Quality of Life Research (2005) 14: 1869–1879
DOI 10.1007/s11136-005-3865-9

Beyond return to work: Testing a measure of at-work disability in workers with musculoskeletal pain

Dorcas E. Beaton[1,2,3,4], Carol A. Kennedy[1,2,3,5] & the Workplace upper extremity research group
[1]*Institute for Work and Health;* [2]*Mobility Program Clinical Research Unit, St Michael's Hospital (E-mail: beatond@smh.toronto.on.ca);* [3]*Martin Family Arthritis Care and Research Centre, St Michael's Hospital;* [4]*Department of Occupational Therapy, Graduate Departments of Health Policy, Management and Evaluation and Rehabilitation Sciences, University of Toronto;* [5]*Department of Physical Therapy, University of Toronto, Toronto, Ontario, Canada*

Accepted in revised form 10 March 2005

Abstract

Background: There are a limited number of validated measures of a patient's perception of the level of difficulty that they are having at their job. The purpose of this study was to evaluate the psychometric properties of the 16-item Work Limitations Questionnaire (WLQ-16). *Methods*: A sample of 42 workers reporting to a workplace occupational health unit with upper limb or low back pain were enrolled in an observational study. Participants were assessed at baseline, 4 and 12 weeks post reporting. Psychometric testing (distributions, Cronbach's alpha, construct validity and responsiveness to change in problem and pain) was done using the baseline and 12-week data. *Results*: The WLQ-16 had evidence of internal consistency, construct validity and responsiveness. Some ceiling effect was found in the domains of mental-interpersonal and output demands. Physical demands suffered from missing values in 18/42 due to not-applicable content. Construct validity revealed that there was less discrimination at the higher (less limited) end of the scale. Responsiveness was present, though less than found with other measures of function and pain. *Conclusions*: The WLQ-16 shows promise as a measure of at-work disability. Further testing to evaluate the ceiling effect and responsiveness to constructs of change more closely related to work disability is recommended.

Key words: Outcomes, Reliability, Reproducibility of results, Responsiveness, Validity, Work disability

Introduction

Health-related loss of work productivity is increasingly recognized as an important aspect of the critical interface of the effect of health disorders on the economy [1–5]. Traditionally, we have considered presence or absence from work as an indicator of lost productivity that, while important, is only part of the picture. In some disorders, including musculoskeletal conditions, many workers are able to retain at least some level of paid employment but may not be able to work at full capacity. Stewart [1] reports that more lost productivity costs can be attributed to people still at work, rather than to those absent from work. The ability to capture this "at-work" disability is necessary in order to obtain a full picture of lost productivity to monitor the effect of a disorder on a person, as well as to estimate the indirect costs of an illness in economic analyses [6, 7]. Valid and reliable measures of work ability are needed [1].

At-work disability, or "presenteeism" [1, 7, 8], is measurable by objective measures, but most common are self-reports from the workers. There has been an upsurge in the number of measures available including instruments such as the Stanford Presenteeism Scale [8], the Work Instability Scale [2] or the Work Performance Questionnaire [9]. Recent reviews have examined the content and development of these measures. The reviews also

revealed that few instruments have been suitably tested for reliability and validity in musculoskeletal disorders, likely because this is a relatively new field of study [4–7, 10].

One instrument found in most reviews is the Work Limitations Questionnaire (WLQ) [11]. It is a self-report measure of work productivity loss, and is one of the few that focuses only on presenteeism, and that asks about the *proportion of work-time with limitations* rather than the degree of difficulty or severity of the limitation. As such, it could be converted into a monetary value for an economic analysis [12, 13]. There are several versions of the WLQ. At the time we began our study the developers shared a 16-item version being developed for a study of carpal tunnel syndrome (WLQ-16). It was later reduced to 15 items in that study and called the Work Role Functioning (WRF) Scale [14]. Responses on the WLQ-16 are on a five-point scale (ranging from none of the time to all of the time), including "not relevant to my job" option as well (treated as missing data). The various domains have differing numbers of items: physical demands ($n = 4$), time management ($n = 2$), output demands ($n = 4$), and mental/interpersonal ($n = 6$).

As mentioned, the WRF has one less item than the WLQ-16. It is the item on "doing more than one task" from the mental-interpersonal domain. The WRF scale also subdivides the mental-interpersonal domain into mental ($n = 2$) and social demands ($n = 3$) where these, plus the additional item, are combined into mental-interpersonal in the WLQ-16. Like the WRF, we retained a higher score (out of 100) as being better and indicative of less limitations for comparability. More recent WLQ versions have reversed the scores such that a score of 100 indicating more limitations[12, 15]. Previous publications have described the construct validity and internal consistency of other versions of the Work Limitations Questionnaire [11, 12, 16]. No articles were found that focused on the measurement properties of the WLQ-16. Only one article was found that used the very similar WRF, but it did not specifically describe the measurement properties [14].

The purpose of this project was to assess the internal consistency, construct validity and responsiveness of the 16-item version of the Work Limitations Questionnaire in a convenience sample of workers reporting musculoskeletal pain to their workplace.

Methods

We recruited participants from those workers visiting the occupational health unit at a large urban newspaper with a new episode (not having reported this pain in last three months) of soft tissue pain or discomfort in the upper extremity or back. The population included editorial (writing and editing), advertising, circulation, and human resources/finance operations staff, but did not include the press operation who were located at another facility. Thus, our sample was in largely computer-oriented occupations. We excluded persons with any associated traumatic injury (laceration, fracture) or persons unable to complete a questionnaire in English (based on self-report). No restrictions were placed on the type of care after enrollment. Persons meeting these screening criteria were considered eligible.

The occupational health nurse provided eligible participants with a package that included a description of the study, a consent form, and a questionnaire package. She then asked if a research coordinator could call them to discuss participating in the study. Those agreeing were identified and a phone number recorded. Study personnel began contacting them at their work number within five days and continued for up to three weeks. At three weeks we felt we were no longer measuring true baseline, and considered them as not recruited. Those who did return a completed baseline questionnaire within the three-week period were later mailed a second one-month package and then a final 12-week package. Twelve weeks was chosen because previous studies have shown that many people with musculoskeletal disorders improve within that timeframe (low-back pain [17], neck pain [18–20] and compensated upper-limb soft tissue disorders [21]) and we believed they would therefore have measurable functional change as well.

Measures

At baseline, we gathered demographic information and information on the location and severity of the respondent's symptoms. At baseline and at each follow-up, each participant completed the WLQ-16 (described above) as well as measures of their disability (QuickDASH Outcome Measure for

upper-limb disorders [22] (www.dash.iwh.on.ca) or Roland–Morris Scale for low back pain [23, 24]), self-efficacy (Lorig's pain and coping subscales only) [25] and rating of the severity of their problem (0–7, 7 = very severe). The QuickDASH and Roland–Morris scales were included to allow a description of overall disability as well as at-work disability. The self-efficacy scale was used as a construct of perceived ability to cope with pain which we hypothesized would be related to fewer limitations at work. At each follow-up we added two items asking how much their pain and their overall problem had changed (0–10, 5 = no change, 10 = much better). In the subset of persons with upper-limb pain, the optional four-item work module of the DASH/QuickDASH was also completed in the subset of persons with upper limb pain (0–100, 100 = unable to work). This provided another measure of difficulties encountered at work which we thought should be associated with the WLQ scores.

Other standardized measures and prognostic factors were included in the questionnaire, but were not used in the psychometric analyses presented in this paper.

Analysis

Sample description

We used univariate statistics (mean, standard deviations, frequency counts) to describe participants in this study in terms of their demographics and work disability.

Item and scale level analysis

We then did an item and scale level analysis of the WLQ-16. To do this, we described the response pattern to each item and also calculated the item to total correlations within each subscale. Items were expected to be correlated more than $r = 0.4$ with their respective subscale scores. The scale scores were created and the distribution examined for ceiling and floor effects – defined as more than the 15% at the minimum or maximum score [26]. We calculated Cronbach alpha coefficients to estimate the internal consistency and precision of measurement within each subscale and for the total score [27].

Construct validity

In assessing construct validity we sought to test if the scale was measuring what it purported to measure [28]. We evaluated construct validity by comparing WLQ-16 scores with other indicators of the impact of the disorder. First, we correlated the WLQ-16 scores with two validated region-specific measures of disability (Roland Morris scale for low back pain [23, 24, 29–31] and the shortened version of the Disabilities of the Arm, Shoulder and Hand (QuickDASH) for upper-limb pain [32–34]. Our *a priori* hypothesis was that we could consider a moderate (rho > 0.50) correlation indicative of construct validity.

The optional work module of the DASH (four items) and the single item score on ability to do daily activities were also correlated against the work-limitations questionnaires in those who were experiencing some upper-limb problem. Again, at least moderate correlations were expected. In addition, in those with upper-limb symptoms, we used a single item from the DASH that asked about the respondent's degree of difficulty doing their usual work. This was rated on a five-point scale and we did an analysis of variance to evaluate if the WLQ-16 scores had logical and statistically significant gradients across perceived difficulty at work.

Responsiveness

Responsiveness is the ability of an instrument to accurately detect change when it has occurred [35]. The first step is establishing groups of participants who had indeed experienced a change. Change scores on the instrument in this subgroup should therefore also change. We used anchor-based approaches to assess responsiveness [36–38]. At follow-up, participants completed two items that were transitional questions asking them if they had experienced a change in: (a) their overall problem or (b) their pain intensity between baseline and their 12-week follow-up. These became our indicators that a perceived change had occurred.

We then used two analytic approaches as described by Deyo [39]. The first is the correlational approach. Spearman ranked correlations were calculated comparing the subscales of the WLQ-16 and the two transitional indices of self perceived

change in problem and in pain. At least moderate correlations would indicate associations between a change in score and the transitional index. The second approach was statistical. We looked for ways of isolating subjects who had improved (or not) in their problem. We again used the transitional indices of change in pain and in overall problem but this time created a threshold that indicated if an individual was "improved." We used a score of greater than six out of ten on the transitional scale where five was no change and ten is much better. We then segregated the sample of "improved" and "not improved" according to this guideline and calculated an effect size statistic for that subgroup. According to Cohen's guidelines for paired data [40], the mean change in the score was divided by the standard deviation of the change in the score - which is equivalent to what is also called the Standardized Response Mean (SRM) [41, 42]. We hypothesized that a responsive measure would have at least a small (0.2) effect size in those deemed improved, and that the effect size would be larger than that found in the unimproved subgroup.

Differences in the SRM were calculated as was the ratio of the $SRM_{improved}/SRM_{not\ improved}$. We anticipated that most persons would experience improvement in their condition during the study time, hence we dichotomized the transitional scales only into improved versus not improved. Not improved was therefore either no change or worse. Numbers would not justify trying to estimate sensitivity to deterioration, though this would be important in future studies.

Responsiveness statistics for the WLQ-16 subscales were then contrasted with those for two other scales used in the study: functional status (either QuickDASH or Roland–Morris depending on body region affected) and Lorig's self-efficacy for pain management. These additional responsiveness statistics were used to help anchor the amount of change being detected on the WLQ-16 subscales relative to these more commonly used indicators.

Results

Sample description

One hundred and twenty-seven people reported to the occupational health nurse with an eligible disorder. Forty-five were contacted, agreed to participate, and returned a package. The greatest attrition was the inability to contact these persons, and to obtain a baseline questionnaire within three weeks of reporting, our a priori cut-off for a valid baseline report. Only three were lost to follow-up over the 12 week period leaving us with 42 for analysis, adequate for our correlational study.

Of the 42 participants, 23 were female. The average age was 42.5 years (standard deviation 10.1). The majority ($n = 25$) had only one area in pain. The most common diagnoses were repetitive strain injury or RSI ($n = 12$), tension neck ($n = 9$), and rotator cuff tendonitis ($n = 8$). On average, the group improved with their overall problem (mean = 7.89/10) and their pain (mean = 7.89/10) over the course of the study (see Table 1). For about half, it was their first episode of such pain. Only two respondents had lost time from work due to their presenting problem, suggesting that this is a group still continuing to work. Hence we were measuring work limitations in the "at-work" population.

Baseline scores

The responses to each item in the four subscales of the WLQ-16 are summarized in Table 2. Distributions are skewed with more people responding in the higher categories (no time with limitations). The lifting and carrying item in physical demands had a high number of missing data because of the high endorsement of the "not relevant for my job" option. Item to scale correlations were moderate in time management and physical demands (0.59–0.79) and higher (> 0.75) in the output demands and mental-interpersonal subscales. It is generally recommended that these exceed a value of 0.4.

The baseline summary statistics (mean, standard deviation, etc.) for each subscale are shown in Table 3. Each domain had at least six persons with missing scores, usually due to missed items, or non-relevant items. All scales used the full breadth of the scale with the exception of physical demands (0–87.5), and all had equivalent variances. Output demands and mental-interpersonal subscales had the highest mean scores (84.4 and 82.2, respectively) and highest numbers at ceiling (maximum score, 17 and 10, respectively) suggesting limitations were experienced less frequently in these

Table 1. Descriptive statistics on the indicators of change used as constructs for the responsiveness study

Construct for change	Distribution in sample	Dichotomization for responsiveness analysis	Number considered "improved"
Self-rated change (Change in problem)	Mean: 7.89	>6 on an 11 point numeric rating scale of change in overall rating of their upper limb or low back problem.	Yes: 28
	Standard Deviation: 2.60 Median: 9 25th percentile 7 75th percentile 10		No: 9 missing: 8
Self-rated change (Change in pain)	Mean: 7.89	>6 on 11 point scale of change in pain	Yes: 29
	Standard Deviation: 2.59 Median: 9 25th percentile 7 75th percentile 10		No: 8 missing: 8

Change index was 0–10 point transitional scale where 5 = no change, and 10 = much better >6 was considered indicative of a perceived improvement.

Table 2. Item level responses to the Work Limitations Questions (16 item version) at baseline presentation to the occupational health unit with new disorder. Domain and number of scale scores missing are shown to the left

WLQ-16 Item		Sub-Scale*	Missing or n/a	Response: % of time having difficulty doing item					Mean (1–5 Scale)	Item to sub Scale correlation
				1 100%	2	3 50%	4	5 0%		
1	Sticking to schedule	tm	5	1	4	7	11	14	3.89	0.59
2	Not taking extra breaks	tm	4	0	8	11	8	11	3.58	0.59
3	Lifting carrying	pd	18	3	3	8	7	3	3.17	0.69
4	Bending, twisting	pd	8	3	9	8	7	7	3.18	0.79
5	Using hand tools	pd	5	3	4	9	14	7	3.49	0.66
6	One position	pd	3	6	9	8	8	8	3.08	0.72
7	More than one task	mi	3	1	0	10	6	22	4.23	0.87
8	Concentrating	mi	3	2	3	4	15	15	3.97	0.87
9	Remembering things	mi	2	1	0	5	5	29	4.53	0.89
10	Talking with people	mi	5	1	2	5	6	23	4.30	0.92
11	Helping others	mi	9	1	1	2	6	23	4.48	0.92
12	Controlling irritability	mi	4	1	0	1	9	27	4.61	0.77
13	No mistakes	od	3	1	1	1	11	25	4.49	0.82
14	Satisfying judge of work	od	2	1	1	5	5	28	4.45	0.93
15	Finishing all your work	od	3	1	2	1	10	25	4.43	0.87
16	Feeling sense of accomplishment	od	5	1	4	3	6	23	4.24	0.85

* Subscales: tm: time management, pd: physical demands, mi: mental interpersonal, od: output demands
Frequency of use of various response options, item scores per item, and item to domain total correlations are shown.

Table 3. Work Limitations Questionnaire-16. Mean, standard deviation for different subscales at baseline

Subscale	n usable*	Mean (sd)	Minimum, maximum score	n at floor (0)	n at Ceiling (100)
Physical demands	22	54.8 (26.5)	0–87.5	2	0
Output demands	35	84.4 (23.7)	6.25–100	0	17
Time management	36	69.1 (24.9)	12.5–100	0	7
Mental-interpersonal	29	82.2 (24.1)	0–100	1	10

* number of usable responses deletes those who had missing values, or answered that an item was answered using response option of ''not part of my job'' which was coded as missing. One item, 'Lifting heavy objects' in the physical demands subscale was missing or not relevant in 18 of 42 cases.
Each dimension is scored from 0 (difficult all the time) to 100 (difficult none of the time).

domains. McHorney and Tarlov [26] suggest that the number of workers at the floor or ceiling should be less than 15% of the sample (n = six in our data) [26]. Time management is just over this level (19%). None of the subscale scores had floor effects greater than 15% of the sample. Scales were internally consistent (see Cronbach alpha statistics >0.86 on diagonal of Table 4), with a lower value found in the Time Management domain (alpha = 0.74) likely because it only has two items [43].

Construct validity

Correlations between the four subscales of the WLQ-16 were moderate for physical and output demands ($r = 0.62$) and high ($> \sim 0.75$) for others (see Table 4). The four dimensions of the WLQ-16 were correlated with the QuickDASH or the Roland–Morris questionnaire in a manner consistent with our a priori hypothesis of a moderate relationship ($r \sim 0.50$ or more) between functional status of these measures and the work-limitations subscales. The lowest correlation was between the QuickDASH and the output demands subscale ($r = 0.48$) but this was still close to 0.50. Correlations between self-rated ability to do usual activities and the WLQ-16 subscales were also moderate. For those subjects with upper-limb disorders, the optional work module correlated strongly ($r > 0.75$) with each dimension of the WLQ-16 except for output demands ($r = -0.62$). These results are summarized in the bottom half of Table 4.

In Table 5 the mean subscale scores are shown stratified by the response to the DASH item on perceived difficulties in doing usual work or other daily activities. This item, answered only by those

Table 4. Correlational Matrix

	Physical demands	Output demands	Time management	Mental/interpersonal
Physical demands	Alpha = 0.86	xxxx	xxxx	xxxx
Output demands	0.62	alpha = 0.94	xxxx	xxxx
Time management	0.83	0.73	alpha = 0.74	xxxx
Mental-interpersonal	0.82	0.88	0.81	Alpha = 0.96
Construct QuickDASH (n = 36)*	−0.78	−0.48	−0.70	−0.61
Roland Morris (n = 18)	−0.71	−0.64	−0.66	−0.77
Self-rating ADL (n = 42)	−0.52	−0.61	−0.64	−0.56
Work module (n = 38) (DASH)	−0.86	−0.62	−0.80	−0.76

alpha = Cronbach's alpha coefficient.
* n refers to the number of participants with this scale available. Only persons with back pain completed the Roland Morris Questionnaire, and only those with upper limb disorders completed the QuickDASH and work module. Some people had both back and upper limb pain and completed both.
The top half of the table shows correlations between subscales on the Work Limitations Questionnaire with Cronbach's Alpha coefficient along the diagonal. The bottom half of the table shows correlations of each subscale with constructs hypothesized to represent effect of a work-related disorder. Moderate correlations were expected ($r \sim 0.5$).

Table 5. Mean values for each Work Limitations Questionnaire-16 subscale across perceived level of limitations in doing work activities (DASH item). Done in the upper-limb patients only

WLQ-16 Subscale	Level of perceived work/activity limitations				*F* Statistic *P* value
	No limitation	Slightly limited	Moderately limited	Very limited	
Physical demands	82.8 [A]	64.5 [A,B]	44.5 [B,C]	26.5 [C]	6.54; $P = 0.005$
Output demands	98.1 [A]	95 [A]	85.8 [A]	48.9 [B]	12.2; $P < 0.0001$
Time management	87.5 [A]	75 [A]	61.3 [B]	27.5 [C]	18.8; $P < 0.0001$
Mental-interpersonal	98.1 [A]	95.8 [A]	81.3 [A]	44.2 [B]	17.9; $P < 0.0001$

*bracketed [ ] letters following the mean scores are the Duncan Post-hoc groupings. Levels of limitations sharing the same letter are not significantly different from each other but differ from scores in levels followed by a different letter.

with upper-limb pain, was rated on a scale of one (no limitation) to five (unable). None of the 36 respondents used the fifth response category, so the fourth (very limited) was the maximum. The subscale scores all showed a logical and statistically significant gradient across the levels of perceived limitations, especially for response category four (mean physical demands scale = 82.8 for no limitations, 26.5 for very limited).

Responsiveness

Correlational approach. The correlational approach to responsiveness failed to produce statistically significant Spearman ranked coefficients [39]. Only one correlation, between the WLQ-16 mental-interpersonal subscale and change in pain had even borderline significance ($r = 0.37$, $P = 0.07$), though it was at the magnitude expected. The lack of significant correlations could be due to the fact that few of our persons were having a great deal of work limitations, and therefore there could be a lack of variability in the rankings.

Statistical approach.
The effect sizes (mean change divided by standard deviation of change) for those deemed to have improved or not according to an external standard [36, 39] are summarized in Table 6. The SRM values are consistently higher for those improved over not improved, for example 0.81 versus 0.35 for the physical demands domain (difference in SRM: 0.46, ratio of SRM improved/not improved groups: 2.31) supporting that the WLQ-16 subscales were validly measuring change when it occurred and discriminating that from no change/worse. The SRM values for the QuickDASH and the self-efficacy for pain management scales were higher in magnitude and more discriminating (improved versus not improved) than those of the WLQ-16 subscales. This type of responsiveness is closely linked to the anchors used. In our situation, the QuickDASH and self-efficacy were likely closer conceptually to the transitional indices used as anchors than the WLQ-16 subscales.

Discussion

Questionnaires need to be tested for their reliability, validity and responsiveness in the relevant populations. This study evaluated the internal consistency, item responses, validity and responsiveness of the WLQ-16 in 42 workers presenting to a workplace occupational health centre for pain or soft-tissue discomfort of the upper arm and/or low back. The WLQ-16 has been designed to describe and monitor difficulties experienced by those workers with pain, but who are still able to be at work – the concept of presenteeism. Presenteeism indicators are sensitive to the disability burden still borne by those people who would have otherwise been labeled as "successes" or as "healthy" when using traditional markers of absenteeism or return to work.

This study supported most of the WLQ-16 subscale psychometrics in this population of workers in a largely computer-based environment. Performance of this scale may have been different, likely better, in a more physically varied or demanding job. In our sample the Cronbach alpha coefficients were moderate to very high, the correlations with disability and scales were good, and the instrument was able to show fairly large effect sizes for this situation where only small changes

Table 6. Standardized response mean (SRM = mean change score/standard deviation of the change) of different constructs for change and different subscales of the Work Limitations Questionnaire

Construct of change	Improved problem?				Improved in pain intensity?			
	No	Yes	Diff.	Ratio	No	Yes	Diff.	Ratio
Physical demands (n = 14)	0.35	0.81	0.46	2.31	0.10	0.85	0.75	8.5
Output demands	0.23	0.59	0.36	2.56	0.23	0.59	0.36	2.6
Time management	0.14	0.72	0.58	5.14	0.08	0.79	0.71	9.9
Mental-interpersonal	0.27	0.74	0.47	2.7	0.15	0.75	0.60	5.0
Quick DASH	0.13	1.10	0.97	8.46	0.48	1.39	0.91	2.9
Self-efficacy (pain)	0.29	1.16	0.87	4	0.12	1.23	1.11	10.3

* the maximum number of values for change in the physical demands subscale was n = 14 due to a high number of respondents saying that the items in physical demands were not relevant for their job.
Diff: mathematical difference between standardized response means of the unchanged and change samples, Ratio: ratio of SRM changed and unchanged.
For comparison, the responsiveness of the QuickDASH and pain self-efficacy scales are also shown. SRM would expected to be larger in magnitude in those with improved problem (improved = yes), or improved pain intensity in comparison to those deemed not improved.

were expected. There were, however, some limitations. The "not relevant" option proved problematic in the physical demands domain of this measure where many people endorsed it for lifting/carrying which resulted in a high number of missing values.

In three of the four subscales more than 15% of our participants scored at the ceiling (100%, no limitations) exceeding current guidelines [26]. Ceiling effects are not wrong if the people in the study truly are healthy, however most of our sample [26, 42] met the National Institute of Occupational Safety and Health (NIOSH) surveillance case definition. The criteria for this definition include: having pain for greater than one week or more than once per month over the past year, and that the pain has been of moderate or worse severity over the past week. NIOSH case definition usually suggests a fairly substantial burden [44] however our current findings might indicate that this burden is not tapped by the WLQ-16. Alternatively, it could be that the pain was not limiting their work. Only work limitations are captured in the WLQ-16. Similar ceiling effects were also described with the WRF scale [14].

The weakest of the domains in the WLQ-16 in terms of psychometric properties was output demands. There are two possible reasons: first, the measure itself is having problems, or second, the domain was not relevant in this group and therefore was not as discriminating or responsive as the others. The items in that scale relate to not making mistakes, pleasing the supervisor, finishing work, and feeling a sense of accomplishment. Over 50% of the people in our study were reporting no difficulty (ceiling) in each of the output demand items (see Table 2), confirming that the people in our sample did not encounter limitations in output demands or they did not attribute them to their musculoskeletal pain. This lack of variability in response would likely account for the weaker statistical performance, however this should be tested in other samples. Ceiling effects, though problematic, are not always invalid if they do reflect a true lack of difficulty. Future work will attend to this domain to see if it is truly not an issue for this type of sample, or if the scale is not performing well.

Responsiveness of the WLQ-16 subscales was shown against two external indicators of change – transitional indices of change in their overall problem, and change in pain. Given that this was a work setting and that the participants were presenting with musculoskeletal disorders, we felt these were valid anchors for change in work disability as well. We calculated responsiveness statistics for those participants who felt they had improved, and for those who were not improved. We then used the differences in SRM values and a ratio of the two values to demonstrate that the

WLQ-16 was responsive to changes in pain and in overall problem. The subscales were also able to differentiate between people who had improved or not improved.

We found that the QuickDASH and self-efficacy were more responsive when considering changes that workers perceived in their overall problem. Interestingly this advantage can be attributed to the not improved group where the scores were much more stable in the QuickDASH and self-efficacy scale than the WLQ-16 domains. The magnitude of the effect size for the ''Improved group'' is relatively the same for both the construct of pain or problem. The use of the difference in SRM and the ratio of SRM between improved and not improved subgroups in this analysis allowed us to consider the validity of the changed scores by giving more emphasis to those with a larger magnitude of difference between changed and unchanged workers.

We did not see the expected results in responsiveness using the correlational approach of Deyo (39) despite seeing what we expected using the statistical approaches (i.e., SRM). The lack of correlation between change scores and other indicators of change could be due to a lack of variability in rankings which would lead to less variance and lower correlations. Change scores for those at or near the ceiling would also be small, despite perhaps large improvements in pain or overall problem. This would alter a correlation more than the effect size statistic. Alternatively it could be that our transitional indices were too disorder-related, and not close enough to the concept of work limitations to allow for a more sensitive correlation across the breadth of the scales. We believe this is less likely the case because of the larger effect sizes found when using a statistical summary.

There were limitations to our work. Our population was focused on those in the computer environment, and on persons with musculoskeletal pain only. Although as technology continues to advance, this represents a growing segment of the workforce and the most relevant and costly syndrome in that workforce, it still means our results may not be generalizable to a more physically demanding work environment or a different disorder. We also had a small sample, though with good and complete data. We encountered difficulties reaching our busy workforce and the inability to make that first contact is where we lost many of our eligible participants. Our sample is adequate for each of the descriptive and correlational statistics we undertook. We were not dependent on inferential statistics; so concern over Type I or Type II errors and erroneous stochastic significance is minimal. We felt a study of this size offered insights into the performance of the questionnaires and did provide us with good insight into the issues of ceiling effects in the sample.

Standardized outcome measures that quantify the effect of a disorder on workers who are still at their job are necessary components of an outcome assessment. They fulfill three needs: for a comprehensive understanding of the impact of a disorder on work, for an indicator of indirect costs for economic analyses, and for outcomes in studies of prognosis or treatment effectiveness. This study examined the psychometric properties of one measure that quantifies the amount of time a worker is experiencing a limitation. The study has shown indications of the the reliability, validity and responsiveness of the WLQ-16. We have raised some concerns about the measure, in particular the output demands subscale, and have made suggestions for ongoing research in varying populations. The next logical step would be to address these concerns, but also to do concurrent comparisons of the properties of this instrument and some of the other instruments now available (4, 6, 9, 10, 14).

Acknowledgements

The authors would like to acknowledge the contribution and support of the Repetitive Strain Injury (RSI) Watch committee at the Toronto Star and the study participants. This study was funded by an operating grant from The National Institute for Occupational Safety and Health (NIOSH) (# 5 R01 0H003708-03), and from operating funds of the Institute for Work & Health, Toronto. Dr Beaton was supported by a Canadian Institute of Health Research (CIHR) PhD Fellowship during the data collection for this project, and by a CIHR New Investigator's award during analysis and writing.

References

1. Stewart WF, Ricci JA, Leotta C. Health-related lost productive time (LPT): Recall interval and bias in LPT estimates. J Occ Environ Med 2004; 46: S12–S22.
2. Gilworth G, Chamberlain AM, Harvey A, Woodhouse A, Smith J, Smyth GM, et al. Development of a work instability scale for rheumatoid arthritis. Arthritis Rheum 2003; 49: 349–354.
3. Boden LI, Biddle EA, Spieler EA. Social and economic impacts of workplace illness and injury: Current and future directions for research. Am J Ind Med 2001; 40: 398–402.
4. Loeppke R, Hymel PA, Lofland JH, Pizzi L, Konicki DL, Anstadt GW, et al. Health-related workplace productivity measurement: General and migraine-specific recommendations from the ACOEM expert panel. J Occ Environ Med 2003; 45: 349–359.
5. Verstappen SMM, Bijlsma JWJ, Verkleij H, Buskens E, Blaauw AAM, Ter Borg EJ, et al. Overview of work disability in rheumatoid arthritis patients as observed in cross-sectional and longitudinal surveys. Arthritis Rheum 2004; 51: 488–497.
6. Prasad M, Wahlqvist P, Shikiar R, Shih Y. A review of self-report instruments measuring health-related work productivity. A patient-reported outcomes perspective. Pharmacoeconomics 2004; 22: 225–244.
7. Lofland JH, Pizzi L, Frick KD. A review of health-related workplace productivity loss instruments. Pharmacoeconomics 2004; 22: 165–184.
8. Scottsdale, AZ. Measuring employee productivity: A guide to self-assessment tools. Institute for Health and Productivity Management; 2001.
9. 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 Occ Environ Med 2004; 46: S23–S37.
10. Mayne TJ, Howard K, Brandt-Rauf PW. Measuring and evaluating the effects of disease on workplace productivity. J Occ Environ Med 2004; 46: S1–S2.
11. Lerner D, Amick III BC, Rogers WH, Malspeis S, Bungay K, Cynn D. The work limitations questionnaire. Med Care 2001; 39: 72–85.
12. Lerner D, Adler DA, Chang H, Berndt ER, Irish JT, Lapitsky L, et al. The clinical and occupational correlates of work productivity loss among employed patients with depression. J Occ Environ Med 2004; 46: S46–S55.
13. Sullivan S. Making the business case for health and productivity management. J Occ Environ Med 2004; 46: S56–S61.
14. Amick III BC, Habeck RV, Ossmann J, Fossel AH, Keller R, Katz JN. Predictors of successful work role functioning after carpal tunnel release surgery. J Occup Environ Med 2004; 46: 490–499.
15. Lerner D, Benjamin CA, Lee JC, Rooney T, Rogers WH, Chang H, et al. Relationship of employee-reported work limitations to work productivity. Med Care 2003; 41: 649–659.
16. Amick BC, III, Lerner D, Rogers WH, Rooney T, Katz JN. A review of health-related work outcome measures and their uses, and recommended measures. Spine 2000; 25: 3152–3160.
17. Spitzer WO, LeBlanc FE, Dupuis M, Abenhaim L, Belanger AY, Bloch R, et al. Scientific approach to the assessment and management of activity-related spinal disorders: A monograph for clinicians. Report of the Quebec task force on spinal disorders. Spine 1987; 12(7S): s4–s55.
18. Spitzer, WO, Skovron, ML, Salmi, LR, Cassidy, JD, Duranceau, J., Suissa, S., and Zeiss, E. Scientific monograph of the Quebec Task Force on Whiplash-Associated Disorders: Redefining ''whiplash'' and its management. Spine 1995; 20(8S): 1S–73S.
19. Cote P, Cassidy JD, Carroll L, Frank JW, Bombardier CA systematic review of the prognosis of acute whiplash and a new conceptual framework to synthesize the literature. Spine 2001; 26: E445–E458.
20. Cote P, Hogg-Johnson S, Cassidy JD, Carroll L, Frank JW. The association between neck pain intensity, physical functioning, depressive symptomatology and time-to-claim-closure after whiplash. J Clin Epidemiol 2001; 54: 275–286.
21. Beaton DE. Examining the clinical course of work-related musculoskeletal disorders of the upper extremity using the Ontario Workers' Compensation Board administrative database. Toronto: University of Toronto, 1995.
22. Beaton DE, Wright JG, Katz JN, Amadio P, Bombardier C, Cole DC, et al. Development of the QuickDASH: comparison of three item-reduction approaches. J Bone Joint Surg – American 2005; 87: 1038–1046.
23. Roland M, Morris R. A study of the natural history of back pain – part I: Development of a reliable and sensitive measure of disability in low-back pain. Spine 1983; 8(2): 141–144.
24. Roland M, Fairbank J. The Roland–Morris Disability Questionnaire and the Oswestry Disability Questionnaire. Spine 2000; 25: 3115–3124.
25. Lorig K, Chastain RL, Ung E, Shoor SM, Holman HR. Development and evaluation of a scale to measure perceived self-efficacy in people with arthritis. Arthritis Rheum 1989; 32: 37–44.
26. McHorney CA, Tarlov AR. Individual patient monitoring in clinical practice: Are available health status surveys adequate? Qual Life Res 1995; 4: 293.
27. DeVellis RF. A consumer's guide to finding, evaluating, and reporting on measurement instruments. Arthritis Rheum 1996; 9: 239–245.
28. Nunnally JC, Bernstein IH. Validity. Psychometric Theory. 3rd edn. New York, McGraw-Hill, 1994, 83–113.
29. Bombardier C, Hayden J, Beaton DE. Minimal clinically important difference. Low back pain: Outcome measures. J Rheumatol. 2001; 28: 431–438.
30. Deyo RA, Battie M, Beurskens AJ, Bombardier C, Croft P, Koes BW, et al. Outcome measures for low back pain research: A proposal for standardized use. Spine 1999; 23: 2003–2013.
31. Stratford PW, Binkley J, Soloman P, Finch E, Gill C, Moreland J. Defining the minimum level of detectable

change for the Roland-Morris questionnaire. Phys Ther 1996; 76: 359–368.
32. Beaton DE, Katz JN, Fossel AH, Tarasuk V, Wright JG, Bombardier C. Measuring the whole or the parts. Validity, reliability and responsiveness of the DASH Outcome Measure in different regions of the upper extremity. J Hand Ther 2001; 14: 128–146.
33. Beaton DE, Davis AM, Hudak P, McConnell S. The DASH (Disabilities of the Arm, Shoulder and Hand) Outcome Measure: What Do We Know about It Now? B J Hand Ther 2001; 6: 109–118.
34. Solway S, Beaton DE, McConnell S, Bombardier C. The DASH Outcome Measure: User's Manual. 2nd edn. Toronto: Institute for Work & Health; 2002.
35. De Bruin AF, Diederiks JPM, De Witte LP, Stevens FCJ, Philipsen H. Assessing the responsiveness of a functional status measure: The Sickness Impact Profile versus the SIP68. J Clin Epidemiol 1997; 50: 529–540.
36. Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. J Clin Epidemiol 2001; 54: 1204–1217.
37. Deyo RA, Patrick DL. The significance of treatment effects: The clinical perspective. Med Care 1995; 33: AS286–AS291.
38. Lydick E, Epstein RS. Interpretation of quality of life changes. Qual Life Res 1993; 2: 221.
39. 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.
40. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 1st edn. New York: Academic Press; 1988.
41. Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1998; 50: 239–246.
42. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of short and longer health status instruments. Med Care 1992; 30(10): 917–925.
43. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd edn. Toronto: McGraw-Hill, Inc.; 1994.
44. Beaton DE, Cole DC, Manno M, Bombardier C, Hogg-Johnson S, Shannon HS. Describing the burden of upper-extremity musculoskeletal disorders in newspaper workers: What difference do case definitions make? J Occup Rehab 2000; 10: 39–53.

Address for correspondence: Dorcas E. Beaton, Mobility Program Clinical Research Unit, St Michael's Hospital, Toronto, Ontario, Canada
Phone: +1-(416)-864-6060 x 6701; Fax: +1-(416)-927-4167
E-mail: beatond@smh.toronto.on.ca