Work Performance Assessed by a Newly Developed Japanese Version of the Work Limitation Questionnaire in a General Japanese Adult Population

Misa TAKEGAMI^{1,2}, Shin YAMAZAKI³, Annabel GREENHILL⁴, Hong CHANG⁴ and Shunichi FUKUHARA^{2,3}

¹Department of Preventive Medicine and Epidemiologic Informatics, National Cerebral and Cardiovascular Center, Japan, ²Institute for Health Outcomes and Process Evaluation Research (iHope International), Japan, ³Department of Healthcare Epidemiology, Kyoto University School of Medicine and Public Health, Japan and ⁴Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, USA

Abstract: Work Performance Assessed by a Newly Developed Japanese Version of the Work Limitation Questionnaire in a General Japanese Adult Population: Misa TAKEGAMI, et al. Department of Preventive Medicine and Epidemiologic Informatics, National Cerebral and Cardiovascular Center—Background: The Work Limitations Questionnaire (WLQ) was recently developed to measure health-related decrements in ability to perform job roles among employed individuals. The purpose of this study was to develop and test a Japanese version of the WLQ. **Methods:** Developing the Japanese version of the WLQ involved translations, back-translations, and a pilot study. Using data obtained from a nationwide survey, 4,600 people aged ≥ 20 years were selected from the entire population of Japan by stratified random sampling. We ultimately used data from a total of 1,358 workers out of 2,266 subjects who filled out the self-administered questionnaire. We computed the proportion of missing data, measured internal consistency reliability, and tested for convergent and discriminant validity, concurrent validity, known-groups validity, and the factor structure of this instrument. **Results:** For the Japanese version of the WLQ, the percentages of missing values for each scale ranged from 3.6% to 7.8%. Internal consistency reliability was high, and Cronbach's α was ≥ 0.7 for all subscales. Subjects with headache and orthopedic pain had significantly higher WLQ subscale scores than subjects without. Higher WLQ subscale scores were associated with depressive symptoms as measured with the Hospital Anxiety and Depression Scale ($p < 0.001$). **Conclu-**

sions: The Japanese WLQ provides reliable and valid information on at-work disability for group-level comparisons and tracking therapeutic outcomes. (J Occup Health 2014; 56: 124–133)

Key words: Reliability, Validity, Work disability, Work Limitations Questionnaire, Work productivity

Work disability remains an important public health and social policy issue, as work productivity is an important aspect of the critical interface of the effect of health disorders on the economy^{1,2}. The total cost of work disability to business and industry is comprised of not only the direct medical costs but also the even greater health-related productivity costs, which typically amount to two to three times the medical costs³. Work consequences of health problems have been increasingly quantified in population-based and clinical studies^{4,5}. Indeed, rates of employee work absences are widely used to indicate and, specifically, to estimate productivity loss due to missed work time; however, the productivity loss should include not only absenteeism but also presenteeism, which is the health-related productivity loss while at paid work^{1,6}.

The Work Limitations Questionnaire (WLQ) addresses the impact of health problems while on the job, what is known as presenteeism. The WLQ is a self-administered questionnaire measuring the degree to which health problems affect job performance (on-the-job work disability) and the work productivity impact of these work limitations in the previous two weeks⁷. The WLQ was developed to fill a gap in measurement of the consequences of health problems on working adults⁸.

The WLQ consists of the following four subscales,

Received Apr 17, 2013; Accepted Dec 7, 2013

Published online in J-STAGE Feb 28, 2014

Correspondence to: M. Takegami, Department of Preventive Medicine and Epidemiologic Informatics, National Cerebral and Cardiovascular Center 5-7-1 Fujishiro-dai, Suita, Osaka 565-8565, Japan (e-mail: takegami-kyt@umin.ac.jp)

each addressing the impact of physical and emotional health problems on performance of a specific category of work tasks: physical demands (6 items), which covers ability to perform tasks involving bodily strength, movement, endurance, coordination and flexibility; time management (5 items), which addresses difficulty handling a job's time and scheduling demands; mental-interpersonal demands (9 items), which addresses cognitively-demanding tasks and on-the-job social interactions; and output demands (5 items), which concerns reduced work quantity, quality and timeliness. (see Appendix)

The WLQ inquires about the level of difficulty in performing specific job demands. The time, mental-interpersonal, and output scale items address the amount of time that physical or emotional health problems made the performance of specific demands difficult. The physical scale refers to the amount of time the employee was able to perform a demand without difficulty due to health problems. Scale response options are as follows: "all of the time (100%)"; "a great deal of the time"; "some of the time (approximately 50%)"; "a slight bit of the time"; "none of the time (0%)"; and "does not apply to my job". Each of the four scale scores are computed as the mean of the non-missing responses and converted to 0 (not limited) to 100 (limited all of the time). For scales containing missing and "does not apply" responses, the half-scale imputation rule is applied. The WLQ productivity loss score, which indicates the percentage of at-work productivity loss for a given group or individuals compared with a benchmark sample of healthy employees, can be estimated using the weighted sum of the scores from four WLQ scales⁹⁾. This score reflects the estimated difference in percentage of productivity between the measured group and the benchmark. Several previous studies have already cited evidence of the WLQ's reliability and validity^{7, 10)}. In an employee population, WLQ scale scores were able to predict objectively measured work productivity⁸⁾, and it has been shown to be sensitive to changes in health status over time¹¹⁾. The WLQ's potential utility in providing new insights into the outcomes of patients has been shown, and a Japanese version has been developed by Sompo Japan Insurance Inc.¹²⁾ The Sompo version has already been validated using an Internet-based survey; however, the subjects in their validation study were only IT company workers and nurses, all of which were less than 50 years old. Its usefulness for other kinds of workers and community residents using a self-administered questionnaire remains unknown. Here, we reported the development of a Japanese version of the WLQ and testing of its reliability and validity.

Methods

Translation of the WLQ and pilot test

The Japanese version of the WLQ was developed to conform with standard methods that have been adopted internationally¹³⁾, including forward translation, back-translation, and examination of the translation quality, with the content of the translated questionnaire reviewed by one of the original developers of the English version. After review, the translated WLQ questionnaire was tested in 10 Japanese workers to identify problems with its cross-cultural equivalence and practicality. Following subsequent appropriate revision aimed at adapting the questionnaire to the Japanese culture without changing the intent of the original developers, the Japanese version of the WLQ was used in the current validation study.

Subjects and data collection

In this cross-sectional study, a total of 4,600 people aged ≥ 20 years were selected from the entire population of Japan by stratified random sampling. Data were obtained from the stratified sampling, including the district of residence (Japan was divided into nine districts for this survey) and the population of the city of residence (five classifications according to population: metropolis; cities with a population of more than 150,000, between 50,000 and 150,000, or less than 50,000; and rural districts). Between November and December 2007, a visit was made to the subjects' homes to distribute and subsequently collect self-administered questionnaires. Informed consent was obtained, and the study was approved by the Institutional Review Board of Fukushima Medical University.

Measurements

The self-administered questionnaire included the Japanese version of the WLQ and questions on the following items: demographic items (date of birth, sex, smoking status, frequency of alcohol consumption, marital status, educational level, annual household income); work-related items (kind of occupation, work hours per week, absence due to health problems); self-reported comorbid conditions; various symptoms including headache, low back pain, knee pain, limb pain; general health status (as assessed using the Medical Outcome Study Short-Form 36-Item Health Survey [SF-36], a valid and reliable instrument for measuring health-related quality of life)^{14, 15)}; and daytime sleepiness (as assessed using the Epworth Sleepiness Scale [ESS], a valid and reliable self-administered questionnaire for measuring subjective daytime sleepiness)^{16–18)}. The ESS score indicates stronger daytime sleepiness, and scores below 10

are considered to indicate no problem. In addition, depression was assessed as well using the depression subscales of the Hospital Anxiety and Depression Scale (HADS), which consists of seven items, the individual scores of which are summed to give an overall score ranging from 0 (no depression) to 21 (severe depression)¹⁹⁾. Two different cutoff scores (≥ 8 and ≥ 11) were established to identify subjects with mild or severe depression, respectively²⁰⁾.

Data analysis

We analyzed the data of all subjects who were marked as “employed or self-employed” in the national survey. All statistical analyses were performed with SPSS 12.0J for Windows (SPSS, Inc., Chicago, IL, USA) and SAS version 9.1 for Windows (SAS Institute Inc., Cary, NC, USA).

Descriptive analysis and item analysis

After computing the percentage of missing values for each question, we noted all items for which 90% or more of the subjects gave the same response. If the item had more than 10% of values missing, characteristics were compared between responders and nonresponders (subjects who either did not respond or answered, “does not apply to my job”) using a logistic regression model. We also examined whether or not each scale score’s distribution of responses was strongly skewed (large ceiling or floor effect). Within a relatively healthy working population, skewing towards the floor is expected, because the WLQ measures work limitations (in contrast to higher levels of work performance). We calculated the mean and standard deviations as well as missing values for each subscales score.

Reliability

Cronbach’s alpha coefficient was used as the index of internal consistency reliability²¹⁾, with values of approximately 0.7 or higher generally accepted as evidence of good internal consistency reliability.

Validity

To assess factor validity, we used factor analysis with principal factor analysis and promax rotation. Factor analysis was conducted using the WLQ subscale scores to test the hypothesis of unidimensionality.

The use of multi-trait analysis to evaluate convergent and discriminant validity has been described previously in detail²²⁾. Briefly, each item is hypothesized to belong to only one multi-item subscale. For each item, correlations between the score on that item and the scores on all the subscales are computed. Then, for each item, if the correlation between the

score on that item and the score on the subscale to which that item belongs is 0.4 or higher, that item is said to have passed the test of convergent validity. In addition, if the correlation between the score for each item and the score on the scale to which that item belongs is greater than the correlation between the item’s score and the scores on all of the scales to which that item does not belong, then that item is said to have passed the test of discriminant validity²³⁾.

To assess concurrent validity, we computed correlations between scores on the WLQ subscales and those on the SF-36 subscales. We hypothesized that the WLQ “physical demands” subscale scores would be associated more strongly with the SF-36 subscale scores measuring the similar domains of “physical functioning” than “mental health”. In the same way, the “mental-interpersonal demands” scores of the WLQ would be associated more strongly with “mental health” than “physical function” of the SF-36.

To test for known-groups validity, we examined the association between WLQ scales scores and the depression subscale score of the HADS. In addition, we also assessed whether or not WLQ scales scores varied based on presence of chronic symptoms such as headache, low back pain, knee pain, limb pain, or daytime sleepiness. An association with the WLQ is reported for all of these indicators, and comparisons can be made with the results of the present study. We used analysis of covariance (ANCOVA) to estimate the adjusted mean differences between groups and estimated least-squares means of WLQ subscale scores with adjustments for sex, age, and the number of comorbid chronic diseases, hypothesizing that WLQ scores of subjects with chronic pain were lower than subjects without, and that headache influenced physical demand scores to a lesser extent than orthopedic pain.

Results

Translation of the WLQ and pilot test

Two items were changed to better suit a Japanese audience. The item “get going easily at the beginning of the workday” was changed to “get going energetically from commencement of work”, as one of the developers pointed out that this item should capture the notion of having difficulty feeling either or both physically or mentally energized to work. In addition, to the item “do your work without stopping to take breaks or rests”, we added “(except for prescribed breaks, lunch breaks, etc.)”, as some subjects misinterpreted the question in the pilot test. Further, “10 lbs” was rewritten as “5 kg”, as the metric system is far more familiar to Japanese subjects than the imperial system.

While the WLQ items generally ask respondents

to rate their level of difficulty in completing a task, respondents are instead asked to rate their level of ability with regard to “physical demands”. In the original WLQ, a different question format was used to help remind readers that the questions in the “physical demands” subscale differed from previous ones, with questions in the “physical demands” subscale supplied as separate questions, while those in other subscales were supplied in table format. However, after translation, it was determined that writing separate questions for each item in the “physical demands” scale would prove too lengthy and confusing; therefore, we kept the questions in this section in the table format and added in bold face “was able to” to each item in this section to avoid any mistakes by respondents. No subject of the pilot test confused their level of ability with their level of difficulty in “physical demands”. All of the above changes were discussed with, and approved by, one of the original WLQ developers.

Subjects for psychometric tests

The survey targeted 4,600 people, with 2,308 ultimately responding to the questionnaire (50.2% response rate). Data from 1,358 worker (58.8%) were analyzed. The subjects' characteristics are described in Table 1.

Item analysis

Percentages of missing values for each scale and proportions of responses at the floor (lowest possible score) and ceiling (highest possible score) are shown in Table 2. Percentages of missing responses ranged from 0.4 to 0.8%, with a fairly large percentage of subjects responding, “does not apply to my job” (3.6–24.8%). For 8 of 25 items (32%), we noted $\geq 10\%$ missing values, including responses of “does not apply to my job”. Greater numbers of missing values in these 8 items correlated with demographical and work-related factors (Table 3). Older age and fewer work hours were associated with missing values. In the “physical demands” and “mental-interpersonal demands” subscales, type of occupation tended to be associated with responses of “does not apply to my job”. For example, a significantly high number of responses of “does not apply to my job” to the question regarding the ability to “lift, carry, or move objects” were noted in nonmanual laborers. None of the questions had response rates of 90% or more in any response category.

Reliability

Cronbach's alpha, which is the index of internal consistency reliability, was ≥ 0.7 for all scales (Table 4). In the “time demands” scales, alpha coefficients were calculated for each 4-question scale made by eliminat-

Table 1. Subject characteristics (n=1,358)

Characteristics	
Male, %	57.7
Age, mean (SD)	46.5 (13.2)
Smoking, %	34.5
Drinking almost every day, %	27.1
Married, %	70.6
Highest - level of education completed, %	
Elementary or junior high school	8.1
High school	40.8
College	17.3
University or graduate school	21.8
Not ascertained or other	12.0
Annual income (million), %	
<3	31.5
3–5	25.6
5–7	12.8
≥ 7	10.5
Missing	6.5
Occupation, %	
Nonmanual	50.8
Service	26.6
Manual	22.6
Hours worked per week, %	
<20	23.4
20–30	8.0
30–40	22.2
40–50	24.6
≥ 50	18.6
Missing	2.2
Hours worked per week, mean (SD)	37.1 (19.6)
Missing work days due to health problems in past 1 month, %	28.6
Comorbid conditions, mean (SD)	1.3 (1.5)
Hypertension, %	15.9
Diabetes, %	5.4
Cerebrovascular disease, %	1.1
Cardiovascular disease, %	2.6
Symptom	
Chronic headache, %	45.6
Low back pain, %	37.5
Knee pain, %	16.5
Limb pain, %	11.9
Excessive daytime sleepiness, %	26.5
Depressive symptom, %	19.9

ing one of the 5 questions, with values ranging from 0.89 to 0.90. Alpha coefficients for other scales were calculated in the same manner and showed similar results, indicating that none of the questions had an

Table 2. Percentages of missing data and of responses at the floor and ceiling effects (n=1,358)

WLQ scales	Items, n	Mean [†]	SD	% missing	% Floor (no limitation)	% Ceiling (severe limitation)
Time demands	5	8.4	16.0	7.8	55.5	0.5
Physical demands	6	21.4	31.1	5.6	43.7	7.5
Mental-interpersonal demands	9	11.2	14.5	3.8	35.2	0.0
Output demands	5	11.5	16.1	3.6	44.2	0.2
WLQ % productivity loss score [‡]	25	3.4	3.4	12.7	0.0	21.0

WLQ: Work Limitations Questionnaire. [†] Minimum scale score (least limited)=0; maximum scale score (most limited)=100. [‡] WLQ % productivity loss score indicates the percentage of at-work productivity loss, and 0% represents no productivity loss.

Table 3. WLQ items with 10% or more of missing values and demographic factors

WLQ items with 10% or more of missing values	Missing		Odds ratios and their 95% CI for each missing value using logistic regression models					
	% missing	% not applicable	Female (vs. male)	Age, per 10 yr increase	Income, less than 3 millions (vs. more)	Work hours, less than 40 hours per week (vs. more)	Occupation, nonmanual (vs. manual)	Occupation, service (vs. manual)
Time demands								
Work required hours	0.4	9.9	0.9 (0.5–1.4)	1.2 (1.1–1.4) [†]	1.1 (0.7–1.7)	1.3 (0.9–2.0)	0.8 (0.4–1.3)	1.1 (0.7–1.9)
Start on work soon after arriving	0.8	9.9	0.9 (0.5–1.4)	1.3 (1.2–1.6) [‡]	1.1 (0.7–1.8)	1.3 (0.9–2.0)	1.1 (0.6–1.8)	1.3 (0.7–2.2)
Work without breaks or rests	0.3	13.4	1.1 (0.7–1.7)	1.3 (1.1–1.5) [‡]	1.4 (0.9–2.0)	1.4 (1.0–2.0)	1.0 (0.6–1.6)	1.6 (1.0–2.5)
Physical demands								
Work / move around work locations	0.6	17.8	1.8 (1.2–2.7) [†]	1.3 (1.2–1.5) [‡]	1.9 (1.3–2.8) [‡]	2.0 (1.4–2.8) [‡]	0.5 (0.4–0.8) [‡]	0.8 (0.5–1.3)
Lift, carry, move objects, ≥10 lb	0.3	24.5	2.2 (1.6–3.3) [‡]	1.2 (1.1–1.3) [‡]	0.9 (0.6–1.3)	1.6 (1.2–2.1) [†]	3.2 (2.0–4.8) [‡]	2.0 (1.3–3.3) [†]
Use handheld tool, equipment	0.6	11.6	1.7 (1.1–2.8) [†]	1.5 (1.3–1.8) [‡]	1.6 (1.0–2.5)	1.3 (0.9–2.0)	0.5 (0.3–0.7) [†]	0.7 (0.4–1.1)
Mental-interpersonal demands								
Speak in person / on phone	0.6	15.9	1.7 (1.1–2.6) [†]	1.4 (1.3–1.6) [‡]	2.2 (1.5–3.3) [‡]	1.9 (1.3–2.8) [‡]	0.2 (0.1–0.4) [‡]	0.5 (0.4–0.8) [‡]
Help others to work	0.7	16.1	1.0 (0.6–1.4)	1.7 (1.5–2.0) [‡]	2.0 (1.4–3.1) [†]	1.3 (0.9–1.9)	0.8 (0.5–1.2)	1.0 (0.6–1.5)

WLQ: Work Limitations Questionnaire. [†] $p < 0.05$; [‡] $p < 0.001$.

Table 4. WLQ scaling test results (n=1,358)

WLQ scales	Items, n	Cronbach's alpha	Range of Item-to-Total Correlations	Convergent validity	Discriminant validity	Factor scoring coefficients
Time demands	5	0.915	0.85–0.89	100	100	0.171
Physical demands	6	0.953	0.87–0.93	100	100	0.015
Mental-interpersonal demands	9	0.915	0.67–0.85	100	100	0.478
Output demands	5	0.841	0.75–0.84	100	100	0.397

WLQ: Work Limitations Questionnaire.

unusually strong influence on the internal consistency.

Validation testing

Item-level factor analysis provided support for our hypothesis that items correlated more highly with their assigned scales than with other scales. However, of note, items in the “mental-interpersonal demands” scales independently converged into “mental demands” (factor loading: 0.98 to 0.64) and “interpersonal

demands” (factor loading: 0.70 to 0.57) scales.

The success eigenvalues (scree test) from the scale-level factor analysis were 2.198, 0.988, 0.514 and 0.300, and the factor analysis yielded a one-factor structure for all 4 WLQ scales, which was used to obtain the scoring coefficients for the WLQ scales in Table 4. Item-total correlation coefficients are also shown in Table 4. All items passed the tests for convergent and discriminant validity.

With regard to concurrent validity, “physical demands” scores had greater correlation with SF-36 “physical functioning” scores than with “mental

Fig. 1. Associations between Work Limitations Questionnaire (WLQ) subscale scores and scores on the depression domain of the Hospital Anxiety and Depression Scale (HADS), after adjustment for sex, age and the number of chronic diseases. *Two different cutoff scores of the HADS depression subscale score (cutoff of ≥ 8 and ≥ 11) were used to identify the mild depression group and severe depression group. † p -value for trend <0.001 .

health” scores ($r=0.25$ and 0.12 , respectively); in contrast, “mental-interpersonal demands” scores had lesser correlation with the above SF-36 scores ($r=0.50$ and 0.26 , respectively). Further, all greater work limitations were associated with greater depression severity (as measured with the HADS depression scale, Fig. 1; ANCOVA p -value for trend <0.001); for example, the “time demands” score was 6.2 for the group with no depressive symptoms and 22.7 for the group with most severe depressive symptoms ($p<0.001$).

The mean scores and 95% confidential intervals after adjustment for sex, age and number of comorbid conditions are shown in Table 5. All scales scores were significantly higher for those with knee pain than for those without, with similar trends noted for limb pain. As expected, all scale scores were significantly higher for subjects with chronic headache than for those without, with the exception of the “physical demands” scale. Similarly, reports of excessive daytime sleepiness had a significant negative impact on scores for “time demands” ($p<0.001$), “mental-interpersonal demands” ($p<0.001$) and “output demands” ($p<0.001$), but not those for “physical demands” ($p=0.643$).

Discussion

We developed a Japanese version of the WLQ

Table 5. Adjusted mean scores and 95% confidence intervals of the WLQ^{†‡}

		Time demands	Physical demands	Mental-interpersonal demands	Output demands	WLQ % productivity loss score [§]
Headache						
	Presence	10.8 (9.5–12.1)	21.9 (19.3–24.4)	13.2 (12.0–14.4)	13.6 (12.3–14.8)	3.9 (3.6–4.2)
	Absence	5.8 (4.5–7.0)	20.8 (18.4–23.3)	9.0 (7.9–10.1)	9.0 (7.8–10.2)	2.8 (2.6–3.1)
	p value	<0.001	0.566	<0.001	<0.001	<0.001
Low back pain						
	Presence	10.1 (8.6–11.5)	22.4 (19.5–25.2)	12.9 (11.7–14.2)	13.0 (11.6–14.5)	3.8 (3.5–4.2)
	Absence	6.9 (5.8–8.0)	20.6 (18.4–22.8)	9.8 (8.8–10.8)	10.1 (9.0–11.3)	3.0 (2.8–3.3)
	p value	<0.001	0.326	<0.001	0.002	<0.001
Knee pain						
	Presence	11.6 (9.4–13.9)	27.5 (23.2–31.8)	13.6 (11.7–15.6)	14.3 (12.1–16.5)	4.3 (3.8–4.8)
	Absence	7.4 (6.5–8.4)	19.9 (18.0–21.8)	10.5 (9.7–11.4)	10.7 (9.7–11.6)	3.1 (2.9–3.4)
	p value	<0.001	0.002	0.005	0.003	<0.001
Limb pain						
	Presence	14.3 (11.7–17.0)	26.5 (21.5–31.4)	15.4 (13.1–17.6)	15.8 (13.2–18.3)	4.7 (4.2–5.3)
	Absence	7.4 (6.4–8.4)	20.4 (18.5–22.2)	10.5 (9.6–11.3)	10.7 (9.7–11.6)	3.2 (2.9–3.4)
	p value	<0.001	0.025	<0.001	<0.001	<0.001

WLQ: Work Limitations Questionnaire. † Adjusted for sex, age and the number of chronic diseases. ‡ Minimum scale score (least limited)=0; maximum scale score (most limited)=100. § WLQ % productivity loss score indicates the percentage of at-work productivity loss, and 0% represents no productivity loss.

and documented its psychometric characteristics in a Japanese population. Overall, our findings demonstrated that the Japanese version has good validity and reliability. This validation study of the Japanese version of the WLQ targeting Japanese employees highlights several issues that were not found in the validation study of the original version.

Missing values due to nonresponders were extremely low across all domains (range: 0.4% to 0.8%); however, a relatively high rate of missing ($\geq 10\%$) due to selecting the response choice of “does not apply to my job” in the 8 items was associated with older age, lower income, fewer work hours, and groups of occupation classification (Table 3). Care should be taken when interpreting WLQ data obtained from those characteristics subjects with those characteristics.

The high proportion of responses of “does not apply to my job” in these items may be attributed that responders tending to select “does not apply to my job” in response to questions about activities with which they had little or no experience at work. For example, nonmanual labor and service workers tended to select “does not apply to my job” for questions inquiring about ability to lift, carry, or move objects weighing more than 10 lbs. In contrast, many manual workers responded “does not apply to my job” to questions inquiring about difficulty speaking with people in person, in meetings, or on the phone. As such, the half-scale rule is particularly useful when administering this questionnaire for Japanese working conditions.

A relatively large percentage of subjects were scoring at the WLQ scale floor (no limitation). Our study design—a population-based study—may have contributed to this. In fact, 35.4% subjects had no chronic disorders, and only 5.1% subjects had been hospitalized in the past year. In addition, the standard deviation and ceiling effect of the physical scale score are larger than those of the other three scales scores. In the WLQ questionnaire, the response items for the time, mental-interpersonal and output scales ranged from “limited all of the time” to “not limited at all”, while the response items for the physical scale ranged from “able to do all of the time” to “able to do none of the time”. The respondents might have the tendency to continue answering the way they did for the other scales and mistakenly answered the items in the physical scale. We might need to change the design to minimize the response error due to the direction of response choices in the physical scale.

The results of psychometric tests indicated that the Japanese version of the WLQ has good internal consistency reliability, factor validity, and known-groups validity. Further, given that the time demands, physical demands and mental-interpersonal demands

subscales had Cronbach’s alpha values of more than 0.9, the number of items in these subscales could be decreased without compromising reliability. Indeed, Beaton *et al.* tested a 16-item WLQ and demonstrated its reliability and validity within populations with musculoskeletal problems²⁴. Lerner *et al.* have developed an eight-item short-form WLQ, which is used extensively in the United States.

WLQ subscale scores varied based on the presence of various symptoms, including headache, low back pain, knee pain, limb pain, and daytime sleepiness. We also noted a dose-response relationship between the severity of depression and all WLQ subscale scores. In addition, Lerner *et al.* showed that the pattern of work limitation was logically consistent with the characteristics of chronic headache and rheumatoid arthritis; headaches limited work performance more than rheumatoid arthritis did in the “time demands” and “mental-interpersonal demands” subscales, as headaches involve sleep disturbance, fatigue, and pain, all of which disrupt activities, as well as symptoms of depression and irritability²⁵. Our present findings in subjects with headaches versus those with orthopedic pain were similar to those of the original study¹⁰. Mulgrew, *et al.* reported that strong associations were evident between subjective daytime sleepiness and three of four scales of the WLQ among patients with obstructive sleep apnea (time, mental-interpersonal and output demands), but not physical demands⁵. Our study supported these results.

The work productivity loss score estimated the percent difference in output between employees who have health-related work limitations and employees who do not. However, the economic impact of work disability should be evaluated based on a combination of both absenteeism as well as presenteeism. Future studies will be needed to develop a method of assessing total economic impact, including both absenteeism and presenteeism.

Another Japanese version of the WLQ has been developed in Japan by Sompo Japan Insurance Inc.¹²; this version was also developed in coordination with the developer of the original WLQ. Thus, it is similar to our version, but uses more formal language. There are two important differences between the Sompo version and our version. First, the validation study for the Sompo version was an Internet-based survey, whereas we conducted a self-administrated questionnaire survey. Second, the setting of the validation study for the Sompo version was one company and one hospital, whereas our validation study was a population-based survey. Our version might be more suitable for manual labors and older workers than the Sompo version. Our version might also be more

useful in a population-based study than the Sompō version. The subjects in the validation study for the Sompō version were much more limited in terms of demographics than the subjects in our study, which included 373 IT company workers and 337 nurses. The proportions of female employees were 29.0 and 95.5%, respectively, and 92.6% of subjects were less than 50 years old. Moreover, we used many hypotheses to test the known-groups validity using various common symptoms. On the other hand, the known-groups validity of the Sompō version was tested only using job stress. There were also a few differences in the instructions and responses options. In the Sompō version, “your physical health or emotional problems” in the instructions was literally translated as “*kenkou-jotai ya kanjou teki na monda*”. However, we instead used “*sintaiteki aruiha sinriteki na riyuu*”, which means “for physical and mental problems” in line with the SF-36^{14, 15}. In addition, the response options use the expression “*sisho ga atta*” in the Sompō version, while we used “*muzukasikatta*” instead, both of which mean “made it difficult for you to do”, but it seems that ours may be easier to understand, making the document more accessible to a wider general readership.

The strength of the present study is in the fact that the data were obtained from a representative sample of the Japanese general population. Therefore, the average WLQ subscale score and productivity loss score can easily be regarded as norm scores for Japanese workers. However, several limitations to our study warrant mention. First, the study design was cross-sectional, and we did not assess test-retest reliability or responsiveness to detect changes over time within groups with some condition. Secondly, this study did not look at the relationship with other scales that measure work performance. This is because, to our knowledge, no scales that measure presenteeism, including the Health and Work Performance Questionnaire (HPQ), had been developed in Japan at the time this study was started. Third, in this study, the percentage of subjects whose weekly work time exceeded 40 hours was 43.2%. This may have been because overall, the percentages of respondents by sex and age differed from the population distribution of Japan. Among the male subjects in this study, those in their 20 s and 30 s accounted for 10.5 and 17.4%, respectively. In contrast, the percentages of men in their 20 s and 30 s were 15.9 and 19.8%, respectively, in the 2008 national population census. Conversely, the proportion of men in their 40 s–70 s was higher in this study than in the national census. Similarly, the percentage of women over 40 years old was higher in this study than in all of Japan. Although it does not affect the psychometric properties of the scale,

work time is reported to affect work performance²⁶, and when interpreting the mean values of the WLQ in this study, caution is needed with respect to their use as standard values for evaluation of work performance, especially in regular full-time work such as in companies. Finally, we used the work productivity loss score with the original scoring algorithm, and the weights were obtained from analysis of the relationship between WLQ scale scores and actual employee productivity in American workers⁹). Although it was found that productivity scores were significantly different between known groups in this study, future studies will be necessary both to compare the WLQ with other presenteeism measures and to test the estimation of WLQ productivity loss scores in Japanese employees.

Conclusions

Psychometric testing indicates that data obtained with the Japanese version of the WLQ are sufficiently reliable and valid in a population-based study. Although several issues arose during validation that we should consider when using the WLQ, we believe that the WLQ represents an important tool for assessing the social and economic impact of chronic health programs.

Acknowledgments: We are grateful to Dr. Debra Lerner, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, for her valuable suggestions. This study was supported by grants from the Institute for Health Outcomes and Process Evaluation Research (iHope International).

References

- 1) Loeppke R, Hymel PA, Lofland JH, et al. Health-related workplace productivity measurement: general and migraine-specific recommendations from the ACOEM Expert Panel. *J Occup Environ Med* 2003; 45: 349–59.
- 2) 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.
- 3) Brady W, Bass J, Moser R, Jr., Anstadt GW, Loeppke RR, Leopold R. Defining total corporate health and safety costs--significance and impact. Review and recommendations. *J Occup Environ Med* 1997; 39: 224–31.
- 4) Wang PS, Simon GE, Avorn J, et al. Telephone screening, outreach, and care management for depressed workers and impact on clinical and work productivity outcomes: a randomized controlled trial. *Jama* 2007; 298: 1401–11.
- 5) Mulgrew AT, Ryan CF, Fleetham JA, et al. The impact of obstructive sleep apnea and daytime sleepiness on work limitation. *Sleep Med* 2007; 9:

- 42–53.
- 6) Lofland JH, Pizzi L, Frick KD. A review of health-related workplace productivity loss instruments. *Pharmacoeconomics* 2004; 22: 165–84.
- 7) Lerner D, Amick BC, 3rd, Rogers WH, Malspeis S, Bungay K, Cynn D. The Work Limitations Questionnaire. *Med Care* 2001; 39: 72–85.
- 8) Lerner D, Amick BC, 3rd, Lee JC, et al. Relationship of employee-reported work limitations to work productivity. *Med Care* 2003; 41: 649–59.
- 9) Lerner D, Chang H, Rogers WH, Benson C, Schein J, Allaire S. A method for imputing the impact of health problems on at-work performance and productivity from available health data. *J Occup Environ Med* 2009; 51: 515–24.
- 10) Lerner D, Reed JI, Massarotti E, Wester LM, Burke TA. The Work Limitations Questionnaire's validity and reliability among patients with osteoarthritis. *J Clin Epidemiol* 2002; 55: 197–208.
- 11) Amick BC, 3rd, 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: 3152–60.
- 12) Ida H, Nakagawa K, Miura M, Ishikawa K, Yakura N. Development of the Work Limitations Questionnaire Japanese Version (WLQ-J): fundamental examination of the reliability and validity of the WLQ-J. *Sangyo Eiseigaku Zasshi* 2012; 54: 101–7.
- 13) Acquadro C, Jambon B, Ellis D, Marquis P. Language and Translation Issues. In: Spilker B, editor. *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1996. p. 575–85.
- 14) Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992; 30: 473–83.
- 15) Fukuhara S, Bito S, Green J, Hsiao A, Kurokawa K. Translation, adaptation, and validation of the SF-36 Health Survey for use in Japan. *J Clin Epidemiol* 1998; 51: 1037–44.
- 16) Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep* 1991; 14: 540–5.
- 17) Johns MW. Reliability and factor analysis of the Epworth Sleepiness Scale. *Sleep* 1992; 15: 376–81.
- 18) Takegami M, Suzukamo Y, Wakita T, et al. Development of a Japanese version of the Epworth Sleepiness Scale (JESS) based on item response theory. *Sleep Med [Controlled Clinical Trial Research Support, Non-U.S. Gov't]*. 2009; 10: 556–65.
- 19) Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983; 67: 361–70.
- 20) Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale. An updated literature review. *J Psychosom Res* 2002; 52: 69–77.
- 21) Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951; 16: 297–334.
- 22) Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 1959; 56: 81–105.
- 23) Hays RD, Hayashi T. Beyond Internal Consistency Reliability: rationale and user's guide for Multitrait Analysis Program on the microcomputer. *Behavior Research Methods, Instruments and Computers* 1990; 22: 167–75.
- 24) Beaton DE, Kennedy CA. Beyond return to work: testing a measure of at-work disability in workers with musculoskeletal pain. *Qual Life Res* 2005; 14: 1869–79.
- 25) Adler DA, McLaughlin TJ, Rogers WH, Chang H, Lapitsky L, Lerner D. Job performance deficits due to depression. *Am J Psychiatry* 2006; 163: 1569–76.
- 26) Collins JJ, Baase CM, Sharda CE, et al. The assessment of chronic health conditions on work performance, absence, and total economic impact for employers. *J Occup Environ Med* 2005; 47: 547–57.

Appendix. The sample items from the Work Limitations Questionnaire (WLQ)

In the past 2 weeks, how much of the time did your physical health or emotional problems make it difficult for you to do the following?

	All of the Time (100%)	Most of the Time	Some of the Time (About 50%)	A Slight Bit of the Time	None of the Time (0%)	Does Not Apply to My Job
a. do your work without stopping to take breaks or rests	☐ ₁	☐ ₂	☐ ₃	☐ ₄	☐ ₅	☐ ₀
b. stick to a routine or schedule	☐ ₁	☐ ₂	☐ ₃	☐ ₄	☐ ₅	☐ ₀
c. keep your mind on your work	☐ ₁	☐ ₂	☐ ₃	☐ ₄	☐ ₅	☐ ₀
d. speak with people in person, in meetings or on the phone	☐ ₁	☐ ₂	☐ ₃	☐ ₄	☐ ₅	☐ ₀
e. handle the workload.	☐ ₁	☐ ₂	☐ ₃	☐ ₄	☐ ₅	☐ ₀

Note: Items a. and b. are from the Time Demands scale. Items c. and d. are from the Mental-Interpersonal Demands scale. Item e. is from the Output Demands scale.

In the past 2 weeks, how much of the time were you **ABLE TO DO** the following without difficulty caused by physical health or emotional problems?

(Mark one box on each line a. through f.)

	All of the Time (100%)	Most of the Time	Some of the Time (About 50%)	A Slight Bit of the Time	None of the Time (0%)	Does Not Apply to My Job
a. walk or move around different work locations (for example, go to meetings)	☐ ₁	☐ ₂	☐ ₃	☐ ₄	☐ ₅	☐ ₀
b. use hand-held tools or equipment (for example, a phone, pen, keyboard, computer mouse, drill, hairdryer, or sander)	☐ ₁	☐ ₂	☐ ₃	☐ ₄	☐ ₅	☐ ₀

Note: Items a. and b. are from the Physical Demands scale.