score between baseline and 6 months follow-up of the WLQ. The MIC based on the ROC-curve is defined as the optimal cut-off value of sensitivity and specificity (i.e. the value at the upper left corner of the ROC curve).[17] Responsiveness is then determined as the AUC value of the plotted ROC curve.

We assessed if the external anchor was adequate by calculating the correlation between the external anchor and the mean change score between baseline and 6 months follow-up and we considered a correlation coefficient of > 0.5 sufficient.[22]

Results

The sample consisted of 53 cancer survivors. Table 2 presents the participants' characteristics at baseline. Participants were on average 46.7 ± 7.6 years old. Eighty-seven percent of the participants were female. Additionally, 84% of the participants had an occupation that consisted mainly of mentally demanding tasks. The response rate at 4 weeks follow-up was 94% (N=50) and the response rate at 6 months follow-up was 85% (N=45). Reasons for non-response were cancer recurrence (N=1), not being able to

work prior to filling in the second or third questionnaire (N=1), and were in 6 cases unknown (N=6). Variables that we used for calculating correlations were not normally distributed. Therefore, we used the Spearman's correlation coefficient in all cases.

Table 2. Participant characteristics at baseline.

Participant characteristic*		**Population N=53**	**Subpopulation of stable participants N=34**
Demographic characteristics			
Age (years)		46.7 ± 7.6	48.1 ± 6.8
Gender (% female)		87	85
Marital status (% married or living with partner)		74	71
Education level (%)	Low	8	9
	Intermediate	52	62
	High	40	29
Clinical characteristics			
Cancer diagnosis (%)	Breast cancer	49	41
	Colon cancer	28	35
	Vulva cancer	6	6
	Cervix cancer	8	6
	Other	9	12
Cancer treatment (%)	Surgery	94	91
	Chemotherapy	72	71
	Radiotherapy	56	47
	Hormone therapy	32	27
	Other	11	9
Years since cancer diagnosis (%)	<1	17	12
	1-5	56	53
	>5	27	35
Number of co-morbidities		1 (0-9)	1 (0-9)
Work-related characteristics			
Breadwinner position (% sole or shared)		79	67
Type of occupation (%)	Public health	21	21
	Administrative	15	24
	Management	13	12
	Other	51	43
Mainly mentally demanding work (%)		84	77
Mainly physically demanding work (%)		16	23
Number of working hours according to contract (1 - 40)		32 (3-40)	30 (4-40)
Years since work resumption (%)	< 0.5 year	27	15
	0.5 year – 1 year	12	12
	1 year – 3 years	29	29
	> 3 years	33	44

Distribution of the WLQ

The mean and standard deviation of the WLQ was 22.9 ± 17.2, the score for time management demands was 26.3 ± 21.7, for physical demands 14.2 ± 15.4, for mental-interpersonal demands 23.4 ± 18.7, and 26.7 ± 20.3 for output demands. Participants reported 'does not apply to my job' 5 times (2%) for time management demands, 40 times (13%) for physical demands, 10 times (2%) for mental-interpersonal demands, and once (0.4%) in output demands. No missing values were found for time management demands, 6 for physical demands, 1 for mental-interpersonal demands, and no missing values were found for output demands.

Internal consistency

The Cronbach's alpha was 0.93 for the WLQ, for time management demands 0.78, for physical demands 0.88, for mental-interpersonal demands 0.94, output demands 0.92.

Reproducibility

Of the 53 participants, 34 participants were included in the reproducibility analyses since 3 participants were lost to follow-up and 16 participants were not stable (2 improved, 11 slightly improved, and 3 slightly deteriorated). This group had a significant longer time to work resumption and a lower score on the WLQ compared to unstable participants (data not shown).

Test-retest reproducibility: The single measures ICC_agreement for the WLQ and the subscales ranged between 0.65 and 0.74 (Table 3).

Table 3. Test-retest reproducibility of stable participants (N=34).

	Baseline	**4 weeks follow-up***
WLQ (mean ± SD)	19.1 ± 13.5	20.6 ± 11.8
ICC (95% CI)	-	**0.74 (0.54 - 0.86)**
Time management demands (mean ± SD)	21.4 ± 20.6	23.4 ± 17.5
ICC (95% CI)	-	**0.71 (0.49 - 0.84)**
Physical demands (mean ± SD)	13.3 ± 13.5	14.6 ± 14.4
ICC (95% CI)	-	0.65 (0.39 - 0.81)
Mental-interpersonal demands (mean ± SD)	18.8 ± 14.0	19.6 ± 13.7
ICC (95% CI)	-	**0.72 (0.51 - 0.85)**
Output demands (mean ± SD)	21.5 ± 14.2	25.3 ± 15.8
ICC (95% CI)	-	0.69 (0.47 – 0.84)

*ICC values that met the a priori criterion are presented in bold.

Level of agreement

The SEM_agreement and the SDC_ind for the 34 stable participants were respectively 6.50 and 18.02 for the WLQ, 10.31 and 28.58 (time management demands), 8.28 and 22.95 (physical demands), 7.26 and 20.12 (mental-interpersonal demands), and 8.40 and 23.28 (output demands). The mean differences of the sum score at baseline and 4 weeks follow-up of the WLQ did not differ statistically from zero neither did the mean differences of the subscales (p-values ranged between 0.057 - 0.67). None of the correlations between the means and the differences exceed the 0.5 (range -0.07 - 0.24). In Figure 1, the means of baseline and 4 weeks follow-up and the differences between baseline and 4 weeks follow-up of the WLQ as well as the 95% LoA were shown in a Bland and Altman plot.

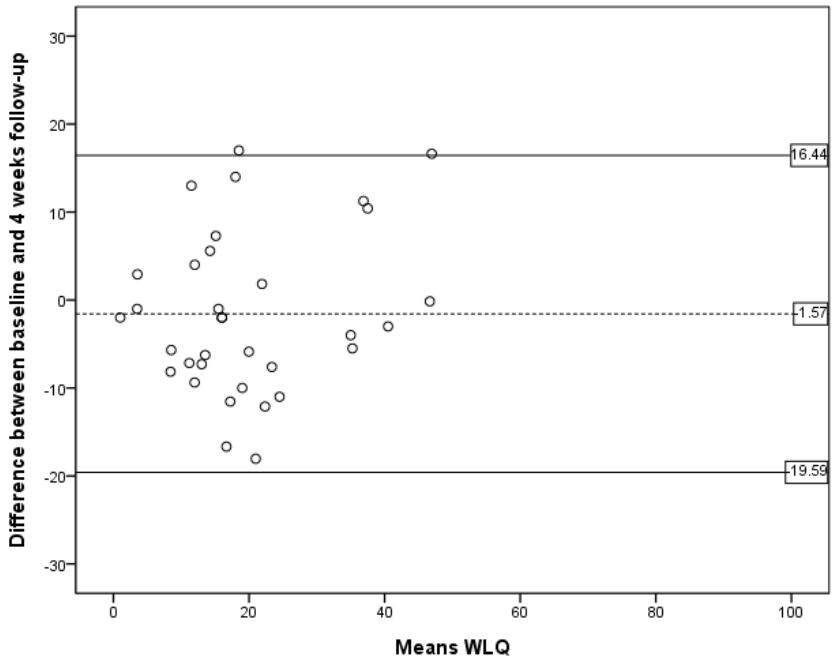

Figure 1. Bland and Altman plot of stable participants (N=34).

Validity

Correlations between the WLQ and the VAS overall work functioning ranged from -0.30 to -0.69 (Table 4), correlations between the WLQ and work ability (WAI) ranged from -0.48 to -0.77. The correlation between the WLQ subscale physical demands and physical work ability was -0.50 and the correlation between the WLQ subscale mental-interpersonal demands and mental work ability was -0.52. Correlations between the WLQ and the VAS overall quality of life ranged from -0.30 to -0.56. The correlation between the WLQ subscale physical demands and the RSCL subscale physical symptom distress was 0.45 and the correlation between the WLQ mental-interpersonal demands and the RSCL subscale psychological distress was 0.45.

Table 4. Correlation between WLQ and related constructs (N=53).

Comparable construct ***(a priori criterion)***	**WLQ**	**Spearman's correlation coefficient***	**P-value**
VAS overall work functioning (r > 0.8)	WLQ	-0.65	< 0.001
	Time management demands	-0.62	< 0.001
	Physical demands	-0.30	0.030
	Mental-interpersonal demands	-0.63	< 0.001
	Output demands	-0.69	< 0.001
Overall work ability (WAI) (r < -0.6)	WLQ	**-0.76**	< 0.001
	Time management demands	**-0.77**	< 0.001
	Physical demands	**-0.48**	< 0.001
	Mental-interpersonal demands	**-0.65**	< 0.001
	Output demands	**-0.69**	< 0.001
Physical work ability (WAI) (r < -0.6)	Physical demands	-0.50	< 0.001
Mental work ability (WAI) (r < -0.6)	Mental-interpersonal demands	-0.52	< 0.001
VAS overall quality of life (r < -0.4 – -0.6)	WLQ	**-0.49**	< 0.001
	Time management demands	**-0.56**	< 0.001
	Physical demands	-0.30	0.031
	Mental-interpersonal demands	**-0.43**	0.002
	Output demands	-0.35	0.011
Physical symptom distress scale (RSCL) (r < 0.4 – 0.6)	Physical demands	**0.45**	< 0.001
Psychological distress scale (RSCL) (r < 0.4 – 0.6)	Mental-interpersonal demands	**0.45**	< 0.001

*Correlation coefficients that met the *a priori* criterion are presented in bold.

Two participants (4%) had the lowest possible score on the WLQ, 12 participants (23%) on time management demands, 17 participants (33%) on physical demands, 6 participants (11%) on mental-interpersonal demands, and 6 participants (11%) on output demands. One participant (2%) had the highest possible score on the output demands. None of the participants had the highest possible score on the WLQ score or on any subscale.

Responsiveness

Of the 53 participants, 39 participants were included in the responsive analysis since 9 participants were lost to follow-up and 6 participants reported having deteriorated. Of these 39 participants, 19 participants reported being stable and 21 participants reported having slightly improved or improved. The MIC for improvement of the WLQ based on the mean changed method was 4.2 and the MIC for improvement of the WLQ based on the ROC-curve method was 4.0 (sensitivity 71% and specificity 31%). These MIC values did not exceed the SDC_ind (18.02) but did exceed the SDC_group (3.09) of the WLQ. The AUC-value of the ROC curve was 0.68 (Figure 2). The correlation between the external anchor and the mean change score between baseline and 6 months follow-up was 0.49.

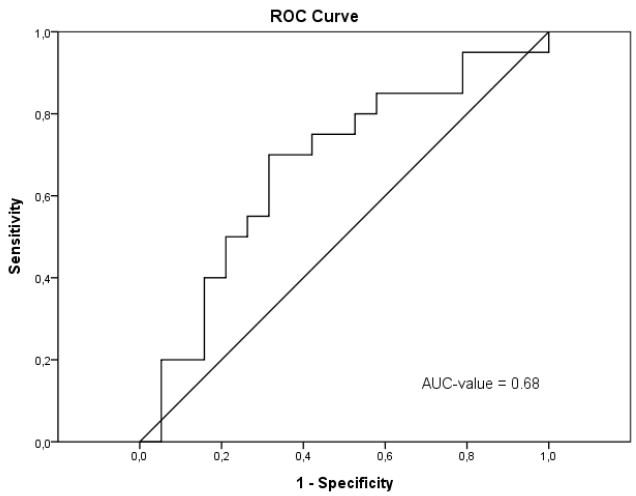

Figure 2. ROC-curve of participants who reported being stable or improved (N=39).

Discussion

We found sufficient reproducibility at the group level but not at the individual level. There was no indication of systematic bias or proportional bias. Internal consistency and construct validity for the WLQ and its subscales were sufficient or slightly less than sufficient. There was a floor effect for one subscale but there were no ceiling effects. Responsiveness was sufficient.

Strengths and limitations

The strength of our study is the determination of the measurement error (SEM) and SDC as these important characteristics of reproducibility were not determined for the original version of the WLQ in any population or any cross-cultural translation.[9 10] We determined both the SDC for clinical or individual use and for group and research use and we used several complementary methods to determine the minimal important change (MIC) for patients. Another strength is the inclusion of patients who worked less than 20 hours a week as the measurement properties for these employees were not determined in the original study by Lerner et al.[8] We were able to include a varied sample of cancer survivors that showed both stability and improvement over time, which is a prerequisite for determining reproducibility and responsiveness.

A limitation of our study is the small sample size for reproducibility analyses caused by a rather high proportion of participants who reported being unstable between baseline and 4 weeks follow-up. This is probably due to those participants who resumed work just before completing the questionnaire and whose work functioning was apparently still improving. However, we intended to determine measurement properties in a varied population because the WLQ will most likely be used in this population. Even though we used an external anchor that consisted of one item only, where the use of multiple items may be more adequate, we could show that the external anchor was adequate for both reproducibility and responsiveness analysis.

Reproducibility

Test-retest reproducibility analysis of the WLQ showed sufficient ICC values for the use at group level and almost sufficient ICC values for physical demands (0.65) and output

demands (0.69). These findings are in accordance with the original study by Lerner et al.[8] The ICC values are however insufficient for the use at individual level.

Level of agreement analysis showed large values for SEM_agreement (6.50 - 10.31) and for SDC_ind (18.02 - 28.58) compared to the range of the scale (0 - 100). The Bland and Altman plot showed equally large LoA values. The moderate level of agreement of the WLQ could be due to weakness in our study such as lack of stability of participants. We do not think that this is the case because we could show that the external anchor measured stability adequately. Taking into account recall bias we also believe that the time span of 4 weeks to measure stability was also appropriate. The lack of sufficient level of agreement could also be due to the cross-cultural adaptations but due to lack of comparable cross-cultural adaptations studies this is difficult to shown. Even though most cross-cultural adaptations affect more often the validity of a questionnaire[23] in this case, the level of agreement might be affected by differences in work culture and legalisation of sick-listed employees. In the Netherlands, employees work substantially more often part-time compared to other countries. The answer categories of the WLQ are related to a percentage of the working time, which may be more difficult to fill in when working part-time. To avoid this problem in future studies, we suggest using answer categories that are not related to time or percentage. Again another explanation could be that the properties of the scale itself led to the moderate level of agreement of the WLQ. Streiner and Norman[24] state that improving the scale design reduces the residual variance ($\sigma^2_{residual}$), which in turn leads to improved levels of agreement. That the answer categories of the WLQ can problematic is indicated by our finding that some participants mentioned that some questions were difficult to fill in. Beaton et al and Roy et al also suggested that problems with the physical demands subscale are caused by the fact that it is the only subscale that has reversed answer categories.[11 25] Therefore, we believe that reversing the answer categories of this subscale can lead to improved reproducibility. Furthermore, Walker et al suggest that problems with the subscale physical demands can be caused by the large number of 'does not apply to my job' and the consequent imputation with the scale mean while imputation with 'no limitations' may be more adequate.[26] Also in our study for more than 50% of 'does not apply to my job' answers, a participant was not consistent in reporting the same answer category at either baseline or 4 weeks follow-up (data not shown). Therefore, we suggest that when

an item does not apply to someone's job it needs to be filled in as 'no limitations' or imputed with 'no limitations' instead of the scale mean. These improvements of the scale design probably will lead to improved reproducibility.

Validity

The WLQ showed moderate to good construct validity. As expected, we found overall better construct validity for work-oriented constructs than for disease-oriented constructs. Furthermore, we found only for the VAS overall work functioning that the majority of the correlations coefficients were below what we had hypothesised. We assumed good validity for the VAS overall work functioning because VAS scales have been used in various situations and have shown good validity.[13] However, the construct validity of the VAS overall work functioning was not determined previously and it may explain the lower correlation, indicating that the VAS overall work functioning measures a slightly different work functioning construct than the WLQ. In contrast, we found the expected correlation between overall work ability (WAI) and the WLQ. This finding gives support for the work-oriented construct validity of the WLQ. We also found insufficient correlations for the subscale physical demands. This might be due to the lack of variation on this subscale in this specific population.

The floor effects and 'did not apply to my job' values that were found for the physical demands subscale are an indication that these questions are either not relevant for this population or do not address physical demands adequately. This finding has been reported previously.[8 11]

Responsiveness

Responsiveness analysis based on the AUC of the ROC-curve indicated that responsiveness of the WLQ was moderate. We found and AUC-value of 0.68 while 0.7 is considered sufficient. Roy et al found similar responsiveness results based on the AUC value (0.72) of the WLQ among patients with rheumatoid arthritis but found a higher MIC value (13) based on the ROC curve method.[25] The MIC value is influenced to a great extent by a population and should therefore be determined for each population separately.[27]

Generalizability

The scores on the RSCL were comparable to normative data of cancer survivors who were disease free for more than three years and were worse than in the general population (data not shown).[28] The WLQ-index score of our sample of 6.7 ± 5.0 was comparable to work limitations found in other studies among breast cancer survivors (5.5 ± 4.0)[3] and brain tumor survivors (5.6 ± 4.4)[4] and worse work functioning in comparison to non-cancer controls 2.8 ± 2.7.[4] Therefore, we think that our findings are generalizable to cancer survivors in general. However, in our sample there were more cancer survivors with less physically demanding jobs and a high education level. This means that the findings may not apply to cancer survivors with physically demanding jobs and low education. It would be worthwhile to study the measurement properties also in this population.

Implications for research and practice

The WLQ can be useful instrument for use in a population of cancer patients in evaluating interventions in research projects but its measurement properties should be improved for the subscale physical demands. For clinical practice, the measurement error is too large to measure change that is important to cancer patients.

Acknowledgement

We would like to thank the cancer survivors for their participation. We would also like to thank the Dutch cancer patient organisation (NFK), the Dutch breast cancer organisation (BVN), and the Dutch breast cancer organisation for young breast cancer survivors (stichting Amazones) for the possibility to post a notice on their website and we would like to thank Prof. dr. J.H.G. Klinkenbijl and T.W. Klinge for their help in recruiting cancer survivors. The study is granted by the Stichting Insituut Gak and is part of the research programme “Pathways to work” (www.verbeteronderzoek.nl).

References

1. **De Boer A,** Taskila T, Ojajärvi A et al. Cancer survivors and unemployment - A meta-analysis and meta-regression. *JAMA* 2009;301:753-62.
2. **Spelten ER,** Sprangers MAG, Verbeek JHAM. Factors reported to influence the return to work of cancer survivors: a literature review. *Psycho-oncology* 2002;11:124-31.
3. **Hansen JA,** Feuerstein M, Calvio LC et al. Breast cancer survivors at work. *J Occup Environ Med* 2008;50:777-84.
4. **Feuerstein M,** Hansen JA, Calvio LC et al. Work productivity in brain tumor survivors. *J Occup Environ Med* 2007;49:803-11.
5. **Lavigne JE,** Griggs JJ, Tu XM et al. Hot flashes, fatigue, treatment exposures and work productivity in breast cancer survivors. *J Cancer Surviv* 2008;2:296-302.
6. **Kessler RC,** Greenberg PE, Mickelson KD et al. The effects of chronic medical conditions on work loss and work cutback. *J Occup Environ Med* 2001;43:218-25.
7. **Short PF,** Vasey JJ, Belue R. Work disability associated with cancer survivorship and other chronic conditions. *Psychooncology* 2008;17:91-7.
8. **Lerner D,** Amick BC, Rogers WH et al. The Work Limitations Questionnaire. *Med Care* 2001;39:72-85.
9. **Roy JS,** Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for musculoskeletal disorders: a systematic review. *J Rehabil Med* 2011;43:23-31.
10. **Abma FI,** van der Klink JJ, Terwee CB et al. Evaluation of the measurement properties of self-reported health-related work-functioning instruments among workers with common mental disorders. *Scand J Work Environ Health* 2011.
11. **Beaton DE,** Tang K, Gignac MA et al. Reliability, validity, and responsiveness of five at-work productivity measures in patients with rheumatoid arthritis or osteoarthritis. *Arthritis Care Res (Hoboken)* 2010;62:28-37.
12. **de Vet HC,** Terwee CB, Knol DL et al. When to use agreement versus reliability measures. *J Clin Epidemiol* 2006;59:1033-9.
13. **de Boer AG,** van Lanschot JJ, Stalmeier PF et al. Is a single-item visual analogue scale as valid, reliable and responsive as multi-item scales in measuring quality of life? *Qual Life Res* 2004;13:311-20.
14. **Ilmarinen J,** Tuomi K. Work ability of aging workers. *Scand J Work Environ Health* 1992;18 Suppl 2:8-10.
15. **de Zwart BC,** Frings-Dresen MH, van Duivenbooden JC. Test-retest reliability of the Work Ability Index questionnaire. *Occup Med* 2002;52:177-81.
16. **de Haes JC,** van Knippenberg FC, Neijt JP. Measuring psychological and physical distress in cancer patients: structure and application of the Rotterdam Symptom Checklist. *Br J Cancer* 1990;62:1034-8.
17. **Terwee CB,** Bot SD, de Boer MR et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;60:34-42.
18. **Bland JM,** Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307-10.
19. **Mantha S,** Roizen MF, Fleisher LA et al. Comparing methods of clinical measurement: reporting standards for bland and altman analysis. *Anesth Analg* 2000;90:593-602.
20. **Rutjes AW,** Reitsma JB, Coomarasamy A et al. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess* 2007;11:iii, ix-51.
21. **Terwee CB,** Roorda LD, Dekker J et al. Mind the MIC: large variation among populations and methods. *J Clin Epidemiol* 2010;63:524-34.
22. **Liang MH.** Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. *Med Care* 2000;38:II84-II90.
23. **Schellingerhout JM,** Heymans MW, Verhagen AP et al. Measurement properties of translated versions of neck-specific questionnaires: a systematic review. *BMC Med Res Methodol* 2011;11:87.
24. **Streiner DL,** Norman GR. Health measurement scales: a practical guide to their development and use. 4th ed. Oxford: Oxford University Press, 2008.
25. **Roy JS,** MacDermid JC, Amick BC, III et al. Validity and responsiveness of presenteeism scales in

chronic work-related upper-extremity disorders. *Phys Ther* 2011;91:254-66.
26. **Walker N,** Michaud K, Wolfe F. Work limitations among working persons with rheumatoid arthritis: results, reliability, and validity of the work limitations questionnaire in 836 patients. *J Rheumatol* 2005;32:1006-12.
27. **Revicki D,** Hays RD, Cella D et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008;61:102-9.
28. **de Haes JC,** Olschewski M, Fayers P et al. Measuring the quality of life of cancer patients with the Rotterdam Symptom Checklist (RSCL), a manual. 1996. Groningen, the Netherlands, Northeren Centre sfor Healthcare Research.