Industrial Health 2020, **58**, 375–387

Original Article

Construct validity and test-retest reliability of the World Mental Health Japan version of the World Health Organization Health and Work Performance Questionnaire Short Version: a preliminary study

Norito KAWAKAMI[1]*, Akiomi INOUE[2], Masao TSUCHIYA[1], Kazuhiro WATANABE[1], Kotaro IMAMURA[1], Mako IIDA[3] and Daisuke NISHI[1]

[1]Department of Mental Health, Graduate School of Medicine, The University of Tokyo, Japan
[2]Department of Public Health, Kitasato University School of Medicine, Japan
[3]Department of Psychiatric Nursing, Graduate School of Medicine, The University of Tokyo, Japan

Received June 3, 2019 and accepted March 9, 2020
Published online in J-STAGE March 14, 2020

Abstract: The aim of the study was to investigate test-retest reliability and construct validity of the World Mental Health Japan (WMHJ) version of World Health Organization Health and Performance Questionnaire (WHO-HPQ) short version according the COSMIN standard. We conducted two consecutive surveys of 102 full-time employees recruited through an Internet survey company in Japan, with a two-week interval in 2018. We calculated Pearson's correlation (r) of measures of the WHO-HPQ with other presenteeism scales (Stanford Presenteeism Scale, Work Functioning Impairment Scale, and perceived relative presenteeism), health and psychosocial job conditions. We tested the test-retest reliability (intraclass correlation, ICC) among those who reported no change of job performance during the follow-up. Among 92 (90%) respondents, the absolute presenteeism significantly correlated with WFun and perceived relative presenteeism (r=−0.341 and −0.343, respectively, *p*=0.001) and psychological distress (r=−0.247, *p*=0.018). The absolute/relative absenteeism did not significantly correlate with the other covariates. The test-retest reliability over a two-week period was high for the WHO-HPQ absolute presenteeism (ICC, 0.73), while those for absolute/relative absenteeism measures were moderate. The study found an adequate level of test-retest reliability, but limited support for the construct validity of the absolute presenteeism measure of the WMHJ version of the WHO-HPQ. Further research is needed to investigate the construct validity of the WHO-HPQ measures in a larger sample.

Key words: Productivity, Absenteeism, Presenteeism, Test-retest reliability, Consensus-based standards for the selection of Health Measurement Instruments (COSMIN)

Introduction

Measuring work productivity loss has become increasingly important in research on mental health. It has been shown that a societal cost of mental disorders is large[1)]. Work productivity loss has been to shown to be one of the

*To whom correspondence should be addressed.
E-mail: nkawakami@m.u-tokyo.ac.jp

largest components of the societal cost[2, 3]. Employers are also concerned with the cost effectiveness or cost benefit of a workplace intervention[4]. Recent studies of workplace interventions tend to focus on the impact of intervention programs on work performance[5].

Work productivity loss consists of two components: absenteeism and presenteeism. Absenteeism is the number (or the proportion) of lost workdays per a certain period; presenteeism is the reduction of on-the-job performance[6]. A number of instruments have been developed to measure absenteeism and presenteeism[7]. Some are developed to measure absenteeism and presenteeism of workers with specific health conditions, such as depression[8] and musculoskeletal disorders[9]. Most instruments have been shown to be reliable and valid, while a further effort is needed to comprehensibly assess the psychometric properties of these instruments[7, 9]. In Japan, absenteeism and presenteeism of workers due to health problems is a concern both for occupational health and productivity loss[10]. Well-established instruments of absenteeism and presenteeism, such as the Stanford Presenteeism Scale (SPS)[11] and Work Limitation Questionnaire (WLQ)[12] have been translated and tested for the reliability and validity in Japan[13–15]. Some instruments, such as Work Functioning Impairment Scale (WFun), were developed even originally in Japan[14] and extensively used in other countries[16].

Although many measures have been developed to assess both absenteeism and presenteeism among employees, most of them focus solely on presenteeism[11, 14]. The World Health Organization Health and Work Performance Questionnaire (WHO-HPQ) is the only instrument that measures both absenteeism and presenteeism[6]. The original English version of WHO-HPQ has been validated against administrative records of a company and performance ratings by supervisors among employees with different occupations[6, 17]. It was also validated among patients with arthritis[18, 19].

The original English version of WHO-HPQ has been translated into several languages (Portuguese for use in Brazil; Spanish; French; Hebrew; and Japanese) (https://www.hcp.med.harvard.edu/hpq/info.php). However, not many studies have been conducted to validate the translated versions. The Persian version was tested for the validity of absenteeism measures with administrative records of a company[20]. The Japanese version was found to correlate with and predict sick leave due to mental disorders[21, 22]. However, no translated version has been fully tested for its psychometric properties, such as test-retest reliability and construct validity including associations with other absenteeism/presenteeism scales[23].

The World Mental Health Japan (WMHJ) Survey conducted in 2002–2006 included some items from the short version of WHO-HPQ[24], and this was used to estimate work productivity loss due to mental disorders[25]. This WMHJ version of WHO-HPQ was translated earlier using a slightly different wording from the recent Japanese translation by Suzuki *et al*[21, 22]. The WMHJ version has not yet been tested for reliability or validity.

The study aim was to preliminarily evaluate the reliability and validity of the WHO-HPQ based on its version used in the WMHJ by investigating its test-retest reliability, construct validity, and responsiveness in a sample of employees, according to the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN)[23]. We conducted two consecutive surveys of a small sample of full-time workers in Japan with a two-week interval to see if the WHO-HPQ absenteeism and presenteeism measures showed acceptable levels of test-retest reliability. We also tested if the measures correlated with eight selected covariates that are thought to be related to absenteeism and presenteeism, including other measures of presenteeism, in order to know the construct validity (hypothesis testing)[23].

For testing the construct validity, we first calculated correlations between the WHO-HPQ measures and other absolute presenteeism scales, i.e., SPS[11], WFun[13], and perceived relative presenteeism to know if the WHO-HPQ absolute presenteeism measure correlate with these three scales. However, while the WHO-HPQ absolute presenteeism measure captures presenteeism from any reason[6], these other presenteeism scales specifically asked presenteeism from health-related problems[11, 13]. Thus, the associations is expected to be only moderate. The WHO-HPQ measures of relative presenteeism would correlate with perceived relative presenteeism more than with SPS or WFun that measure absolute presenteeism. The WHO-HPQ absolute and relative absenteeism measures would not correlate with these other presenteeism scales, because a sick worker with no absenteeism may have greater presenteeism. Second, we tested the correlations between the WHO-HPQ measures and five possible predictors (two health conditions and three psychosocial job conditions. We assumed that WHO-HPQ absenteeism measures would correlate positively with psychological distress and depression/anxiety disorder[10] and negatively with job control, and supervisor and coworker support[26]; Similarly, the WHO-HPQ presenteeism measures, for which the greater score implies better work performance[6], would

correlate negatively with psychological distress and depression/anxiety disorder, and positively with job control, and supervisor and coworker support[27].

Subjects and Methods

Sample

A sample (N=102) of employees was drawn from a large pool (n>100,000) of registered members of a large Internet survey company in Japan. The inclusion criteria were: currently being employed full-time by a company or organization; and being aged between 20 and 60 yr old. They were asked to respond to two anonymous Internet surveys within a two-week interval (T1 and T2), because a 2–4 wk period was the most recommended time interval for test-retest reliability[28]. In previous studies, the interval used for the test-retest reliability for measures of absenteeism and presenteeism varied, e.g., one day[29], 1–2 wk[19, 30, 31], and one month or longer[32]. A short interval, such as one day, may overestimate the test-retest reliability because respondents remember their initial responses[28]. While a longer interval may underestimate the test-retest reliability due to a change in a target condition, a previous study reported that the test-retest reliability of a health-related quality of life scale was similar for two different intervals of two days and two weeks[33]. In addition, even for a longer interval, limiting respondents to those who reported no change during the follow-up would help correctly estimating the test-retest reliability[23]. Thus we decided to use a two-week interval for estimate the test-retest reliability among respondents who reported no change in their work performance.

The sample size was planned so that a moderate correlation (Pearson's r=0.3 between a scale score and other variables)[34] could be statistically significant ($p<0.05$, two tailed) with the power of 0.80 and 80% of valid responses.

The study aim and procedure were fully informed to participants and consent was obtained. The study plan was reviewed and approved by the Research Ethics Committee of the Graduate School of Medicine/Faculty of Medicine, The University of Tokyo (No. 2953-(4)).

Measures

The short-version of WHO-HPQ

In addition to items that were already translated into Japanese in the WMHJ Survey in 2008, the authors (N.K., A.I. and M.T.) translated some other items on the short version of WHO-HPQ. The authors again reviewed all the items and made modifications based on discussion and feedback from three employees selected at companies for which some of the authors (N.K. and D.N.) worked as occupational physicians. The final version was back-translated to English by a commercial translator and reviewed by Professor Kessler, the researcher who developed the WHO-HPQ.

Briefly, absolute absenteeism is defined as total hours lost from work in a certain time frame; and relative absenteeism is hours lost from work relative to the total work hours. Absolute presenteeism is defined as work performance (i.e, the quality of work) rated by a respondent on a 0–10 scale; relative presenteeism is self-rated performance relative to work performance done by most coworkers that are also rated by the same respondent. According to the scoring manual, the following measures were calculated. See the items and calculation formula in the Appendices 1 and 2, respectively):

1) Absenteeism
a) Using four-week estimates
Absolute absenteeism in the past four weeks
Relative absenteeism in the past four weeks
b) Using seven-day estimates
Absolute absenteeism in the past seven days
Relative absenteeism in the past seven days
2) Presenteeism (work performance)
Absolute presenteeism
Relative presenteeism (ratio)
Relative presenteeism (subtraction)

The survey at T1 used the WMHJ version of the WHO-HPQ. For testing the construct validity, we used measures at T1 in principle. After a preliminary analysis of all the collected data from the T1 survey revealed that respondents seemed to rate similarly on B9 (work performance of most people working on a similar job) and B11 (their own work performance), since the correlation between these two questions was strong (r=0.653). Thus, for the T2 survey, we added one sentence to B9 to clarify for respondents that the question asked about their co-workers' job performance, not their own. The modified version was used in the second survey at T2. We used the relative presenteeism measures at T2 for testing the construct validity. This modification also made it impossible to calculate the relative presenteeism measures comparative for the two surveys. For this reason, we could not calculate the test-retest reliability for the relative presenteeism measures.

Other presenteeism scales

The SPS[11] and WFun[13] were measured in the survey at T1. The SPS was translated into Japanese and already vali-

dated[35]. The SPS score ranged from 10 to 50, with higher scores being indicative of greater presenteeism. The score was calculated only when a respondent endorsed any of 13 chronic conditions[35]. The WFun is a seven item self-rated scale to measure work performance on the job at present, developed and validated in Japan[13, 14, 36]. The total score of WFun ranged from 4 to 28, with higher scores being indicative of greater presenteeism. In the survey at T2, one question, A13 from the Clinical Trials Baseline Version of the WHO-HPQ, was added to ask respondent' perceived relative presenteeism using a seven-point response option, with a greater score being indicative of poorer relative presenteeism (see Appendix 1). To ascertain the responsiveness, in the survey at T2, a single-item question asked if a respondent had better or poor work performance compared to two weeks ago, with a seven-point response option (e.g., 1=much better, 4=no change, and 7=much worse).

Other covariates

For mental health conditions, K6 was measured to assess psychological distress[37, 38]. In addition, the SPS listed 13 chronic conditions[11]. One item from the list was used to determine if a respondent had depression/anxiety disorder. Selected psychosocial job conditions, i.e., job control, and supervisor and coworker support, were measured using the Brief Job Stress Questionnaire, which has been well-validated in Japan[36]. Information of sex, age, occupation, and educational attainment was also collected in the survey at T1.

Statistical analysis

The authors made corrections to some apparently careless input values. For instance, if a respondent was not a manager, his/her expected work hours (B2) were set as 40 h per week, which is the legal requirement for regular work hours in Japan. For respondents who reported 0 h on B6, the response was replaced with an estimation based on B5a to B5e. Minimum, maximum, and average scores of WHO-HPQ measures were calculated.

For testing the construct validity, the COSMIN taxonomy integrates convergent, discriminant and known groups validity into one single concept, i.e., the "hypothesis testing"[23]. Also in the COSMIN taxonomy, the criterion validity is only used when it is compared with a "gold" standard, such as objectively measured absenteeism and presenteeism. In the present study, we investigated only the hypothesis testing for the validity (see the list of hypotheses in the Appendix 3). Pearson's correlation coefficients for the measures of the WHO-HPQ (absolute and relative absenteeism measures, both for 4 wk and 7 d, at T1, absolute presenteeism at T1, and relative presenteeism measures, both ratio and subtract, at T2) were calculated with the three other presenteeism scales and the five covariates to examine the construct validity (hypothesis testing).

Test-retest reliability was measured by the intraclass correlation coefficients (ICCs) of the measures from the WHO-HPQ at T1 and T2 in a one-way random model, only for those who reported no changes in their work performance between T1 and T2 (4=no change on the 7-point single question on the self-reported changes of work performance), following the definition of test-retest reliability of COSMIN[23]. The responsiveness was tested by the Pearson's correlation coefficients between the T1–T2 changes of the WHO-HPQ measures and the self-reported changes in work performance assessed at T2. Because the Internet survey required the participants to answer all items, there were no missing values on any variables or items. The IBM SPSS Software (ver. 22) was used for the analyses. The statistical significance of the correlations was assessed with two-side test with an alpha level of 0.05.

Results

Participants

All participants at T1 participated in the survey at T2. The following respondents were excluded from the analysis: those who reported 97 h employed per week (n=3); who had a large discrepancy between the reported value on B6 and one estimated from B5a-B5e (more than 2SDs, i.e., 152) (n=7). The final sample for the analysis included 92 respondents (Table 1). Half were women, with an average age of 43.1 yr old. Most of them were white-collar workers such as clerks, professionals and technicians, and managers. They worked about eight hours longer in the past week than the labor standard work hours (i.e., 40 h per week) on average. Almost half were university graduates. About 60% were currently married and had a child. Less than half (45.7%) had any of 12 chronic medical conditions. Back/neck disorders were most frequent (16.3%), followed by depression/anxiety disorder (10.9%).

Construct validity

Average values of both absolute and relative absenteeism measures were negative, indicating that respondents worked longer than they were expected on average (Table 2). The scores of all absolute and relative absenteeism measures showed a unimodal, right-skewed distribution. No significant correlation was observed between any

Table 1. Demographic, occupational and health-related characteristics of the respondents (n=92)

	n	%	Mean	SD
Sex (women)	43	46.7		
Age (yr)			43.1	11.2
20–34	24	26.1		
35–49	38	41.3		
50–60	30	32.6		
Occupation				
Managers	18	19.6		
Professionals/technicians	19	20.7		
Clerks	33	35.9		
Service workers	9	8.7		
Production/machine operators	14	15.2		
Work hours in the past week			47.9	12.0
Education (university or higher)	48	52.2		
Marital status (married)	57	62.0		
Having a child (any)	54	58.7		
Chronic conditions				
Allergy	7	7.6		
Stomach/bowels	6	6.5		
Asthma	2	2.2		
Back/neck disorders	15	16.3		
Heart or circulatory	2	2.2		
Depression/anxiety disorder	10	10.9		
Diabetes	3	3.3		
Arthritis/joint pain	8	8.7		
Migraine/chronic headache	9	9.8		
Hearing impairment	4	3.3		
Vision impairment	3	3.3		
Skin diseases	7	7.6		
Others	3	3.3		
Any of the above	42	45.7		
WFun presenteeism score T1 (7–35)			15.1	6.6
SPS presenteeism score T1 (1–50)			32.9	6.5
Perceived relative presenteeism T2 (1–7)			3.0	1.4
Psychological distress (K6) T1 (0–24)			12.1	5.5
Job control score T1 (3–12)			7.9	2.0
Supervisor support score T1 (3–12)			6.7	2.1
Coworker support score T1 (3–12)			6.8	2.2

SD: standard deviation.

of the absenteeism measures and any of the other presenteeism scales. For the hypothesis of the correlations with health conditions and psychosocial job conditions, only supervisor support significantly and negatively correlated with four-week absolute absenteeism (p=0.044).

The scores of the absolute and relative presenteeism measures showed a unimodal, left-skewed distribution. Average scores of absolute presenteeism measures at T1 were about 60. Relative presenteeism (subtraction) at T2 was small but positive, indicating that participants rated their work performance slightly better than others on average. For the hypotheses of correlations with the other presenteeism scales, the absolute presenteeism measure significantly and negatively correlated with WFun and perceived relative performance (p=0.001); it also marginally significantly and negatively with SPS (p=0.050). For the hypothesis of a negative correlation with health conditions and psychosocial job conditions, absolute presenteeism significantly and negatively correlated only with K6 (p=0.018), and marginally significantly and positively with job control (p=0.055).

For the hypothesis of a correlation with perceived relative presenteeism, perceived relative presenteeism significantly and negatively correlated with WHO-HPQ relative presenteeism measures (both ratio and subtract) at T2 (p<0.001). For the correlations with health conditions and psychosocial job conditions, none of these variables significantly correlated with relative presenteeism measures. Sex, age or education did not significantly correlate with any WHO-HPQ measures (p>0.05, data available upon request).

Test-retest reliability and sensitivity to change

A total of 64 (63%) participants reported no changes in work performance during the past two weeks. These participants had significantly lower prevalence of chronic conditions than participants who reported the changes (n=17) (39% and 61%, respectively). Otherwise, no statistically significant difference between these two groups. The ICC calculated for the participants reported no changes was high enough for absolute presenteeism (Table 3). ICCs were moderate for four-wk absolute and relative absenteeism; ICCs were slightly greater for the seven-day absenteeism measures. The change in work performance in two weeks significantly and positively correlated with the change in absolute absenteeism in the total sample (n=92, r=0.252, p=0.015), but not with the change in other measures (r=0.085 − 0.174, p>0.05).

Discussion

The aim of the study was to preliminarily investigate test-retest reliability and construct validity of the WMHJ version of WHO-HPQ in Japan. The absolute and relative absenteeism measures and the absolute presenteeism measure of the WMHJ version of the WHO-HPQ were stable over a two-week period (test-retest reliability). Among eight correlations hypothesized, the absolute presenteeism

Table 2. Pearson's correlation coefficients of the WHO-HPQ measures of absenteeism and presenteeism with other presenteeism measures, health conditions, and psychosocial job conditions in a sample of full-time employees in Japan (n=92)†

				Pearson's correlation coefficients (*p* values)							
				Other presenteeism measures			Health conditions		Psychosocial job conditions		
	Average	SD	Min, Max	Stanford Presenteeism Scale (SPS) (T1) ‡	WFun (T1)	Perceived relative presenteeism (T2)	K6 (T1)	Depression/ anxiety (T1)	Job control (T1)	Supervisor support (T1)	Coworker support (T1)
Absenteeism (4 wk)											
Absolute absenteeism T1 (hr)	−21.2	40.7	−232.0, 64.0	−0.01 (0.950)	0.014 (0.898)	−0.11 (−0.298)	−0.094 (0.373)	−0.021 (0.843)	0.09 (0.393)	−0.210* (0.044)	−0.145 (0.168)
Relative absenteeism T1	−0.14	0.27	−1.45, 0.19	−0.013 (0.935)	0.03 (0.778)	−0.091 (0.390)	−0.072 (0.496)	−0.016 (0.877)	0.074 (0.486)	−0.196 (0.061)	−0.147 (0.161)
Absenteeism (7 d)											
Absolute absenteeism T1 (hr)	−36.6	42.5	−176.0, 40.0	−0.114 (0.474)	−0.097 (0.355)	−0.011 (0.917)	−0.103 (0.330)	0.101 (0.340)	−0.002 (0.981)	−0.037 (0.724)	0.038 (0.716)
Relative absenteeism T1	−0.23	0.26	−1.10, 0.17	−0.117 (0.460)	−0.101 (0.339)	−0.012 (0.91)	−0.105 (0.321)	0.09 (0.391)	0.004 (0.968)	−0.034 (0.75)	0.042 (0.692)
Presenteeism (work performance)											
Absolute presenteeism T1 (0–100)§	60.5	17.8	0.0, 100.0	−0.304 (0.050)	−0.341** (0.001)	−0.343** (0.001)	−0.247* (0.018)	−0.07 (0.507)	0.201 (0.055)	0.025 (0.813)	0.162 (0.123)
Relative presenteeism T2 (ratio)§	1.05	0.33	0.25, 2.00	0.041 (0.794)	−0.052 (0.622)	−0.358** (<0.001)	0.061 (0.562)	−0.183 (0.081)	0.146 (0.164)	−0.029 (0.787)	0.045 (0.667)
Relative presenteeism T2 (subtract)§	1.8	20	−8.00, 8.00	−0.004 (0.981)	−0.079 (0.453)	−0.383** (<0.001)	0.043 (0.685)	−0.138 (0.190)	0.199 (0.057)	−0.033 (0.751)	0.054 (0.611)

†T1: assessed at the baseline; T2: assessed at the follow-up survey two weeks later.

‡Only asked when a respondent had any of 13 health problems (N=42).

§A greater score is indicative of better work performance.

*p<0.05, **p<0.01.

Table 3. Test-retest reliability (ICC) between two surveys with a 2-wk interval among respondents who reported no change in work performance between T1 and T2 (n=64)

	ICC	95%CI
Absenteeism:		
Absolute absenteeism (4 wk)	0.610	0.429–0.743
Relative absenteeism (4 wk)	0.527	0.336–0.690
Absolute absenteeism (7 d)	0.649	0.480–0.771
Relative absenteeism (7 d)	0.647	0.478–0.770
Presenteeism:†		
Absolute presenteeism	0.730	0.591–0.827

ICC: intraclass correlation; 95%CI: 95% confidence intervals.
†Relative presenteeism measures were not tested for the test-retest reliability because the measures were modified at T2.

measure significantly correlated with WFun and perceived relative presenteeism and with psychological distress; this measure also marginally significantly correlated with SPS and job control. These findings provided some, but limited, support for the construct validity of this measure. As hypothesized, relative presenteeism measures significantly correlated with perceived relative presenteeism. Otherwise, none of WHO-HPQ measures significantly correlated with the variables for hypothesis testing.

We found significant and moderate correlations of the absolute presenteeism measure with WFun and perceived relative presenteeism, and also its marginally significantly and moderately correlation with SPS. The findings are consistent with our hypothesis. The WHO-HPQ presenteeism measure asks presenteeism derived from any reasons, including health problems and work environment[29], while SPS and WFun assess presenteeism only due to health problem. This could explain the moderate correlation between absolute presenteeism scores of the WHO-HPQ and SPS and WFun. The WHO-HPQ presenteeism measure may capture a different aspect of presenteeism than that measured by SPS and WFun. The WHO-HPQ absolute presenteeism measures significantly correlated with K6, and marginally significantly with job control. This is in line with previous findings that presenteeism was associated with poor mental health conditions[26] and job control[27]. However, since only three of the eight correlations tested were found significant, the present study provides only limited support for the construct validity of the WHO-HPQ absolute presenteeism measure. The construct validity should be investigated further, in particular, with variables that could be more closely associated with absolute presenteeism, such as a scale of presenteeism from any reason not limited to health problems or health status (e.g., musculoskeletal symptoms). On the other hand, the absolute presenteeism measure was quite stable for a two-week period (0.73 in ICC). This test-retest reliability is better than a moderate two-week test-retest reliability (0.59 in ICC) for this measure[19] and close to that for other global performance measures (0.69–0.78 in ICC) that were previously reported among patients with rheumatic diseases[19]. The higher ICC in this study may be because we limited the sample to participants who did not have change in work performance. These findings suggest that the absolute presenteeism measure of the WMHJ version of the WHO-HPQ is reliable over a short period (e.g., two wk) and valid to measure work performance among Japanese employees.

The WHO-HPQ relative presenteeism measures (both ratio and subtract) significantly and negatively correlated with perceived relative presenteeism. The relative presenteeism measure did not correlate with SPS or WFun that are supposed to assess absolute presenteeism. Previous studies reported that the relative presenteeism was useful in predicting mental health problems in future[21, 22]. However, in this study, we did not find significant correlations between this measure and any of health conditions or psychosocial job conditions, providing little support for the construct validity. More research is needed to investigate the construct and predictive validity of the WHO-HPQ relative presenteeism measures. From our experience, it may also be better to add a small sentence to avoid a respondent misunderstanding that the question B9 that asks job performance of most workers in a job similar, not the respondent's job performance.

The average scores of absolute and relative presenteeism measures in this study were close to those reported in a previous study from Japan[22]. However, the scores were lower than those reported in a previous study in the United States, in which median scores of absolute presenteeism were between 80 and 90 among four different sample of workers[6]. Reporting presenteeism on the WHO-HPQ may be affected by norms and cultures of the workplace in a given country.

The WHO-HPQ absolute absenteeism measure significantly and negatively correlated with supervisor support, as hypothesized. The change in absolute and relative absenteeism correlated with self-reported changes in one's own work performance. The test-retest reliability (ICCs) was also moderate. However, the present study did not find much supporting evidence for the construct validity of the absolute or relative absenteeism measures. In addition, about 10% of the sample reported extremely long

or conflicting responses. Some respondents made clear mistakes in entering hours and days in the WHO-HPQ questions about absenteeism. Some full-time non-manager respondents reported that their contract hours were longer than regular work hours (i.e., 40 h per week). This may be because in Japanese culture[39], employees are often expected to work outside of their regular work hours. However, such inconsistency in reporting regular work hours among participants is likely to lead to a measurement error in relative absenteeism.

Limitations

The following limitations of the study should be noted. Our use of an Internet sample for the sake of convenience may limit the generalizability of the findings, since Internet users tend to have different sociodemographic and psychological characteristics than non-users[40, 41]. In addition, the present sample included a limited proportion of respondents with blue-collar jobs. The reliability and validity of the WHO-HPQ should be replicated and confirmed in a future study with a larger diverse sample of workers. The prevalence of depression/anxiety disorder in this sample was higher than the prevalence reported from a nationally representative survey[42], that may further limit the generalization of the findings. Calculating the response rate was impossible because the employees from registered members joined the survey in order of arrival. This may have caused selection bias in the Japanese working population. Some of the participants may not have answered the questions carefully. The time frame that we used to investigate the test-retest reliability may not be optimal, because four-week time periods of participants' first and second assessments were not same. This could be still the case even if we limited the analysis to participants who reported no change of their work performance in the past two weeks at T2. This could underestimate the test-retest reliability in our study. We did not use doctors' diagnoses of health problems. Self-reported health problems may be more associated with self-reported work performance. We did not use objective measure of absenteeism (e.g., a company record) or presenteeism (e.g., manager's evaluation of job performance of participants) to test the criterion validity. Finally, the selection of covariates to test the construct validity was arbitrary, not systematic. Some covariates may not have been appropriate for selection. This may lead to underestimation of the construct validity of the instrument. The covariates for testing the construct validity should be selected based on a systematic review of the relevant literature in future research.

Practical implications

For practical implications of the study findings, the absolute presenteeism measure of the WMHJ version of the WHO-HPQ may be used as a reliable measure of presenteeism among Japanese workers. However, this measure should be tested further for its construct validity and used with caution that it reflects presenteeism from any reasons, not like other presenteeism scales such as SPS and WFun that assess presenteeism solely from health problems. Further research is needed to clarify the validity of other measures of the WMHJ version of WHO-HPQ.

Conclusion

The study found some support for the construct validity and test-retest reliability of the absolute presenteeism measure of the WMHJ version of the WHO-HPQ among Japanese workers. Further research is needed to clarify the construct validity of other measures of this instrument.

Author Contribution

Conceptualization, N.K., A.I., and M.T.; Methodology, N.K., A.I., M.T., K.W., K.I., and D.N.; Investigation, K.W., and M.I.; Formal analysis, N.K.; Writing, N.K., A.I., M.T., K.W., K.I., I.M., and D.N.; Funding Acquisition, N.K.

Funding

This work was supported by JSPS KAKENHI Grant Number JP 18H04072.

Conflict of Interest

M.T. is an employee of the ADVANTAGE Risk Management Co., Ltd., Tokyo, Japan. Neither the funder nor the employer participated in designing the study, collecting and analyzing the data or preparing the manuscript. Otherwise, the authors declare no conflict of interest.

References

1) Wang PS, Simon G, Kessler RC (2003) The economic burden of depression and the cost-effectiveness of treatment. Int J Methods Psychiatr Res **12**, 22–33.
2) Kessler RC, Berglund PA, Coulouvrat C, Hajak G, Roth T, Shahly V, Shillington AC, Stephenson JJ, Walsh JK (2011) Insomnia and the performance of US workers: results from the America insomnia survey. Sleep (Basel) **34**, 1161–71.
3) Stewart WF, Ricci JA, Chee E, Hahn SR, Morganstein

D (2003) Cost of lost productive work time among US workers with depression. JAMA **289**, 3135–44.
4) Loeppke R, Taitel M, Richling D, Parry T, Kessler RC, Hymel P, Konicki D (2007) Health and productivity as a business strategy. J Occup Environ Med **49**, 712–21.
5) Wang PS, Simon GE, Avorn J, Azocar F, Ludman EJ, McCulloch J, Petukhova MZ, Kessler RC (2007) Telephone screening, outreach, and care management for depressed workers and impact on clinical and work productivity outcomes: a randomized controlled trial. JAMA **298**, 1401–11.
6) Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, Pronk N, Simon G, Stang P, Ustun TB, Wang P (2003) The World Health Organization Health and Work Performance Questionnaire (HPQ). J Occup Environ Med **45**, 156–74.
7) Ospina MB, Dennett L, Waye A, Jacobs P, Thompson AH (2015) A systematic review of measurement properties of instruments assessing presenteeism. Am J Manag Care **21**, e171–85.
8) Despiégel N, Danchenko N, François C, Lensberg B, Drummond MF (2012) The use and performance of productivity scales to evaluate presenteeism in mood disorders. Value Health **15**, 1148–61.
9) Roy JS, Desmeules F, MacDermid JC (2011) Psychometric properties of presenteeism scales for musculoskeletal disorders: a systematic review. J Rehabil Med **43**, 23–31.
10) Nagata T, Mori K, Ohtani M, Nagata M, Kajiki S, Fujino Y, Matsuda S, Loeppke R (2018) Total health-related costs due to absenteeism, presenteeism, and medical and pharmaceutical expenses in Japanese employers. J Occup Environ Med **60**, e273–80.
11) Koopman C, Pelletier KR, Murray JF, Sharda CE, Berger ML, Turpin RS, Hackleman P, Gibson P, Holmes DM, Bendel T (2002) Stanford presenteeism scale: health status and employee productivity. J Occup Environ Med **44**, 14–20.
12) Lerner D, Amick BC 3rd, Rogers WH, Malspeis S, Bungay K, Cynn D (2001) The Work Limitations Questionnaire. Med Care **39**, 72–85.
13) Fujino Y, Uehara M, Izumi H, Nagata T, Muramatsu K, Kubo T, Oyama I, Matsuda S (2015) Development and validity of a work functioning impairment scale based on the Rasch model among Japanese workers. J Occup Health **57**, 521–31.
14) Nagata T, Fujino Y, Saito K, Uehara M, Oyama I, Izumi H, Kubo T (2017) Diagnostic accuracy of the Work Functioning Impairment Scale (WFun): a method to detect workers who have health problems affecting their work and to evaluate fitness for work. J Occup Environ Med **59**, 557–62.
15) Kono Y, Matsushima E, Uji M (2014) Psychometric properties of the 25-item Work Limitations Questionnaire in Japan: factor structure, validity, and reliability in information and communication technology company employees. J Occup Environ Med **56**, 184–8.
16) Fujino Y, Liu N, Chimed-Ochir O, Okawara M, Ishimaru T, Kubo T (2019) Cross-cultural validation of the work functioning impairment scale (WFun) among Japanese, English, and Chinese versions using Rasch analysis. J Occup Health **61**, 464–70.
17) Kessler RC, Ames M, Hymel PA, Loeppke R, McKenas DK, Richling DE, Stang PE, Ustun TB (2004) Using the World Health Organization Health and Work Performance Questionnaire (HPQ) to evaluate the indirect workplace costs of illness. J Occup Environ Med **46** Suppl, S23–37.
18) AlHeresh R, LaValley MP, Coster W, Keysor JJ (2017) Construct validity and scoring methods of the World Health Organization: Health and Work Performance Questionnaire among workers with arthritis and rheumatological conditions. J Occup Environ Med **59**, e112–8.
19) Leggett S, van der Zee-Neuen A, Boonen A, Beaton DE, Bojinca M, Bosworth A, Dadoun S, Fautrel B, Hagel S, Hofstetter C, Lacaille D, Linton D, Mihai C, Petersson IF, Rogers P, Sergeant JC, Sciré C, Verstappen SM, At-work Productivity Global Measure Working Group (2016) Test-retest reliability and correlations of 5 global measures addressing at-work productivity loss in patients with rheumatic diseases. J Rheumatol **43**, 433–9.
20) Pournik O, Ghalichi L, Tehrani Yazdi AR, Tabatabaee SM, Ghaffari M, Vingard E (2012) Reliability and validity of Persian version of World Health Organization health and work performance questionnaire in Iranian health care workers. Int J Occup Environ Med **3**, 33–8.
21) Suzuki T, Miyaki K, Sasaki Y, Song Y, Tsutsumi A, Kawakami N, Shimazu A, Takahashi M, Inoue A, Kurioka S, Shimbo T (2014) Optimal cutoff values of WHO-HPQ presenteeism scores by ROC analysis for preventing mental sickness absence in Japanese prospective cohort. PLoS One **9**, e111191.
22) Suzuki T, Miyaki K, Song Y, Tsutsumi A, Kawakami N, Shimazu A, Takahashi M, Inoue A, Kurioka S (2015) Relationship between sickness presenteeism (WHO-HPQ) with depression and sickness absence due to mental disease in a cohort of Japanese workers. J Affect Disord **180**, 14–20.
23) Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC (2010) The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol **63**, 737–45.
24) Kawakami N, Takeshima T, Ono Y, Uda H, Hata Y, Nakane Y, Nakane H, Iwata N, Furukawa TA, Kikkawa T (2005) Twelve-month prevalence, severity, and treatment of common mental disorders in communities in Japan: preliminary finding from the World Mental Health Japan Survey 2002–2003. Psychiatry Clin Neurosci **59**, 441–52.
25) Tsuchiya M, Kawakami N, Ono Y, Nakane Y, Nakamura Y, Fukao A, Tachimori H, Iwata N, Uda H, Nakane H,

Watanabe M, Oorui M, Naganuma Y, Furukawa TA, Kobayashi M, Ahiko T, Takeshima T, Kikkawa T (2012) Impact of mental disorders on work performance in a community sample of workers in Japan: the World Mental Health Japan Survey 2002–2005. Psychiatry Res **198**, 140–5.

26) Duijts SFA, Kant I, Swaen GMH, van den Brandt PA, Zeegers MPA (2007) A meta-analysis of observational studies identifies predictors of sickness absence. J Clin Epidemiol **60**, 1105–15.
27) Nagami M, Tsutsumi A, Tsuchiya M, Morimoto K (2010) Job control and coworker support improve employee job performance. Ind Health **48**, 845–51.
28) Nunnally JC (1964) Educational measurement and evaluation, McGraw-Hill, New York.
29) Aboagye E, Jensen I, Bergström G, Hagberg J, Axén I, Lohela-Karlsson M (2016) Validity and test-retest reliability of an at-work production loss instrument. Occup Med (Lond) **66**, 377–82.
30) Beemster TT, van Velzen JM, van Bennekom CAM, Reneman MF, Frings-Dresen MHW (2019) Test-retest reliability, agreement and responsiveness of productivity loss (iPCQ-VR) and healthcare utilization (TiCP-VR) questionnaires for sick workers with chronic musculoskeletal pain. J Occup Rehabil **29**, 91–103.
31) Paschoalin HC, Griep RH, Lisboa MTL, Bandeira de Mello DC (2013) Transcultural adaptation and validation of the Stanford Presenteeism Scale for the evaluation of presenteeism for Brazilian Portuguese. Rev Lat Am Enfermagem **21**, 388–95.
32) Walker TJ, Tullar JM, Diamond PM, Kohl HW 3rd, Amick BC 3rd (2017) Validity and Reliability of the 8-Item Work Limitations Questionnaire. J Occup Rehabil **27**, 576–83.
33) Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF (2003) A comparison of two time intervals for test-retest reliability of health status instruments. J Clin Epidemiol **56**, 730–5.
34) Cohen J (1988) Statistical power analysis for the behavioral sciences, L. Erlbaum Associates, Hillsdale.
35) Wada K, Moriyama M, Narai R, Tahara H, Kakuma R, Satoh T, Aizawa Y (2007) [The effect of chronic health conditions on work performance in Japanese companies]. Sangyo Eiseigaku Zasshi **49**, 103–9 (in Japanese).
36) Makishima M, Fujino Y, Kubo T, Izumi H, Uehara M, Oyama I, Matsuda S (2018) Validity and responsiveness of the work functioning impairment scale (WFun) in workers with pain due to musculoskeletal disorders. J Occup Health **60**, 156–62.
37) Furukawa TA, Kawakami N, Saitoh M, Ono Y, Nakane Y, Nakamura Y, Tachimori H, Iwata N, Uda H, Nakane H, Watanabe M, Naganuma Y, Hata Y, Kobayashi M, Miyake Y, Takeshima T, Kikkawa T (2008) The performance of the Japanese version of the K6 and K10 in the World Mental Health Survey Japan. Int J Methods Psychiatr Res **17**, 152–8.
38) Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SLT, Walters EE, Zaslavsky AM (2002) Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med **32**, 959–76.
39) Nakane C (1973) Japanese Society, Penguin, Harmondsworth.
40) Tsuboi S, Yoshida H, Ae R, Kojo T, Nakamura Y, Kitamura K (2015) Selection bias of Internet panel surveys: a comparison with a paper-based survey and national governmental statistics in Japan. Asia Pac J Public Health **27**, NP2390–9.
41) Whitehead L (2011) Methodological issues in Internet-mediated research: a randomized comparison of internet versus mailed questionnaires. J Med Internet Res **13**, e109.
42) Nishi D, Ishikawa H, Kawakami N (2019) Prevalence of mental disorders and mental health service use in Japan. Psychiatry Clin Neurosci **73**, 458–65.

Appendix 1.

1. The World Health Organization Health and Work Performance Questionnaire (WHO-HPQ) short version
https://www.hcp.med.harvard.edu/hpq/ftpdir/absenteeism%20presenteeism%20scoring%20050107.pdf

B3. About how many hours did you work in the past seven days? (If more than 97, enter 97.) Number of hours (00–97)

B4. How many hours does your employer expect you to work in a typical seven-day week? (If the number varies, estimate the average. If more than 97, enter 97.) Number of hours (00–97)

B5. Now please think of your work experiences over the past four weeks (28 days). In the spaces provided below, write the number of days you spent in each of the following work situations.

In the past four weeks (28 days), how many days did you...

B5a. ...miss an entire workday because of problems with your physical or mental health? (Please include only days missed for your own health, not someone else's.)

B5b. ...miss an entire workday for any other reason (including vacation)?

B5c. ...miss part of a workday because of problems with your physical or mental health? (Please include only days missed for your own health, not someone else's.)

B5d. ...miss part of a workday for any other reason (including vacation)?

B5e. ...come in early, go home late, or work on your day off?

B6. About how many hours in total did you work in the past four weeks (28 days)?

B9. On a scale from 0 to 10 where 0 is the worst job performance anyone could have at your job and 10 is the performance of a top worker, how would you rate the usual performance of most workers in a job similar to yours?

B10. Using the same 0-to-10 scale, how would you rate your usual job performance over the past year or two?

B11. Using the same 0-to-10 scale, how would you rate your overall job performance on the days you worked during the past four weeks (28 days)?

2. A question on perceived relative presenteeism from the WHO-HPQ Clinical Trials Baseline Version

A13. How would you compare your overall job performance on the days you worked during the past seven days with the performance of most other workers who have a similar type of job? (Circle the number.)

1. You were much better than other workers
2. You were somewhat better than other workers
3. You were a little better than other workers
4. You were about average
5. You were a little worse than other workers
6. You were somewhat worse than other workers
7. You were much worse than other workers

Appendix 2. List of measures of absenteeism and presenteeism in the World Health Organization Health and Work Performance Questionnaire (WHO-HPQ) short version

	Description	Calculation formula (refer the items used, B4–B11 to the Appendix 1)
Absenteeism (4 wk)		
Absolute absenteeism (hr)	Difference (deficit) of actual work hours compared to standard working hours in the last four weeks	4 × standard working hours per week (B4) − Hours worked in the last four weeks (B6)
Relative absenteeism	Proportion of difference (deficit) of actual work hours compared to standard working hours relative to standard working hours in the last four weeks	[4 × Standard working hours per week (B4) − Hours worked in the last four weeks (B6)]/[4 × Standard working hours per week (B4)]
Absenteeism (7 d)		
Absolute absenteeism (hr)	Difference (deficit) of actual work hours compared to standard working hours in the last seven days	4 × Standard working hours per week (B4) − 4 × Hours worked in the last seven days (B3)
Relative absenteeism		[Standard working hours per week (B4) − 4 × Hours worked in the last seven days (B3)]/[4 × Standard working hours per week (B4)]
Presenteeism (work performance)		
Absolute presenteeism	Work performance (i.e, the quality of work) rated by a respondent	10 × Self-reported work performance (B11, ranging 0–10)
Relative presenteeism (ratio)	Work performance (i.e, the quality of work) rated by a respondent relative to work performance of other most workers at the same job also rated by the respondent, calculated in a ratio	10 × [Self-reported work performance (B11) − Work performance of most workers at the same job (B9)], restricted to the range of 0.25 to 2.
Relative presenteeism (subtraction)	Work performance (i.e, the quality of work) rated by a respondent relative to work performance of other most workers at the same job also rated by the respondent, calculated in a difference	Self-reported work performance (B11)/ Work performance of most workers at the same job (B9)

Appendix 3. List of hypotheses tested for the construct validity of the WHO-HPQ measures ("X" indicates a hypothesized correlation for a set of column and raw variables)†

	Other presenteeism measures			Health conditions		Psychosocial job conditions		
WHO-HPQ measures	Stanford Presenteeism Scale (SPS) (T1)	WFun (T1)	Perceived relative presenteeism (T2)	K6 (T1)	Depression/ anxiety (T1)	Job control (T1)	Supervisor support (T1)	Coworker support (T1)
Absenteeism (absolute)				X	X	X	X	X
Absenteeism (relative)				X	X	X	X	X
Presenteeism (absolute)	X	X	X	X	X	X	X	X
Presenteeism (relative)			X	X	X	X	X	X

† T1: assessed at the baseline; T2: assessed at the follow-up survey two weeks later.