J Occup Health 2015; 57: 521–531

Journal of Occupational Health

Development and validity of a work functioning impairment scale based on the Rasch model among Japanese workers

Yoshihisa FUJINO[1], Masamichi UEHARA[2], Hiroyuki IZUMI[3], Tomohisa NAGATA[4], Keiji MURAMATSU[1], Tatsuhiko KUBO[1], Ichiro OYAMA[5] and Shinya MATSUDA[1]

[1]Department of Preventive Medicine and Community Health, University of Occupational and Environmental Health, Japan, [2]Health Support Center, Brother Industries, Ltd., Japan, [3]Department of Ergonomics, Institute of Industrial Ecological Sciences, University of Occupational and Environmental Health, Japan, [4]Department of Occupational Health Practice and Management, Institute of Industrial Ecological Sciences, University of Occupational and Environmental Health, Japan and [5]Corporate ESH & QA, Asahi Kasei Corporation, Japan

Abstract: Development and validity of a work functioning impairment scale based on the Rasch model among Japanese workers: Yoshihisa FUJINO, *et al.* Department of Preventive Medicine and Community Health, University of Occupational and Environmental Health, Japan—Objectives: The purpose of this study was to develop a new work functioning impairment scale (WFun) and examine its validity. **Methods:** The WFun was developed based on the Rasch model, which consists of seven items. We conducted a pilot study (n=1,000) using an Internet investigation and a field study (n=1,294) in a manufacturing industry, and we additionally collected data from six workplaces from other industries. This series of studies was examined with a Rasch model analyses including item fit statistics as well as hypothesis testing. Convergent validity was used to examined the association of the WFun with the Stanford Presenteeism Scale, SF-8, Work Ability Index, and several types of job disruptions. We also examined differential test functioning. **Results:** All the items showed adequate fit (infit mean-square statistics <1.5). The item reliability was 0.98, and the item separation index was 6.37. The person reliability was 0.86, and the person separation index was 2.32. All tests for convergent validity showed significant differences. All *p* values derived from ANOVA were highly significant ($p<0.001$). No differential test function was observed between groups by age, sex, or job type or between various samples from different workplaces. The intraclass correlation of the estimated Rasch measurements from these groups was 0.99 (95% CI: 0.976−0.992). **Conclusions:** The WFun was confirmed to show good fit to a Rasch model and construct validity. Given that its good fit indicates specific objectivity, this tool will be useful in assessing the ability of individuals to function at work and in evaluating group levels for benchmarking.
(J Occup Health 2015; 57: 521–531)

Key words: Occupational health, Patient-reported outcome, Presenteeism, Quality of life, Rasch model, Work limitations

There has been an increase in interest in presenteeism, which refers to "showing up for work when one is ill"[1, 2]. Three major issues have raised awareness of presenteeism. First, it is natural for employers to be concerned about worker productivity. Productivity is a major determinant of business growth and the stability of economic activities. Loss of productivity is traditionally captured by absenteeism, or time away from work. However, recent reports have claimed that presenteeism determines the major part of productivity loss, not absenteeism[3, 4]. Second, the cost of employee illness is a major concern of employers, and presenteeism is recognized as a major hidden part of this cost[3, 4]. In general, the employer is responsible for the employee's medical treatment costs, at least partly, although there are regional variations. In addition, given the broad understanding that the employer's greatest asset is human capital, many employers provide formal and informal health-care programs. While these are accounted for as a direct cost, many employers have shown interest in the indirect costs included in absenteeism and presenteeism. Third, evaluation of the economic impact of pharmaceutical

Received May 21, 2015; Accepted Jul 10, 2015
Published online in J-STAGE Sep 4, 2015
Appendix tables 1~4, 6, 7: refer to J-STAGE: https://www.jstage.jst.go.jp/browse/joh
Correspondence to: Y. Fujino, Department of Preventive Medicine and Community Health, University of Occupational and Environmental Health, Japan. 1-1 Iseigaoka, Yahatanishi-ku, Kitakyushu, Fukuoka 807-8555, Japan (e-mail: zenq@med.uoeh-u.ac.jp)

and other medical interventions now includes indirect costs, which improves the cost-effectiveness of interventions[5, 6].

These issues arising from presenteeism lead to questions about the monetary value of productivity loss due to presenteeism. Accordingly, many self-reported questionnaires to assess presenteeism have been developed and are used to estimate the monetary value of productivity loss. Brooks *et al*. listed and reviewed 16 presenteeism tools, such as the Work Limitation Questionnaire, Work Productivity and Activity Impairment Questionnaire, and Stanford Presenteeism Scale (SPS)[7, 8]. Although these tools were developed under different concepts and methodologies, many of them emphasize the assessment of productivity by accounting for various perspectives, including time not on task, quality of work (mistakes, peak performance, injury rates, etc.), quantity of work, and personal factors (social, mental, physical, emotional, etc.)[9]. Some of these questionnaires have reported good validity consistent with the objective quantification of productivity[10, 11].

We argue that there is a need to complement these previous tools for presenteeism with tools based on the perspective of worker health management in occupational health practice. It is important to distinguish between productivity and the ability to function at work. Productivity is regarded as the output of the production function, where the ability to function and other production factors are the input. Previous presenteeism questionnaires appear to have captured productivity but to have unintentionally ignored this distinction. It is possible for a worker who is suffering from health problems to retain his productivity by receiving appropriate support from the workplace. Alternatively, the work environment and job design, such as assembly-line systems, can enable the impact of a worker's health problems on productivity to be minimized. Accordingly, assessment of productivity, a close concept of labor output, may overlook issues that relate to occupational health practice.

In addition, there is growing recognition that a patient's ability to work is a legitimate clinical outcome in clinical settings. Waddell and Burton[12] have proposed that the goals of the biomedical model are to alleviate symptoms, in contrast to the aims of clinical management subject to the biopsychosocial model, which should aim to manage symptoms and bring about a return of function.

We therefore argue for the importance of assessing health-related ability to function at work. Functioning refers to "the ability of the individual to perform particular defined tasks"[13]. This concept is closer to quality of life, but it is not the same as worker performance and behavior. In other words, we attempted to translate the severity of worker health problems into the degree to which a worker will experience a limitation of functioning at work due to these problems.

We adopted latent trait theory to measure these unobservable characteristics. This refers to "constructs" or "latent traits" in psychology[14]. In contrast, previous presenteeism questionnaires were developed based on classical test theory.

Whether based on classical test theory or latent trait theory, this type of questionnaire is called a health-related patient-reported outcome. The development and validation process of these questionnaires has recently been standardized[15, 16]. We developed an original 7-item questionnaire to assess a worker's health-related ability to function at work in accordance with consensus-based standards for the selection of health measurement instruments (COSMIN)[16]. We named this questionnaire the "work functioning impairment scale" (WFun). Here, we report the process of development and validity of the WFun.

Subjects and Methods

This study consisted of three stages, a development stage, a pilot-testing stage, and a field-testing stage, in accordance with the process proposed by de Vet *et al*.[15] (Appendix 1). This study was approved by the Ethics Committee of the University of Occupational and Environmental Health, Kitakyushu, Japan.

In the development stage, we conceptualized what we wanted to measure, and generated item candidates. Content validity, including face validity, was checked via a focus group discussion by the authors, six occupational physicians, and colleagues who were not medical experts.

We conceptualized that work impairment is a state in which the worker's ability to function at work is impaired by health problems. The questionnaire aims to measure this construct. Measuring this construct will enable identification of workers who are suffering from work impairment and assessment of the degree to which their ability to function at work is damaged. The details of the background of this construct are discussed in the Discussion section.

Generating item candidates

The measurement theory of constructs distinguishes the reflective and formative models[17]. Both the latent trait model and classical test theory are based on the reflective model[15]. To generate item candidates, we summarized the concept of work impairment representing a reflective model and a formative model (Appendix 2). The reflective part was assumed to consist of four subscales, namely sociability, execution of work, physical and mental tolerance, and motivation. We generated 30 items related to each subscale

via focus group discussion (Appendix 3). In addition, we generated 7 items related to particular job disruptions in the formative part according to the work transition model proposed by Gignac[18]. In the work transition model, job disruptions lead to productivity loss and absenteeism and further to work changes and an exit from the work force. The validation study used the following 7 items: "I have changed my work routine", "I have postponed a troublesome task", "I have changed a work schedule", "I have asked other staff to undertake a part/all of my task", "Work content or amount has changed", "My work hours have been changed", and "I could not take on some work due to poor health conditions". Five response categories were set: 1, not at all; 2, one or more days a month; 3, about one day a week; 4, two or more days a week; and 5, almost every day.

Pilot-testing stage

The pilot study was carried out by an Internet investigation targeting 1,000 registered monitors. We requested a commercial testing company to carry out an Internet test user investigation. Of 2 million registered Internet test users, an email requesting participation was sent to approximately 20,000. Screening items included the statements "I am currently employed" and "I have some health issues". Registered users who matched the screening items were assigned to 10-year age groups (20 s, 30 s, 40 s, 50 s and 60 s) by sex, with 100 people in each group, and the first 1,000 responses were collected to ensure full recruitment of all groups. Respondents were asked about their age, sex, occupation, and employment type, and they were asked to complete the prepared question items we generated, the SPS, and 8-item Short Form Health Survey (SF-8). All items had to be answered, so there were no missing answers. A follow-up investigation 3 months later was used to determine changes in employment conditions. Of the 1,000 subjects, 867 subjects responded to this follow-up survey.

The SPS has been validated for the assessment of self-reported absence, work impairment, and loss of work attributable to a diagnosed primary health condition[8, 19, 20]. A work impairment scale (WIS) was developed using the responses to 10 SPS items asking about the frequency or intensity of particular manifestations of the primary health condition and the effect of the manifestations on work. The WIS is considered to measure the degree to which a health condition diminishes subject's input into their job, such as their energy, ability to work with colleagues, and ability to focus. In addition, the SPS includes a work output score (WOS), a single-item global assessment that asks the subject to estimate the percent of "usual" productivity they could achieve during 4 weeks despite their primary health condition. Yamashita *et al*. reported about the reliability and validity of the Japanese version of the SPS[21].

The SF-8 is a generic quality of life tool. It includes 8 items, and each item concerns a specific domain of the subject's physical and mental well-being. The tool has been comprehensively examined with regard to its validity, responsiveness, and reliability[22–24].

Assessing the conceptual structure of the reflective model

As an initial step of the pilot stage, we assessed the conceptual structure of our assumed reflective model and confirmed that the 30 candidate items we generated were related to the model. This may provide a rationale for item generation. An exploratory factor analysis and confirmatory factor analysis were performed using the pilot-study data. Given ordinal categorical data, exploratory factor analysis with promax rotation was performed based on polychoric correlations. For confirmative factor analysis, structural equation modeling was performed. We employed a 4-factor 2-stage model (shown as the reflective model in Appendix 2) based on the development concept which consists of sociability, execution of work, physical and mental tolerance, and motivation, which was also confirmed by the exploratory factor analysis.

Item selection using Rasch model analysis

The Rasch model is a widely used statistical method for estimating latent abilities by studying item responses[25–28]. It provides a mathematical framework based on the assumption of unidimensionality and local independence against which test data can be compared. Estimates and standard errors of person ability and item difficulty are calculated on a common equal-interval logit scale. The Rasch model uses one parameter to estimate person ability (the number of correct responses by a person) and item difficulty (the number of correct responses to an item) to evaluate the probability that person n will succeed in an item.

We conducted a preliminary Rasch model analysis for 30 items to reduce the number of items (Appendix 4). Items with either outfit >1.5 or infit >1.5 were initially excluded. We then selected items that were non-disease-specific and non-job-specific, accounting for word nuances. Seven items were finally selected and then further assessed in depth by the Rasch model. The correlation between the raw sum score of the final 7 items and that of the 30 items was 0.98.

We then called this 7-item questionnaire the

"work functioning impairment scale" (WFun), which consisted of the following items: I haven't been able to behave socially, I haven't been able to maintain the quality of my work, I have had trouble thinking clearly, I have taken more rests during my work, I have felt that my work isn't going well, I haven't been able to make rational decisions, and I haven't been proactive about my work. (The Japanese version of the WFun is available from the corresponding author upon request.)

Rasch analysis was performed using WINSTEPS version 3.81.0[29]. Data were fitted to the Rasch rating scale model using joint maximum likelihood estimation, in which the rating scale structure was defined to be equal for all items[30]. Fit statistics in combination with principal components analysis of the residuals were used to test the unidimensionality assumption. The fit criteria for this study were set at 1.5 for the infit and outfit mean-square statistics, respectively[31]. The criterion used to confirm unidimensionality was that the first contrast had to have an eigenvalue of <2. Local independence was checked by residual correlations between the items.

The reliability of the instrument was examined using the person separation reliability statistics in the Rasch analysis. The person separation index represents the ability of a given test to separate persons into different strata[28]. The index must exceed 2 to attain the desired level of reliability of at least 0.80[32].

Rating scale analysis included category frequencies, average measures, category fit, and threshold estimates. According to reported guidelines[33], an item is considered appropriate for the rating scale when the rating scale has an outfit mean-square statistics of <2. The guidelines also indicate that a five-category rating scale requires advances of at least 1.0 logit between step calibrations.

Differential test functioning

In theory, Rasch modeling assumes that the measures produced by a model are not sample dependent for the test items, a property called "specific objectivity"[28, 30, 34]. To examine this property, we evaluated differential test functioning. First, we estimated Rasch measurements corresponding to raw scores separately for the following groups in the pilot study: sex (men and women), age category groups (20 s, 30 s, 40 s, 50 s, and 60 s), and job type (mainly desk work, jobs mainly involving interpersonal communication, and mainly labor). Second, we collected an additional six samples from different workplaces in a variety of types of industry (Appendix 4). We then estimated Rasch measurements corresponding to the raw scores separately for the pilot study, field study, and the six additional samples. Absolute consistency was examined by ICC (2,1), which is a form of intraclass correlation[35].

Hypothesis testing

According to COSMIN, hypothesis testing must be examined when a gold standard is not available. Hypothesis testing includes convergent and discriminant validity: convergent validity examines the positive correlation of the developed measurement with similar constructs, while discriminant validity examines the lack of correlation of the developed measurement with groups regarded as having no difference in the level of work functioning impairment. Mean raw scores between groups are compared by ANOVA. In addition, we used ANOVA for a linear trend test to examine convergent validity. We also hypothesized that the mean WFun score would be 14 or lower in the healthiest group of each category, and assumed it would be 21 or higher in the poorest health group of each category. The WFun score ranges from 7 to 35, and approximately 50% of the subjects scored 14 points or lower, while approximately 20% scored 21 points or higher.

1) Hypothesis testing in the pilot study

We examined convergent and discriminant validity using the pilot study data and the WIS, WOS from the SPS, SF-8, and the seven types of job disruption that we assumed were based on the formative model for the test of convergent validity. The WIP and SF-8 were classified into five categories, and the WOS was classified into four categories. We adopted sex, age, job type, employment type, and annual income for the test of discriminant validity.

2) Hypothesis testing in the field study

In according with the guidance, we further conducted a field study. Subjects were collected from a manufacturing industry that produced air conditioner machinery. Approximately 1,294 subjects responded to a self-administered questionnaire that included basic characteristics (age, sex, and job type), the WFun, the WAI, and the job disruptions that we assumed based on the formative model. Discriminant validity examined the association of the WFun with sex, age, and job type, and convergent validity examined the association of the WFun with the WAI. The WAI questionnaire is a standardized instrument used in both research and practice in occupational health. It consists of seven items, and its score is classified as poor, moderate, good, or excellent.

Results

Assessing the conceptual structure of the reflective model

Subject characteristics in the pilot study are shown in Appendix 5. Exploratory factor analysis extracted

four factors (Appendix 6) that appropriately expressed the hypothesized subscales, namely sociability, execution of work, physical and mental tolerance, and motivation. Consistency within each subscale was good, with Cronbach's α scores of 0.94, 0.94, 0.90, and 0.96, respectively.

We also performed a confirmatory factor analysis for the second-order factor model with work impairment. This confirmatory factor analysis revealed that the a priori hypothesized four subscale structure had an adequate fit, given a comparative fit index (CFI) of 0.96, root mean square error of approximation (RMSEA) of 0.06, and a standardized root mean square residual (SRMR) of 0.03. These fit indices reveal that the model exhibited good fit according to the guidelines proposed by Hu, in which a CFI close to 0.95 or higher, RMSEA close to 0.06 or lower, and SRMR close to 0.08 or lower are representative of good-fitting models[36]. Further, good fitting of the second-order factor model implies that the score of the four factors can be combined into one overall score for work functioning[37].

Table 1. Item fit statistics of Rasch analysis for seven items

Item	Measure	SE	Infit mean-square	Outfit mean-square
4	–0.29	0.05	1.25	1.26
9	–0.20	0.05	0.94	1.05
11	–0.38	0.05	0.85	0.86
15	0.39	0.05	1.39	1.38
19	–0.15	0.05	0.88	0.85
21	0.40	0.05	0.86	0.80
26	0.23	0.05	0.84	0.79

Rasch model analysis

The individual-item fit statistics for the seven items are presented in Table 1. All the items showed adequate fit. Item reliability was 0.98, and the item separation index was 6.37. Person reliability was 0.86, and the person separation index was 2.32. These values imply good reliability of person and item.

A principal components analysis of the residuals showed that the Rasch dimension explained 62% of the variance in the data. The largest secondary dimension explained 7.8% with the eigenvalue of 1.4, which implies that unidimensionality is satisfied. No residual correlations exceeded 0.3, implying local independency[38].

The rating scale analysis is summarized in Table 2. Average measures advanced monotonically with category. The threshold increased more than 1.0 logit between categories, and the infit and outfit statistics appeared appropriate. These findings imply that the category rating scale worked well.

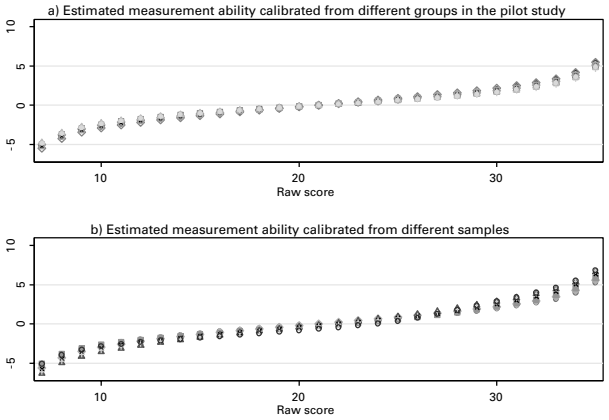
Differential test functioning

Figure 1 shows the estimated Rasch measurements corresponding to the raw scores, which were separately estimated for each subgroup. The upper figure shows 11 overlapped lines of the estimated Rash measurements from subgroups in the pilot study, namely for total subjects, sex (men and women), age group (20 s, 30 s, 40 s, 50 s, and 60 s), and job type (office work, interpersonal communication, and manual work). The ICC (2,1) was 0.99 (95% CI: 0.991–0.997).

The lower figure shows 8 overlapped lines of the estimated Rasch measurements from different samples and companies for the pilot study, field study, and six workplaces. The ICC (2,1) was 0.99 (95% CI:

Table 2. Summary of the rating scale analysis

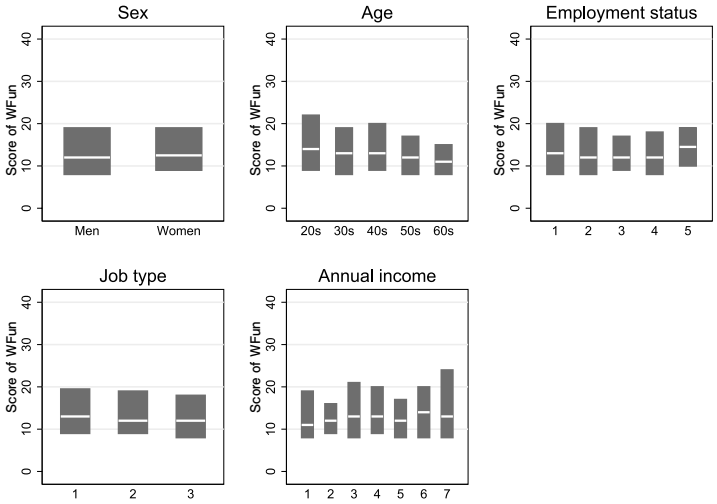
Category label (score)	Observed count %	Observed average	Infit mean-square	Outfit mean-square	Threshold
Pilot study (n=1,000)					None
Not at all (1)	45	–3.82	1.25	1.16	NONE
One or more days a month (2)	27	–2.36	0.83	0.87	–3.04
About one day a week (3)	15	–0.48	0.86	0.84	–0.84
Two or more days a week (4)	9	1.06	0.94	0.94	0.74
Almost everyday (5)	5	2.75	1.27	1.21	3.15
Field study (n=1,294)					
Not at all (1)	36	–2.65	1.27	1.2	None
One or more days a month (2)	29	–1.62	0.71	0.78	–2.43
About one day a week (3)	19	–0.33	0.75	0.73	–0.55
Two or more days a week (4)	11	0.68	0.95	1.04	0.60
Almost everyday (5)	4	1.71	1.49	1.53	2.38



The upper figure shows 11 overlapped lines of the estimated Rash measurements from subgroups in the pilot study, namely for total subjects, sex (men and women), age group (20s, 30s, 40s, 50s, and 60s), and job type (office work, interpersonal communication, and manual work). The ICC(2,1) was 0.99 (95% CI: 0.991–0.997).

The lower figure shows 8 overlapped lines of the estimated Rasch measurements from different samples and companies for the pilot study, field study, and six workplaces shown in Table 2. The ICC(2,1) was 0.99 (95% CI: 0.976–0.992).

Fig. 1. Differential test functioning by characteristics and samples.



The box indicates the interquartile range, and the white line indicates the median.

Employment status:1, full-time; 2, non-regular employee; 3, part-time; 4, self-employed; 5, contract(outsourcing)

Job type: 1, mainly desk work; 2, mainly jobs involves interpersonal communication; 3, mainly labor

Annual income: 1, <1 million JPY; 2, 1–2 million JPY; 3, 2–3 million JPY; 4, 3–5 million JPY; 5, 5–7 million JPY; 6, 7–10 million JPY; 7, 10 million JPY<

Fig. 2. Discriminant validity by sex, age, employment status, job type, and income among the pilot study subjects.

0.976–0.992). These results imply no differential test functioning.

Discriminant validity

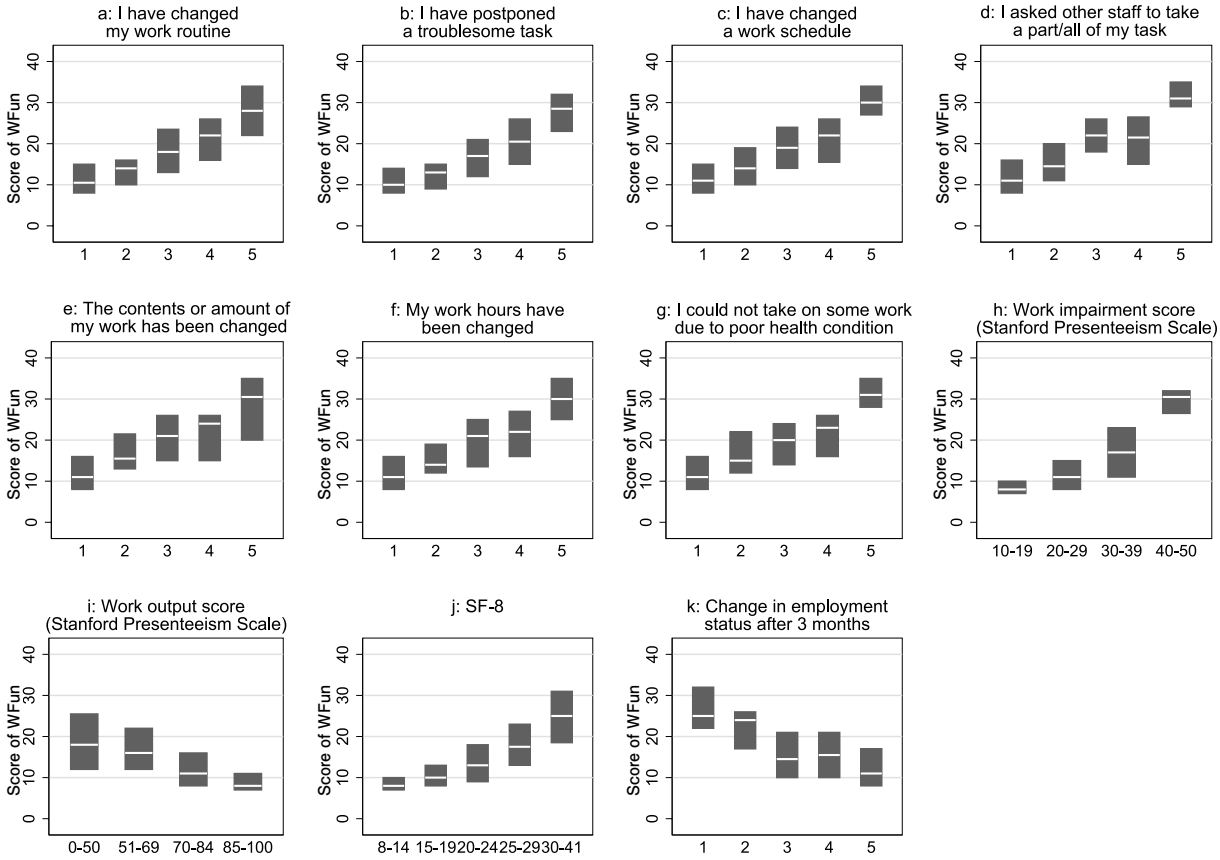
There were no significant differences in raw scores between the groups by sex, age category, employment type, job type, or annual income (Fig. 2 and Appendix 7). The p values derived from ANOVA were 0.58 for sex, 0.17 for employment type, 0.28 for job type, and 0.21 for annual income. A significant difference was seen in raw scores between age groups (p<0.001), and subjects aged 50 and over showed a lower score.

Convergent validity

All tests for convergent validity showed significant differences. Results from the pilot study are shown in Fig. 3 and Appendix 7, and those from the field study are shown in Table 3. All p values for trend were highly significant (p<0.001). In Table 3, the mean raw scores were 14 or less in the healthiest group in each category, while they were 21 or higher in the poorest health group.

Discussion

We developed a seven-item tool named the WFun to measure the degree to which the ability to function at work is impaired by health problems. The WFun



The box indicates the interquartile range, and the white line indicates the median.

For (a) to (g): 1, not at all; 2, one or more days a month; 3, about one day a week; 4, two or more days a week; 5, almost everyday

For (k): 1, I missed work more than 1 day due to bad health condition; 2, I took a leave of absence; 3, I left a job and currently not working; 4, I moved to a different job; 5, non of the above.

Fig. 3. Convergent validity by job disruptions, Stanford Presenteeism scale, SF-8, and change in employment status in the pilot study.

was confirmed to show good fit to a Rasch model and construct validity. No differential test function was observed between groups by age, sex, or job type or between various samples from different workplaces.

We developed the concept of the WFun based on a health-related conceptual model proposed by Wilson and Cleary[13]. This conceptual model distinguishes different levels of measurements from biological to overall quality of life. Biological and physiological disruptions lead to a variety of symptoms, which impact the individual's ability to function. We conceptualized that the construct we want to measure focuses on work impairment, which is a particular part of functioning in daily life.

We developed the WFun based on the Rasch model and confirmed that it had a good fit to the Rasch model. There are several benefits to the Rasch model if the data fit the model[39]. First, the model simply assesses a unique trait, which is referred to as a latent trait. The model also assumes that the probability of a particular person interaction (with regard to rating high or low) is determined only by how difficult the item is, and the subject's ability.

Second, the degree to which the trait has been mastered is represented by the summed rating of the attributes, since for Rasch measurement, the raw score is considered a "sufficient statistic"[40]. The statistical sufficiency of raw scores means that the person total score includes the complete information available about the individual in the specified context with regard to the relevant latent trait, regardless of response pattern.

Third, the Rasch model theoretically assures that the estimate of item difficulty is independent of the sample used for item calibration[28, 30, 34], which is supported by the result showing that there was no differential test functioning in the present study. We argue that this is the most important property of an assessment tool in terms of utility because this property enables objective comparison both between indi-

Table 3. Convergent validity of the raw score according to work ability index and other subjects' characteristics in the field study (n=1,293)

	n	Mean	SD	*p* for trend*
In the past month, how often did you work in a state where you were concerned/ had issues about your physical condition or health?				<0.001
Not at all	472	12.2	6.5	
One or more days a month	374	15.1	5.8	
About one day a week	192	17.3	6.9	
Two or more days a week	143	19.4	6.0	
Almost everyday	86	21.8	7.4	
Work ability index category				<0.001
Excellent	184	11.3	4.1	
Good	643	13.9	5.1	
Moderate	338	18.2	5.5	
Poor	61	25.6	6.2	
How do you rate your current work ability with respect to the physical demands of your work?				<0.001
Very good	255	13.1	6.4	
Rather good	418	14.3	6.1	
Moderate	479	15.6	6.6	
Rather poor	98	21.5	6.9	
Very poor	17	27.2	7.0	
How do you rate your current work ability with respect to the mental demands of your work?				<0.001
Very good	136	11.9	5.8	
Rather good	424	13.3	6.2	
Moderate	514	15.2	6.8	
Rather poor	161	21.5	6.3	
Very poor	32	26.6	6.8	
Self-rated current work ability compared with your highest work ability ever				<0.001
9–10	214	12.6	5.9	
7–8	580	13.9	4.8	
5–6	333	17.0	7.3	
3–4	104	20.8	7.1	
1–2	28	23.4	8.5	
Absenteeism during the last 12 months				<0.001
None	720	14.3	5.4	
1–9 days	469	16.2	6.5	
10–24 days	54	18.2	7.3	
25–99 days	19	21.0	8.3	
100–354 days	4	23.0	3.8	

* *p* values were derived from ANOVA with the linear trend test.

viduals and between different groups.

Fourth, the raw ordinal scores, which have unknown distances between them, can be converted into linear interval measurement scores and presented as a logit function by the Rasch model. Given the almost perfect linear relationship between the Rasch measurement and the raw score, the complexity associated with converting the raw score to a Rasch measurement does not appear warranted. However, while the raw score does not provide any meaning without reference, the Rasch measure confers interpretability based on the probability function.

Hypothesis testing

Both the discriminant and convergent validity of the WFun was confirmed. The basic principle of hypothesis testing is that hypotheses are formulated about the relationships of scores of an instrument with those of other instruments that measure a similar or dissimilar construct or with regard to differences in instrument scores occurring among subgroups of the subjects[15]. We adopted the WIS and the WOS from the SPS and WAI as instruments that measure similar constructs. We also hypothesized that scores of the WFun differed between the subgroups with regard to the SF-8, absenteeism in the past 1 year, employment status after three months, and status of job disruption based on the formative model. Discriminant validity also showed that there was no difference in WFun scores between subgroups by sex, job type, employment type, or annual income. If the WFun is affected by physical strength, for example, it might indicate a difference between sexes. If the WFun is affected by job satisfaction, it might indicate a difference between subgroups according to annual income or employment type. These results are supporting evidence for the WFun's suitability as a measure of the hypothesized construct concerning the ability to function at work.

Limitations

Several limitations of the present study should be mentioned. First, approximately 9% of respondents did not fit the model, implying that these people are not within the targeted population to be measured by the WFun. However, it is reasonable that not all workers experience impaired work functioning. A typical response pattern misfit to the model was the rating of one or two items as grade 4 or 5, while other items were rated as grade 1. This violates the Guttmann scale, which the Rasch model assumes. These subjects may experience particular job problems rather than the work functioning impairment that the WFun assumes. Nevertheless, identifying subjects with misfit might be of clinical use in occupational health practice.

Second, COSMIN requires a further test for responsiveness, which refers to the ability of an instrument to detect change over time in the construct to be measured.

Third, in terms of hypothesis testing, a hypothesis should state not only the direction but also the magnitude of the difference, but it is presently not possible to rationally hypothesize about the magnitude of difference because the WFun is a newly developed tool. Nevertheless, it appears reasonable to expect that the healthiest group would report a better score than the approximate median value of the WFun and that the poorest health group would report a worse score than the highest 80th percentile of the WFun.

Fourth, the WFun was originally generated in the Japanese language. Items shown in English in the Appendix have been translated for presentation purposes and might be found unsuitable by back translation or cross-cultural differentiation. Accordingly, the validity of the WFun is currently limited to Japanese populations. Nevertheless, the WFun was found to have convergent validity with other standardized tools with cross-cultural validation, such as the SPS, WAI, and SF-8. Further study is needed to assess cross-cultural differences in the WFun.

Conclusion

We developed a seven-item tool to measure the ability to function at work and confirmed that it has good construct validity. The tool also has a good fit to the Rasch model and specific objectivity, and it will be useful in assessing the ability of individuals to function at work and also in evaluating group levels for benchmarking.

Funding: This study was funded by the Occupational Health Promotion Foundation (2014), Japan.

Conflict of interest: The authors declare that they have no conflict of interest.

References

1) Aronsson G, Gustafsson K, Dallner M. Sick but yet at work. An empirical study of sickness presenteeism. J Epidemiol Community Health 2000; 54: 502–9.
2) Dew K, Keefe V, Small K. 'Choosing' to work when sick: workplace presenteeism. Soc Sci Med 2005; 60: 2273–82.
3) 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.
4) Loeppke R, Taitel M, Richling D, et al. Health and productivity as a business strategy. J Occup Environ Med 2007; 49: 712–21.
5) Evans CJ. Health and work productivity assessment: State of the art or state of flux? J Occup Environ Med 2004; 46: S3–11.
6) Burton WN, Morrison A, Wertheimer AI. Pharmaceuticals and worker productivity loss: a critical review of the literature. J Occup Environ Med 2003; 45: 610–21.
7) Brooks A, Hagen SE, Sathyanarayanan S, Schultz AB, Edington DW. Presenteeism: critical issues. J Occup Environ Med 2010; 52: 1055–67.
8) Koopman C, Pelletier KR, Murray JF, et al. Stanford presenteeism scale: health status and employee productivity. J Occup Environ Med 2002; 44: 14–20.
9) 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.
10) Kessler RC, Barber C, Beck A, et al. The world health organization health and work performance questionnaire (HPQ). J Occup Environ Med 2003; 45: 156–74.
11) Lerner D, Amick BC, 3rd, Lee JC, et al. Relationship of employee-reported work limitations to work productivity. Med Care 2003; 41: 649–59.
12) Waddell G, Burton AK, Great Britain. Department for Work and Pensions. Is work good for your health and well-being? London: TSO; 2006.
13) Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA 1995; 273: 59–65.
14) Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000; 38: Ii28–42.
15) de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge (UK): Cambridge University Press; 2011.
16) Mokkink LB, Terwee CB, Knol DL, et al. Protocol of the COSMIN study: consensus-based standards for the selection of health measurement instruments. BMC Med Res Methodol 2006; 6: 2.
17) Edwards JR, Bagozzi RP. On the nature and direction of relationships between constructs and measures. Psychol Methods 2000; 5: 155–74.
18) Gignac MA, Cao X, Lacaille D, Anis AH, Badley EM. Arthritis-related work transitions: a prospective analysis of reported productivity losses, work changes, and leaving the labor force. Arthritis Rheum 2008; 59: 1805–13.
19) Ospina MB, Dennett L, Waye A, Jacobs P, Thompson AH. A systematic review of measurement properties of instruments assessing presenteeism. Am J Manag Care 2015; 21: e171–85.
20) Noben CY, Evers SM, Nijhuis FJ, de Rijk AE. Quality appraisal of generic self-reported instruments measuring health-related productivity changes: a systematic review. BMC Public Health 201414115.
21) Yamashita M, Arakida M. Reliability and validity of the Japanese version of the Stanford Presenteeism Scale in female employees at 2 Japanese enterprises. J Occup Health 2008; 50: 66–9.
22) Ware J, Kosinski M, Dewey J, Gandek B. How to Score and Interpret Single-Item Health Status Measures: A Manual for Users of the SF-8 Health Survey. Boston (MA): QualityMetric Incorporated, Lincoln RI; 2001.
23) Fukuhara S, Suzukamo Y. Manual of the SF-8 Japanese Version. Kyoto. Institute for Health Outcomes & Process Evaluation Research; 2004.
24) Tokuda Y, Okubo T, Ohde S, et al. Assessing items on the SF-8 Japanese version for health-related quality of life: a psychometric analysis based on the nominal categories model of item response theory. Value Health 2009; 12: 568–73.
25) RASCH G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology 1966; 19: 49–57.
26) Andrich D. Rasch models for measurement. Newbury Park; London. Sage; 1988.
27) Fischer GH, Molenaar IW. Rasch models: foundations, recent developments, and applications. New York: Springer-Verlag; 1995.
28) Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. 2nd ed. Mahwah (NJ): Lawrence Erlbaum Associates; 2007.
29) Linacre JM. Winsteps® Rasch measurement computer program. Beaverton, Oregon. Winsteps.com; 2014.
30) Wright BD, Masters GN. Rating scale analysis. Chicago (USA): Mesa Press; 1982.
31) Wright BD, Linacre JM. Reasonable mean-square fit values. Rasch Measurement Transactions 1994; 8: 370.
32) Schumacker R, Smith E. A Rasch perspective. Educ Psychol Meas 2007; 67: 394–409.
33) Linacre JM. Investigating rating scale category utility. J Outcome Meas 1999; 3: 103–22.
34) Nering ML, Ostini R. Handbook of Polytomous Item Response Theory Models; 2011.
35) Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86: 420–8.
36) Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal 1999; 6: 1–55.
37) Kim J, Klein DN, Olino TM, Dyson MW, Dougherty LR, Durbin CE. Psychometric properties of the Behavioral Inhibition Questionnaire in preschool children. J Pers Assess 2011; 93: 545–55.
38) Smith RM. Fit analysis in latent trait measurement models. J Appl Meas 2000; 1: 199–218.
39) Conrad KJ, Smith EV, Jr. International conference on objective measurement: applications of Rasch analysis in health care. Med Care 2004; 42: I1–6.
40) Andersen E. Sufficient statistics and latent trait models. Psychometrika 1977; 42: 69–81.

Appendix 5. Basic characteristics of study subjects

	Pilot study	Field study	Additional samples to examine differential test functioning					
			Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Sample description	Respondents to Internet survey	Air conditioner machinery manufacturer	Heavy machinery manufacturer	Health-care institution	Customer service department, electronic device industry	Cashier at a supermarket	Warehouse employee at medical wholesale dealer	Chemical manufacturer
Number of subjects	1000	1294	620	212	29	117	96	294
Response rate (%)	—	95%<	99%	76%	98%<	95%<	95%<	100%
Men (%)	50%	89%	99%	35%	11%	0% (all women)	0% (all women)	68%
Age (mean and SD)	44 (13) *	41(12)	36 (11)	45 (10)	42 (9)	40 (14)	52 (5)	40 (10)
Job type (%)								
Mainly desk work	51%	25%	10%	47%≠	—	—	—	34%
Mainly work involving interpersonal communication	23%	—	—	15%≠	100%	100%	—	
Mainly physical work	26%	23%	90%	—	—	—	100%	45%
Other	—	52% (technical and research staffs)	—	38%≠ (health nurses and medical technicians)	—	—	—	21% (technical and research staffs)

* Subjects were assigned to age groups according to decade of life (20 s, 30 s, 40 s, 50 s and 60's) with 200 people in each group. ≠ Detail information was not available. The figure shows the staff composition of the workplace.